A Simple, Assumption-Free, and Clinically Interpretable Approach for Analysis of Modified Rankin Outcomes

George Howard, DrPH; Jennifer L. Waller, PhD; Jenifer H. Voeks, PhD; Virginia J. Howard, PhD; Edward C. Jauch, MD; Kennedy R. Lees, MD; Fenwick T. Nichols, MD; Volker W. Rahlfs, PhD; David C. Hess, MD

**Background and Purpose**—There is debate regarding the approach for analysis of modified Rankin scale scores, the most common functional outcome scale used in acute stroke trials.

**Methods**—We propose to use tests to assess treatment differences addressing the metric, “if a patient is chosen at random from each treatment group and if they have different outcomes, what is the chance the patient who received the investigational treatment will have a better outcome than will the patient receiving the standard treatment?” This approach has an associated statement of treatment efficacy easily understood by patients and clinicians, and leads to statistical testing of treatment differences by tests closely related to the Mann-Whitney U test (Wilcoxon Rank-Sum test), which can be tested precisely by permutation tests (randomization tests).

**Results**—We show that a permutation test is as powerful as are other approaches assessing ordinal outcomes of the modified Rankin scores, and we provide data from several examples contrasting alternative approaches.

**Discussion**—Whereas many approaches to analysis of modified Rankin scores outcomes have generally similar statistical performance, this proposed approach: captures information from the ordinal scale, provides a powerful clinical interpretation understood by both patients and clinicians, has power at least equivalent to other ordinal approaches, avoids assumptions in the parameterization, and provides an interpretable parameter based on the same foundation as the calculation of the probability value. *(Stroke. 2012;43:664-669.)*

**Key Words:** ordinal analysis ■ modified Rankin score

See editorial, p 621.

The stroke community is in the midst of a spirited discussion regarding optimal approaches to the analysis and interpretation of the modified Rankin (mRS) score, which is commonly used as an outcome in acute stroke clinical trials.1–6 The mRS is a 7-point scale ranging from 0 (no symptoms) to 6 (dead; Table 1). The most severe scores, representing severe disability (mRS = 5) and death (mRS = 6), are frequently pooled, as these outcomes are considered equivalently bad. Two general approaches for analyzing mRS outcomes have been employed:

**Categorical Statistical Approach Where the mRS Is Dichotomized**

Most studies have taken the approach of either dichotomizing the scale at a fixed point (as in the Albumin in Acute Stroke trial7), or of using a sliding dichotomy, where a good outcome is defined as a function of baseline severity (as in the The Paracetamol [Acetaminophen] In Stroke [PAIS] trial8). This approach has the advantage of simplicity of clinical interpretation (and has been strongly encouraged by the U.S. Food and Drug Administration), but the disadvantage of failing to harness information from the entire spectrum of mRS outcomes. Analysis of the mRS dichotomous groups has been generally implemented by $\chi^2$ testing or by logistic regression.

**Ordinal Statistical Approaches of the Distribution Across the Entire mRS**

Two analytic ordinal approaches have been employed, both having the advantage of increased statistical power provided by using information from the entire mRS and by capturing
Table 1. The Modified Rankin Scale

<table>
<thead>
<tr>
<th>Rankin Score</th>
<th>Clinical Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>No symptoms at all</td>
</tr>
<tr>
<td>1</td>
<td>No significant disability despite symptoms; able to carry out all usual duties and activities</td>
</tr>
<tr>
<td>2</td>
<td>Slight disability; unable to carry out all previous activities, but able to look after own affairs without assistance</td>
</tr>
<tr>
<td>3</td>
<td>Moderate disability; requiring some help, but able to walk without assistance</td>
</tr>
<tr>
<td>4</td>
<td>Moderately severe disability; unable to walk without assistance and unable to attend to own bodily needs without assistance</td>
</tr>
<tr>
<td>5</td>
<td>Severe disability; bedridden, incontinent and requiring constant nursing care and attention</td>
</tr>
<tr>
<td>6</td>
<td>Dead</td>
</tr>
</tbody>
</table>

The Proportional Odds Model

The Proportional Odds Model (POM) is a straightforward approach to the analysis of ordinal scales, both with advantages and disadvantages, specifically the following.

The Proportional Odds Model (POM) is a straightforward approach to the analysis of ordinal scales, both with advantages and disadvantages, specifically the following. Considering the differences for different measures of risk, and are only similar in magnitude if the outcome is rare (generally not the case for analysis of mRS outcomes). In addition, the odds ratio provided by POM does not reflect a “simple” odds ratio, but rather an average across multiple thresholds of classification of the outcome, and as such should not be interpreted as a simple odds ratio.

The Cochrane-Mantel-Haenszel Test

The Cochrane-Mantel-Haenszel (CMH) test is a test of the similarity of the mean rank (across the mRS) for 2 groups. The CMH test has the advantage of requiring virtually no assumptions for calculation of the probability value, but the disadvantage of not providing a clinically interpretable parameter.

In the analysis of the ordinal outcome, some studies (such as SAINT-1) used the CMH test treatment differences, and once significance was established, used the POM to provide a clinically interpretable parameter. Although this approach is generally acceptable, it does have the disadvantage of basing the assessment of the significance on a different approach (different metric) than does the estimation of the magnitude of the effect, as the 2 approaches can potentially give disparate results.

Thus, commonly used current approaches require the choice between the easily interpretable dichotomous analyses that fail to capture the information across the entire mRS spectrum, versus ordinal analyses that either provide no measures of clinical efficacy or provide measures that are commonly misinterpreted.

An Alternative Approach

We set out to find an alternative approach that captures information from the ordinal scale, provides a powerful and easily understood clinical interpretation, has statistical power at least equivalent to the CMH or POM, avoids assumptions in parameterization, and has the clinically interpretable parameter based on the same foundation as the calculation of the probability value. This process began by attempting to state simply the treatment effects in terms that patients and clinicians would understand, and we suggest that the simple statement, “if a patient is chosen at random from each treatment group and if they have different mRS outcomes, what is the chance that the patient who received the investigational treatment will have a better outcome than will the patient receiving the standard treatment?” captures the essential information necessary for clinical decision making. This statistical statement of treatment efficacy can be rephrased to be understood easily by both patients and treating clinicians. For example, the findings of the National Institute of Neurological Disorders and Stroke Tissue Plasminogen Activator trial would be explained as, “Of 100 patients treated with tissue-type plasminogen activator (tPA) instead of with placebo, 48 patients will be better with tPA, whereas 31 patients will be better with placebo. Twenty one (21) patients appear the same with either treatment.” We suggest this simple statement provides the information patients value by focusing on the chance that they will do better with a particular treatment. We are also confident that this statement can be
further refined with future experience working with patients with the interpretation of results.

This approach was initially proposed more than 60 years ago as an approach now referred to as the Mann-Whitney U test (MWU). This idea of whether a randomly chosen person from 1 group has better outcome (ie, lower mRS) than does a randomly chosen person from the other group has been shown to be mathematically equivalent to a ranking approach proposed 2 years earlier (the Wilcoxon Rank-Sum test). Both tests were developed for a continuous outcome where ties are impossible (or at least very rare), but this is not the case in the analysis of mRS outcomes. Approaches to handle the ties, however, have been subsequently proposed and are now very well-accepted. Although these 2 tests are among the most commonly used nonparametric tests, normally the focus is on the probability value, with little emphasis on the real-life, patient-centric description of the magnitude of the clinical effect. However, it is this description of the effect magnitude that was the impetus for the proposed alternative approach.

Because our alternative approach arose from discussions with clinicians to develop a statement of clinical effect, it differs slightly from the MWU on the accounting for tied scores. Specifically, the MWU assumes that half the pairs with tied scores are superior for 1 treatment and half are superior for the other (ie, half the tied scores had a lower mRS score for the patient assigned to investigational treatment, and half had lower mRS score for the standard-treatment patient). Although this approach is mathematically attractive, it leads to an awkward clinical statement of treatment differences—perhaps, "if a patient is chosen at random from each treatment group, and we assume that half the pairs of patients who have the same mRS score did in fact have a better outcome in the investigational-treatment group and half had a better outcome in the standard-treatment group, then what is the chance that the patients receiving the investigational treatment will have a better outcome than will the patient with standard treatment?" We consider this statement awkward to the point that understanding and explaining the clinical effect is a barrier.

Hence, our approach differs from the MWU in how it accounts for the tied values, where we consider only untied pairs of patients and the MWU assumes half to be superior in each group. Therefore, the MWU does not precisely test our proposed clinical statement; however, this can be accomplished by a permutation test (ie, randomization test). The permutation test first calculates a test statistic for the observed data. In this case, we used how far the observed proportion of non-tied pairs was from the null hypothesis of 50%. The permutation test randomly assigns treatments to individuals, ensuring no association between treatment and the test statistic (guaranteeing the null hypothesis is true). This process is repeated many times and the distribution of the test statistic under the null hypothesis is estimated. Whether the observed test statistic is unusual under the null hypothesis (that is, the probability value) is simply the location of the observed test statistic in distribution of test statistics under the null hypothesis. For example, if 234 of 1000 test statistics calculated under the null hypothesi are greater than the observed test statistic, then the probability value is 0.234. Although the construction details of the test are slightly complex and performing this test does require a bit of computer programming, we provide SAS code for the calculations in the Supplemental Data. This concept of permutation test is also not new, but was proposed by Fisher in 1935 as part of the pivotal Lady Tasting Tea experiment; the approach was generalized to randomization test (ie, rerandomization test) when the sample size is large and complete enumeration of all outcomes is impractical. Finally, the correlation between probability values from the permutation test and the MWU in the simulations considered for this report was 0.989; and although this estimated correlation will change with the distribution of mRS scores, this high correlation suggests that the MWU provides an approximate probability value for the more correct permutation test.

The calculations underlying our approach can be demonstrated using data from the National Institute of Neurological Disorders and Stroke tPA trial. Here, the distribution of the mRS in the tPA-treated patients is shown in rows, whereas the mRS scores for placebo-treated patients are shown in columns (Table 2). The first patient from the first group is matched up with each patient in the second group, the second patient in the first group with each patient in the second group, and so on to the last patient in the first group matched up with each patient in the second group; hence, the number of pairs of patients is simply the product of the number of patients in each combination. For example, there were 29 tPA-treated patients with a mRS of 3, and 30 placebo-treated patients with a mRS of 2; hence, there are 29×30=870 pairs of patients with this combination of scores. Altogether, there are 30 178 pairs of patients with higher (worse) mRS for placebo, 46 016 pairs of patients with lower (better) scores for tPA, and 19 596 pairs of patients with a tie between placebo and tPA. As such, 48% of pairs of patients (46 016 of 95 790) had lower (better) scores for tPA, 20% had tied scores (19 596 of 95 790), and 31% had lower (better) scores for placebo (30 178 of 95 790). Hence, these data support the statement, "Of 100 patients given tPA instead of placebo, 48 patients will be better with tPA, whereas 31 patients will be better with placebo. Twenty one (21) patients appear the same with either treatment."

The approach is also flexible. For example, if the primary hypothesis requires stratification by a covariate (for example, stroke severity), the analysis can be adjusted by totaling the of pairs of patients that are better, worse, or tied within each stratum of the covariate, and then by summing across the covariate strata. Importantly, this statistic is calculated by summing the pairs as higher/lower within each strata, and only then calculating the statistic. Because the statistic is not calculated within the strata, but rather across all strata, the approach does not require a large sample size within each stratum to meet asymptotic distribution assumptions for each strata. Hence, the approach does not require large sample size and does allow for the simultaneous adjustment multiple covariates. In addition, the potential for effect modification (interaction with treatment) within a trial can be assessed by calculating the proportion with better outcomes within each stratum, and by using a
permutation test testing if there is a large difference between strata.

Not surprisingly, the statistical power to detect treatment effects using this approach is equivalent to that of the CMH or POM. As demonstrated in Table 3, we assumed a standard treatment mRS distribution of the placebo-treated patients in the National Institute of Neurological Disorders and Stroke tPA study, then created a distribution for patients treated with the investigational drug by shifting an increasingly larger proportion of individuals to lower mRS strata (details of the approach for shifting is provided in Supplemental Material). The power of our approach is indistinguishable from the power of the POM, both of which are generally marginally more powerful than is the CMH test. We stress that it is possible also to specify distributions where the CMH will be marginally more powerful than are either this proposed approach or the POM, and it is also possible to specify more extreme distributions where the dichotomous analysis is most powerful.

Finally, Table 4 provides a comparison of the alternative ordinal approaches for 4 recently reported studies. As expected, the statistical test evaluating treatment effects is concordant among the 3 approaches.

**Discussion**

Our analytic approach meets all the goals we established to achieve in defining a statistical approach, specifically: it captures information from the ordinal scale, it provides a clinical interpretation that is easily understood, it has power at least equivalent to that of the CMH or POM, it avoids assumptions in the calculation of the significance of treatment differences, and the clinically interpretable parameter is based on the same foundation as the calculation of the probability value. We suggest that this approach is:
- Superior to the dichotomous approach, as it captures the entire spectrum of the mRS, and therefore will be generally more powerful.
- Superior to the CMH test, as it provides an easily interpretable index of treatment efficacy that is based on the same metric as is the significance test (ie, does not require the...
use of the POM, a test using a different metric, to provide a measure of efficacy).

- Superior to the proportional odds model as it avoids the proportional odds assumption.

We would suggest that those using this approach report results visually as has been done by many studies using stacked horizontal bar graphs (sometimes referred to in the acute stroke literature as Grotta bars), and also provide the estimate of clinical effect and the probability value for difference in treatment.

On the surface, given that ties are not considered in calculations, it would seem that the proposed permutation test would have lower power than would the MWU, CMH, or POM tests, all of which incorporate the observations with ties into probability value calculations and, as such, have a larger sample size. However, the MWU test assigns half the ties to treatment A and half to treatment B, and as such within the MWU, these observations provide no information to the test (ie, implicitly excluding the observations). Preliminary investigations have shown that even in situations where the large majority of observations are tied, the relationship of the probability value from the MWU to include NNT and measures of association from the MWU to include NNT and measures of effect; however, we hope that the measure of association of the proposed approach is sufficiently straightforward as to be easily understood. Hence, we suggest that the choice of the proposed method versus the alternative MWU is not a major statistical packages, including the computation of confidence bounds for effects and adjustment for covariates. Although we offer the program for calculation of the index and probability value (Supplemental Appendix), these extensions of the technique will both require modest additional programming and will be more computationally intensive (noting, however, that computational burden is becoming less of an issue). Others, including coauthors of this work (V.W.R.), are developing methods for re-expression of measures of association from the MWU to include NNT and measures of effect; however, we hope that the measure of association of the proposed approach is sufficiently straightforward as to be easily understood. Hence, we suggest that the choice of the proposed method versus the alternative MWU is primarily a matter of personal preference; it should focus on the benefit of having the basis for the measure of the association and probability value arising from the same metric versus the additional computational burden of the proposed approach.

One could easily (and perhaps correctly) argue that because the proposed approach, the POM, and the CMH are generally concordant, the choice among them is not a major study design issue. However, the discussion in the neurological community regarding the strengths and weaknesses of the analytic approach for the mRS outcome seems to be a never-ending spirited debate. We do not claim that our approach is universally more powerful than the others are, but rather that it is generally as powerful, yet superior, because it avoids the pitfalls present in other approaches.

### Sources of Funding

This work was supported by R01 NS055728 (David C. Hess, Principal Investigator).

### Disclosures

None.
References

A Simple, Assumption-Free, and Clinically Interpretable Approach for Analysis of Modified Rankin Outcomes
George Howard, Jennifer L. Waller, Jenifer H. Voeks, Virginia J. Howard, Edward C. Jauch, Kennedy R. Lees, Fenwick T. Nichols, Volker W. Rahlfs and David C. Hess

Stroke. published online February 16, 2012; Stroke is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2012 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/early/2012/02/16/STROKEAHA.111.632935

Data Supplement (unedited) at:
http://stroke.ahajournals.org/content/suppl/2012/02/17/STROKEAHA.111.632935.DC1

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/
Supplemental Material

SAS code to perform the re-randomization test for treatment differences in mRS scores (or other ordinal outcomes)
/* this program calculates the indices for treatment difference for the mRS and performs a
permutation test of treatment differences. It assumes that data will include two variables:
l) TRT which is a 0/1 indicator variable for treatment, and 2) MRS which is the modified
Rankin Score. It would be wise to create a dataset that only contains these two variables
(reduces the chance that a variable name in the dataset would be the same as a variable
that was used in the program. */

/* the data step calculates the percent that is better */
data perm;
  set tmp end=last;  /* data set with the data is inserted here */
  retain
    a0 a1 a2 a3 a4 a5
    b0 b1 b2 b3 b4 b5 num_g0;
  array a a0 a1 a2 a3 a4 a5;
  array b b0 b1 b2 b3 b4 b5;
  if _n_ = 1 then do;
    a0=0; a1=0; a2=0; a3=0; a4=0; a5=0;
    b0=0; b1=0; b2=0; b3=0; b4=0; b5=0;
    num_g0=0;
  end;
  if trt=0 then num_g0=num_g0+1;
  if trt=0 then a[mrs+1]=a[mrs+1]+1;
    else b[mrs+1]=b[mrs+1]+1;
  if last=1 then do;
    atot=a0+a1+a2+a3+a4+a5;
    btot=b0+b1+b2+b3+b4+b5;
    better =
      a0*b1 + a0*b2 + a0*b3 + a0*b4 + a0*b5 +
      a1*b2 + a1*b3 + a1*b4 + a1*b5 +
      a2*b3 + a2*b4 + a2*b5 +
      a3*b4 + a3*b5 +
      a4*b5 ;
    worse =
      b0*a1 + b0*a2 + b0*a3 + b0*a4 + b0*a5 +
      b1*a2 + b1*a3 + b1*a4 + b1*a5 +
      b2*a3 + b2*a4 + b2*a5 +
      b3*a4 + b3*a5 +
      b4*a5 ;
    same = a0*b0 + a1*b1 + a2*b2 + a3*b3 + a4*b4 + a5*b5;
    ratio=better/(better+worse);
    dif_ratio=abs(ratio-.5);
    m=1;
    pct_better=better/(better+same+worse);
    pct_same=same/(better+same+worse);
    pct_worse=worse/(better+same+worse);
    format pct_better pct_same pct_worse nlpct5.2;
    output;
  end;
  keep a0-a5 b0-b5 atot btot better same worse pct_better pct_same pct_worse ratio
dif_ratio num_g0 m;
run;

proc print data=perm;
  var a0-a5 b0-b5 better same worse pct_better pct_same pct_worse ratio num_g0;
  title 'Listing of the number of people in each mRS category for treatment 0 (A0-A5),';
  title2 'treatment 1 (B0-B5), the number of pairs that are better (BETTER) and worse
  (WORSE),';
title3 'the ratio that are better as a function of the non-ties (BETTER/(BETTER+WORSE))'
title4 'and the number in TRT=0 (NUM_G0)';
run;
/* now re-randomize */
/* assign a random number for ordering */
data reran;
  set tmp;
  length trt mrs reran_rep t 4;
  keep trt mrs reran_rep t m;
  m=1;
/* this is the "random" seed. This can be any large number. If it is set
to zero, the program will create a randomly selected number. The advantage of
this is runs of the program are independent. The disadvantage is that
running the program on the same data set will provide slightly different answers
(not the case if a large number is entered, which will provide completely identical
estimates */
  retain seed 8382732;
  do reran_rep=1 to 10000; /* number of replications of the re-randomzation. Should be
    at least 5000 */
    t=ranuni(seed);
    output;
  end;
run;
/* randomly order the observations */
proc sort data=reran;
  by reran_rep t;
run;
/* now merge in the PERM file to get the number treated with TRT=0 */
data reran2;
  merge reran perm;
  by m;
  keep trt mrs reran_rep num_g0;
run;
/* now randomly assign treatments */
data reran3;
  set reran2;
  by reran_rep;
  keep reran_rep trt mrs;
  length reran_rep trt mrs 4;
  retain nobs;
  if first.reran_rep then nobs=0;
  nobs=nobs+1;
/* now randomly assign treatment */
  if nobs<=num_g0 then trt=0; else trt=1;
run;
/* now get ratios from each of the reran_reps */
data perm2;
  set reran3;
  by reran_rep;
  retain a0 a1 a2 a3 a4 a5
b0 b1 b2 b3 b4 b5;
array a a0 a1 a2 a3 a4 a5;
array b b0 b1 b2 b3 b4 b5;

if first.reran_rep then do;
a0=0; a1=0; a2=0; a3=0; a4=0; a5=0;
b0=0; b1=0; b2=0; b3=0; b4=0; b5=0;
end;

if trt=0 then a[mrs+1]=a[mrs+1]+1;
else b[mrs+1]=b[mrs+1]+1;

if last.reran_rep then do;
atot=a0+a1+a2+a3+a4+a5;
btot=b0+b1+b2+b3+b4+b5;
better =
a0*b1 + a0*b2 + a0*b3 + a0*b4 + a0*b5 +
a1*b2 + a1*b3 + a1*b4 + a1*b5 +
a2*b3 + a2*b4 + a2*b5 +
a3*b4 + a3*b5 +
a4*b5;
worse =
b0*a1 + b0*a2 + b0*a3 + b0*a4 + b0*a5 +
b1*a2 + b1*a3 + b1*a4 + b1*a5 +
b2*a3 + b2*a4 + b2*a5 +
b3*a4 + b3*a5 +
b4*a5;

reran_ratio=better/(better+worse);
diff_rr_ratio = abs(reran_ratio - .5);
m=1;

output;
end;

length reran_rep reran_ratio diff_rr_ratio 4;
keep reran_rep reran_ratio diff_rr_ratio m;
run;

proc chart data=perm_both;
  hbar reran_ratio diff_rr_ratio;
  title 'Distribution of the score under the null';
run;
/* now get the number with bigger deviations */
data perm_both;
  merge perm perm2;
  by m;
  if dif_ratio>diff_rr_ratio then bigger=0; else bigger=1;
run;
/* now get p-value */
proc means data=perm_both mean;
  var bigger;
  title 'the p-value is the proportion with bigger';
  output mean=pr_perm out=perm_test;
run;