Effect Size Measures and Their Relationships in Stroke Studies

Volker W. Rahlfs, PhD CStat; Helmuth Zimmermann, Dipl. Math.; Kennedy R. Lees, MD, FRCP

Many articles in Stroke have considered good statistical practice for adequate planning and high-power analysis for stroke trials. They have discussed which test may be adequate and powerful, proposals for an effect size measure, and proposals for defining number needed to treat (NNT) based on an ordinal scale (see online-only Data Supplement for citations).

The clinical problem is straightforward. We read results of trials that fall into 3 categories: unequivocally neutral or even negative; overwhelmingly positive; or encouraging but open to various interpretations according to the approach taken to statistical analysis and presentation of findings. This dependence on methodology for the third group undermines our confidence in effective treatments and can prompt unjustified repetition of trials of ineffective treatments. A robust, powerful, and universal statistical approach is required.

The statistical problem is more complex. From dozens of available statistical tests, each may be uniquely powerful in certain circumstances. However, trials intended as confirmatory for regulatory approval or for the use in clinical guidelines demand that the analysis plan be prespecified so that the test of choice should minimize assumptions yet maximize power for the anticipated difference between treatment groups. Furthermore, it is not sufficient to indicate that 1 treatment is significantly different from another: clinical research guidelines require that the magnitude of the treatment effect should be declared using the so-called effect size measure accompanied by its measure of precision, the confidence interval.

Thus, there is need for a robust test, preferably for all data types—binary, ordinal, or continuous—and a test-related effect size measure, with a confidence interval that matches the test-related $P$ value. This requirement for an adequate analysis of study data fortunately restricts the plethora of available tests to a small number of useful candidates.

Here, we describe and explain the relationships between 2 tests, of which 1 is well-known and the other is less familiar. Both are suitable for the analysis of binary, ordinal, and continuous data, and both offer associated confidence intervals. Using intravenous thrombolysis trial data, we can illustrate the relative merits of these tests when compared with other approaches. We will demonstrate why dichotomizing ordinal scales are undesirable. Finally, we can show that in principle each well-known effect size measure can be re-expressed as the popular NNT.

Family of Mann–Whitney Measures

One family of effect size measures fulfills both desirable criteria: robustness, that is minimizing assumptions, and relevance for clinical research in which medicinal products are tested against a reference treatment. This family is based on the concept of proversions. Two groups are contrasted by comparing each patient of 1 group with each member of the other and counting the number of cases for which there is superiority for 1 or the other group. The number of proversions for the 2 groups can be related to the total number of comparisons to express a probability. In statistical textbooks, the probabilities of the 2 groups are called $P(X<Y)$ and $P(Y<X)$. If the scale is not a continuous one but an ordered scale with only a few categories, there will also be patients in each group with equal outcomes, yielding the so-called tied values. The number of ties defines the probability $P(X=Y)$.

With these 3 probability values, we can construct several general measures of treatment superiority. Two possible measures and the principles of this procedure were discussed in a recent publication in this journal. Our discussion concentrates on the following 3 further measures:

1. Mann–Whitney measure of superiority

   $$MW = P(X<Y) + 0.5P(X=Y)$$

2. Mann–Whitney difference

   $$MWD = P(X<Y) - P(Y<X)$$

3. Mann–Whitney centered $MWC = MW - 0.5$
Of note, each of these measures can be transformed into 1 of the other 2, but all are robust measures of superiority, differing only marginally in their interpretation when describing the degree of superiority for presenting study results. There is no assumption involved in this procedure, it requires simple counting.

All 3 measures are related to the well-known Wilcoxon–Mann–Whitney (WMW) test. The Mann–Whitney measure of superiority is the probability that a randomly selected patient of 1 group is better off than a randomly selected patient of the other group. The Mann–Whitney difference (MWD) is also in use because when applied to binary data the measure is identical with the risk difference (RD). These are essentially synonymous with other well-known measures of relevance: Kendall’s τ, Somer d, Goodman–Kruskal γ, etc. The Mann–Whitney centered is less familiar but is important for interpretation in relationship to RD and NNT as explained later. All 3 measures can be interconverted or expressed as a standardized mean difference via relation $MW = \Phi(d/\sqrt{d})$, with $\Phi(\cdot)$ meaning the standard normal distribution and $d$ being the standardized difference (Table 1).

### Displaying Outcomes: the Mann–Whitney Measures and the P–P Graph

Stroke researchers have become familiar with display of outcomes, such as Rankin scales, according to paired horizontal stacked bar graphs, sometimes termed Grotta bars by stroke researchers (Figure 1). However, although these readily illustrate at which levels in the ordinal Rankin scale benefit may have occurred, they do not display the total magnitude of effect so clearly.

There is another graph type that can display all Mann–Whitney measures, their relationships, and even relationships to other measures of interest. This is the percentile–percentile plot (P–P plot), better known to stroke researchers as the receiver operating characteristic curve. This graph gives simply the cumulative distribution of the data of 1 group plotted against that of the other group. The MW superiority measure is equal to the area under the curve of the receiver operating characteristic curve. Familiarity with this makes the concept of the MW superiority measure easy to grasp.

### Table 1. Well-Known Benchmark Values of Cohen’s $d$ and the Conversion to Mann–Whitney Superiority Measure, Mann–Whitney Difference and Mann–Whitney Centered (Average Risk Difference)

<table>
<thead>
<tr>
<th>Cohen</th>
<th>MW</th>
<th>MWD</th>
<th>MWC</th>
</tr>
</thead>
<tbody>
<tr>
<td>−0.8</td>
<td>0.29</td>
<td>−0.42</td>
<td>−0.21</td>
</tr>
<tr>
<td>−0.5</td>
<td>0.36</td>
<td>−0.28</td>
<td>−0.14</td>
</tr>
<tr>
<td>−0.2</td>
<td>0.44</td>
<td>−0.12</td>
<td>−0.06</td>
</tr>
<tr>
<td>0</td>
<td>0.50</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>+0.2</td>
<td>0.56</td>
<td>0.12</td>
<td>0.06</td>
</tr>
<tr>
<td>+0.5</td>
<td>0.64</td>
<td>0.28</td>
<td>0.14</td>
</tr>
<tr>
<td>+0.8</td>
<td>0.71</td>
<td>0.42</td>
<td>0.21</td>
</tr>
</tbody>
</table>

$MW$ indicates Mann-Whitney; MWC, Mann–Whitney centered; and MWD, Mann–Whitney difference.

Taking as an example the data from the European Cooperative Acute Stroke Study-III (ECASS-III) trial of intravenous alteplase versus placebo, Figure 2 shows the P–P plot of both cumulative distributions. Because the curve is completely above the diagonal line, this indicates superiority of alteplase (statisticians call this type of difference stochastic ordering when the line is always above or on the line of identity). The area under the curve is identical with the MW measure, which in this example is $MW=0.546$. This is only a small superiority but is typical for clinical stroke research. Of interest, the lower bound of the confidence interval is 0.508 so that the superiority is statistically significant.

The ECASS-III data represent a convenient situation with a monotonically ascending curve, not far from being convex. The graph illustrates practically no difference between treatments at high Rankin values, representing severe disability or death; only scores in the range 0 to 4 show a difference.

We can also see how this graph incorporates the various possible dichotomizations (responder definitions). Each vertical line from the P–P curve to the diagonal is just the RD of a dichotomization.

### Dichotomization

If we restrict our result to only one of the cut points, we have a smaller area under the curve when working with a convex curve, which means loss of patient information. Also, if the overall superiority is small, the fluctuation at the various cut points is so large that dichotomizing the ordinal scale resembles a lottery, particularly as clinical trials intending dichotomization must prespecify their choice of cut point. We choose to prespecify one of the various possible RDs at our peril unless the choice has unique clinical relevance. Certainly, we do not need to dichotomize to derive an NNT, as we explain later. Ordinal analysis is sometimes loosely described as a shift analysis. In fact, the WMW test seeks evidence for more than simply shift: it can also detect a beneficial treatment difference regardless of the pattern of benefit. The P–P plot or Grotta bars will illustrate the pattern, and if a test of where the maximal difference lies should be required, the Kolmogorov–Smirnov test does this, taking into account the number of possible differences. Unfortunately the Kolmogorov–Smirnov test cannot give a confidence interval around the dichotomized analysis.

### Mann–Whitney Measure and Odds Ratio

The Mann–Whitney measure and the odds ratio (OR) are 2 sides of the same coin. One can be derived from the other, which is useful for the interpretation of study results; these in turn can be transformed into other measures of effect size, including standardized difference or NNT. The formula for obtaining MW from the OR is as follows, as could be demonstrated by our research team (H.Z. and V.R.):

$$MW = \frac{OR}{(OR - 1)^2}[(OR - 1) - \ln(OR)]$$

This procedure assumes proportionality of the odds. If this assumption does not apply, it is safer to calculate the MW statistic directly by counting because this is the gold standard of analysis, making no assumptions. A conversion program
implemented by our research group (H.Z. and V.R.) is also available at http://www.idv-software.com/mw-convert/.

We can check and validate these translation procedures. Taking the data of the ECASS-III study, Table 2 presents results in juxtaposition, first MW is derived directly based on the counting process (the number of patients with better outcome) and then also the proportional OR based on the terms of the \( P \) value calculation (so-called test-based result), and finally the MW is derived from OR.

The equivalence is understandable because the P–P plot shows near proportionality of the ORs. Note that the \( P \) value is identical for all measures, \( P=0.0196 \), so that the associated confidence intervals should not cover the null difference (which is 0.5 for MW and 1.0 for OR). The computer outputs for all these calculations are given in the online-only Data Supplement (Tables I and II in the online-only Data Supplement). In passing, we note that our OR was not calculated as usual based on logistic regression but based on the \( P \) value as proposed by Lachin.

Of interest, the MW measure based on pure counting is assumption free, whereas the calculation of the OR and the derived MW measure demands at least near proportionality of effect. Therefore, the direct way is the gold standard among these procedures.

In practical terms, if treatment, such as an endovascular approach, was to increase the proportion of patients with excellent outcome at the expense of a small increase in mortality or severe independence, then the OR for surviving would not be comparable with the OR for recovering fully and proportionality would not occur. The MW statistic could indicate net benefit, but the pooled OR may be misleading.

### Useful Definitions of RD

It is well-known that the simple RD in a 2x2 table is identical with the MWD. Thus, until recently, it was taken for granted that the MWD is a generalized RD, which can be used for...
Table 2. Relationships of Effect Size Measures; Mann–Whitney Superiority Measure, Calculated Directly and Indirectly via Calculation of Odds Ratio (Proportional Odds Ratio)

<table>
<thead>
<tr>
<th></th>
<th>P Value</th>
<th>Effect Size</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mann–Whitney direct</td>
<td>0.020</td>
<td>0.546</td>
<td>0.508–0.585</td>
</tr>
<tr>
<td>Odds ratio direct</td>
<td>0.020</td>
<td>0.751</td>
<td>0.591–0.955</td>
</tr>
<tr>
<td>Mann–Whitney derived</td>
<td>0.020</td>
<td>0.548</td>
<td>0.508–0.587</td>
</tr>
</tbody>
</table>

calculating NNT based on categories of an ordinal scale. However, using geometric and mathematical arguments, it can be shown that the weighted average of the RDs for the whole area (not only of the empirical RDs), the expected RD, is identical to MW–0.5, which is half of the hitherto assumed amount 2x(MW–0.5)=MWD (H. Zimmermann, Dipl. Math., and V.W. Rahlf, PhD, Comments on number-to-treat derived from ordinal scales, submitted manuscript, 2012). The MWD in this situation is larger than any of the possible RDs, a fact which simply is not reasonable. A recent publication of the Optimising the Analysis of Stroke (OAST) Collaboration proposed calculating NNT from ordinal scales based on the difference of the superiority probabilities of the 2 groups being compared. However, this difference of probabilities is just again the MWD so that the given NNT numbers are too small.

The new finding is of considerable importance for interpretation. We get a fresh interpretation for a MW measure based on RDs. For instance, a RD of 6 percentage points is now identical with an MW=0.56, which according to Cohen effect size definition is a small difference. This seems plausible.

There is another useful definition of the RD that gives an absolute upper bound of RD (maximum RD [MRD]). The derivation formula is as follows:

\[
MRD = \left(\sqrt[3]{OR} - 1\right) / \left(\sqrt[3]{OR} + 1\right)
\]

Both RD definitions are of interest for interpretation, the (summarizing) expected RD and the local maximum RD: each should be cited with its corresponding confidence interval. In the following, we give the expected RD and the derived generalized NNT, as well as maximum RD and derived minimum NNT, for the ECASS-III study data:

RD = 0.046NNT = 21.7

Maximum RD = 0.072minimum NNT = 14.0

Of interest is the comparison of different proposed numbers for RD and NNT, all based on the ordinal scale modified Rankin scale of the study ECASS-III. The described maximum RD for ordinal scales is nearly the RD for a well-chosen cut point of the ordinal scale in well-behaved data situations (convex curve in the P–P plot). The newly proposed expected (weighted average) RD is, of course, smaller. It is only half of the value described by Bath et al and much smaller than the RD given by Saver et al. A detailed table with comparisons of numbers for the different procedures is given in the online-only Data Supplement to this article (Table IV in the online-only Data Supplement).

The question might be raised whether the NNT estimate is attenuated or inflated by nondifferential misclassification (judgment error) as is well-known for dichotomous data. However, Lu et al have shown that, based on work with the Glasgow Outcome Scale (GOS) in the field of traumatic brain injury, the effect of misclassification is likely to be small when ordinal scales are used instead of dichotomous scales. This mitigates against the frequently encountered dichotomization of ordinal scales. In any case, clinical trial researchers should strive to avoid misclassification by applying appropriate measures for enhancing reliability of judgment scale data.

Our approach delivers only the net effect of treatment. Although it may be intriguing to express treatment effects as an NNT for benefit contrasted with an NNT for harm, estimation of these opposing effects has a subjective component, whereas the net effect can be calculated without any assumption unsupported by the trial design.

Choice of Test(s)

It is now well recognized that dichotomizing an ordinal or continuous scale and analyzing the resulting binary scale mean discarding patient information that can only be compensated for by substantially enhancing the patient number in clinical trials. The loss may be ≥50%. Dichotomization and analysis of binary data, the so-called responder analysis, are no longer recommended, at least for first-line analysis. The only justification for reporting is the clinical meaningfulness of the treatment effect.

Despite a multitude of tests that are frequently used for the analysis of ordinal data, all of these tests can be reduced to 2 principles: WMW-type (stochastic superiority or rate difference) and OR-type tests (Table 3).

Because the WMW procedures and OR procedures are closely related, permitting 1 result to be re-expressed as the other (at least for the approximately proportional odds situation), there seems little advantage of one compared with the other. For interpretation, it may be helpful to present both results.

Are there reasons in practical work to prefer one or the other? According to current guidelines, there is 1 basic stipulation: each analysis procedure P value must be accompanied by the associated effect size measure and its confidence interval. This requirement is fulfilled by logistic regression, as well as by the WMW test, with its associated family of MW measures for stochastic superiority and their confidence intervals. In addition, 2 other features are desirable for an efficient data analysis: adjustment for covariates and generalization to multiple end points. The first is available for the ordinal logistic regressions. The second is available for logistic regression as well: it was applied in binary form within the National Institute of Neurological Disorders (NINDS) trial of alteplase by Tilley et al and has been described in ordinal form by Whitehead et al. These 2 desirable features are also available for the WMW test. The first is well-known with the name stratified analysis (eg, van Elteren pooling principle of WMW) available as Cochran–Mantel–Haenszel pooling or formal meta-analysis pooling procedures of WMW results. The generalization of the WMW test to multiple end points with more powerful directional tests is known as the Wei–Lachin procedure and was recently discussed in

...
a similar format with name modified O’Brien. These generalizations allow the combination of binary, ordinal, and continuous data into 1 powerful analysis.

Comparing the relative merits of OR and WMW procedures, we recommend the latter because there are fewer assumptions and restrictions when using the WMW principle.
(proportionality, missing treatment by covariate interaction, power loss by applying binary logistic regression). Thus, we could call the WMW test the gold standard test because it is based on simple ranking of data (Wilcoxon principle) or just counting of ordered comparisons (Mann–Whitney principle). The use of the Mann–Whitney effect size measure also allows the calculation of an average RD and then the derivation of a generalized NNT, which is an effect size measure currently required by many guidelines. However, we emphasize that data analysis using ordinal logistic regression in most analyses of stroke trials should give correct and unbiased results because the procedure seems to be reasonably robust about the proportionality assumption. This could be verified with empirical data from studies in the field of stroke.

Sample Size Determination

The sample size calculation method of Whitehead with respect to ordinal data with few categories is well-known. For the WMW test, Frick and Rahlfs developed a more exact method from a proposal by Noether (Program available). Other Approaches

Some other data evaluation methods in use should be mentioned here. Analysis of ordinal scales with only a few categories should not be performed using the differences of medians and the median test. This analysis is not sensitive to small improvements (in most cases there is only a difference of 0 or 1 point). Note that this approach failed for ECASS-I. Although there are more sensitive tests available (Su-Wei procedure), these tests should only be applied when the data situation is one of pure shift (for each individual patient), which cannot be true for bounded scales (eg, 0–6 modified Rankin scale). Although the t test with its test for mean difference is more sensitive for small differences and in well-behaved data situations delivers comparably good results, we would not recommend it as a first-line analysis procedure because of its strong dependence on the specific data situation.

About continuous or quasi-continuous data, we would not recommend parametric procedures, such as t test, linear regression, ANOVA, and ANCOVA, when for this data situation more adequate and more powerful procedures are available. Robust procedures lose little but can gain a lot when applied instead of parametric procedures. An additional advantage of the nonparametric WMW test is that all scales may be analyzed with a comparable effect size measure using the same procedure.

Conclusions

In summary, the MW effect size measure and the derived generalized NNT number deliver an overall picture for the complete ordinal scale, giving the net effect of superiority and inferiority across the whole scale, and the result does not rest on assumptions (gold standard test). This is a substantial advantage when performing confirmatory studies that require prespecified analysis methods. Examination of the data with a P–P plot can offer information about the nature of the difference between groups although Grotta bars provide some corresponding information and are now familiar to stroke researchers. For instance, the P–P graph can show why one dichotomization gives a favorable result, whereas another does not. If one prefers the OR as the first-line analysis, we strongly recommend comparing the result with that of the gold standard analysis. If results agree, there is no objection to calculate the OR.

Figure 3 gives an overview of all discussed effect size measures and their relationship. In conclusion, we recommend the Mann–Whitney effect size measure (and WMW test). This is not only because it is in itself a valuable robust measure of a beneficial treatment effect, but also because it is the missing link in relationships among other well-known effect size measures. Possessing the MW measure, we can have in principle all other well-known measures.

Disclosures

Dr Rahlfs is active as consultant for EVER Neuro Pharma and receives honoraria for this activity. H. Zimmermann is a consultant for mathematical statistics for idv-Data Analysis and Study Planning. Dr Lees chaired the European Stroke Organisation working group on stroke outcome measures; chairs the outcomes adjudication committees for the Clot Lysis: Evaluating Accelerated Resolution of Intraventricular Hemorrhage III (CLEAR III), EuroHyp-1, and SITS-OPEN trials; chairs independent data monitoring committees for trials in stroke sponsored by commercial organizations including Boehringer Ingelheim, Lundbeck, Grifols, Fundació Privada ICTUS Malaltia Vascular; chairs the Virtual International Stroke Trials Archive; has consulted with EVER Neuro Pharma; and participated in the Stroke Academic Industry Roundtables (STAIR) I-VIII but has received no support from any source in connection with the present article.

References

12. Naver JL, Gornhein J, Grotta J, Liebeschid D, Lutsep H, Schwamm L, et al. Number needed to treat to benefit and to harm for intravenous tissue plasminogen activator therapy in the 3- to 4.5-hour

Key Words: effect size measure ▪ Mann-Whitney measure ▪ numbers needed to treat ▪ proportional odds ratio ▪ randomized controlled trial ▪ risk difference
Effect Size Measures and Their Relationships in Stroke Studies
Volker W. Rahlfs, Helmuth Zimmermann and Kennedy R. Lees

Stroke. 2014;45:627-633; originally published online December 26, 2013;
doi: 10.1161/STRÖKEAHA.113.003151
Stroke is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2013 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/45/2/627

Data Supplement (unedited) at:
http://stroke.ahajournals.org/content/suppl/2013/12/26/STROKEAHA.113.003151.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

Effect Size Measures and Their Relations in Stroke Studies

1 Volker W. Rahlfs, PhD Cstat
2 Helmuth Zimmermann, Dipl. Math.
3 Kennedy R Lees, MD FRCP

1 idv, Data Analysis and Study Planning, 82152 Krailling, Germany
2 idv, Data Analysis and Study Planning, 82152 Krailling, Germany
3 Acute Stroke Unit and Cerebrovascular Clinic, Institute of Cardiovascular & Medical Sciences, Gardiner Institute, Western Infirmary & Faculty of Medicine University of Glasgow, Glasgow, G11 6NT, UK

Correspondence: Dr Volker W. Rahlfs, PhD, idv Data Analysis and Study Planning, Konrad-Zuse-Bogen 17, 82152 Krailling, Germany; Fon: +49.89.8508001, Fax: +49.89.8503666, Mail: volker.rahlfs@idvgauting.de
This supplement gives more information for parts of the paper

1. Literature
The background literature on statistical testing in stroke includes papers on which test is optimal for power (Tilley, Marler 1996; The optimizing analysis of stroke trials (OAST) 2007, 2008; Bath, Geeganage, Gray, Collier, Pocock 2008)1,2,3,4, including proposals for an effect size measure such as odds ratio (Lee, Hong, Saver 2010)5, rate ratio (relative risk): (Sillberg, Wells, Perry 2008, Verro, Gorelick, Nguyen 2008)6,7; stochastic superiority: (Liang 2008; Howard, Waller et al. 2012)8,9 or on use of dichotomization versus ordinal data (Bath, Geeganage, Gray, Collier, Pocock 2008)4; and finally consideration of proposals for defining number-needed-to-treat based on ordinal scale, use of a paired or crossover design (matched pairs) (Walter 2001)10; all pairs (Bath et al. 2011)11; joint distribution tables constructed with expert panels (Saver 2004; Saver 2007; Saver, Gornbein et al. 2009; Lansberg, Schrooten et al. 2009)11,12,13,14,15 or the mean difference (Koziol, Feng 2006)16.

2. Facts about Mann-Whitney effect size
All measures of effect size we describe are related to the Wilcoxon-Mann-Whitney (WMW) test. Though the test may be well known, its related effect size measures are less familiar to stroke clinicians. The Mann-Whitney measure is the best-known of several measures all based on simple counting of paired comparisons (provisions); it was rediscovered several times by various authors, many times with another name: measure of stochastic superiority, Vargha and Delany 200017; competing probabilities (CP), Rahadja and Zhao 200918; Individual exceedance probability (IEP), Senn 199719, Senn 200620; Probabilistic index, Acion, et al. 200621, Generalized treatment effect, Tian, Li, Yan 201222; relative effect, Brunner, Munzel, Puri 200223; C (concordance), C-statistic, Harrell’s C, Harrell et al. 198224; AUC of ROC curve, Kraemer et al. 200325.

3. Odds ratio calculation
The proportional odds measure is usually calculated using the logistic regression model. We used a model-free approach. Because of the duality of P-value and confidence interval we can calculate the odds ratio and its confidence interval based on the same values used for calculation of the P-value (Lachin 2000, also Bennett 1983)26,27.

4. P-P plot and example
In addition to stacked bar graphs (Grotta bars) we make use of a graphic tool which has stood the test of time in the statistical literature because it is extremely useful based on its associations with several well-known powerful tests. One graph type can display all Mann-Whitney measures, their relations, and even relations to other measures of interest. This graph type is the percentile-percentile plot, in short P-P plot (Wilk and Gnanadesikan 1968, Holmgren 1995)28,29, also called the receiver operating curve (ROC). This graph gives the cumulative distribution of the data of one group plotted against that of the other group. Bamber (1975)30 showed that the MW stochastic superiority measure is equal to the Area under the Curve (AUC) of the Receiver Operating Characteristics (ROC) curve in medical diagnostic testing. Because of this relation, the concept of MW measure can be quickly and easily grasped by researchers.
As an example from research in the field of stroke we use the data of the study ECASS-III (Bluhmki et al. 2009). The tableau with computer output (Table I) gives the raw counts for the ordinal scale, the percent value of each class, and finally the MW measure with associated confidence interval and the P-value (several different calculations) resulting from the Wilcoxon-Mann-Whitney test.

The associated P-P plot is given in the main part of the paper. As the curve is completely above the diagonal line this is an indication for superiority of alteplase (in a strong sense of stochastic ordering). The area under the curve is identical with the MW measure, which in this example is MW = 0.546. This is only a small superiority, but typical for this area of research. Of interest, the lower bound of the confidence interval is LB-CI = 0.508, so that the superiority is statistically significant.

5. Determination of sample size
The determination of the necessary sample size is important for planning a new study. The sample size calculation method of Whitehead for proportional odds with respect to ordinal data with few categories is well-known (Whitehead 1993). Less well-known is the adequate calculation for the WMW test. An early paper of Noether drew the attention to a procedure based on the MW effect size measure (Noether 1987). Based on Noether’s proposal Frick and Rahlfs (1998) developed a more exact method based on adequate definitions of variance estimators (Program available as Nnpar from idv, www.idv-cro.com). A comparison of several procedures in use to determine their relative merits was published by Lesaffre et al. (1993). A discussion of Noether’s method was given by Vollandt and Horn (1997). A more recent publication dealing with this subject is that of Huang et al. (2008). Their results are in principle identical with those of Frick and Rahlfs. The calculation of sample size for a median difference described by The Optimising Analysis of Stroke Trials (OAST) Collaboration 2008 seems to overestimate grossly the number of necessary patients for a clinical trial. From its very nature (median difference) this procedure does not fit the associated effect size measure.

6. Various effect size measures, all related
There are many other well-known effect size measures, which in most cases are related by simple algebraic formulas. Fig. 3 in the print version gives an overview with arrows indicating possible back and forth conversion. The output of the odds ratio calculation and the derivation of the MW statistic is given here as table II.

A program for conversion of proportional odds ratio (POR) in Mann-Whitney measure and vice versa is available at http://www.idv-software.com/mw-convert/.

We demonstrated the relation of the odds ratio and the MW measure in this paper. We can transform each measure into the other, in most cases with a simple formula (with the exception of the re-expression of MW as proportional odds which can only be performed based on a numerical iterative procedure). Fortunately all these measures can also be reexpressed with confidence intervals using the so-called substitution rule (Daly 1998).

7. Relations to benchmark values.
The most important relation is that of the standardized mean difference (SMD) and the MW measure and the odds. The benchmark values are useful for answering the
question”How large is large”. The benchmark values for SMD are sometimes useful, and when accepted for SMD should also be useful for the derived measures (even when SMD is not normally distributed). Table III gives an overview of some benchmark values for SMD and other important related measures. The table gives equivalent values of proportional odds and MW-measure as well as derived Risk Difference (RD) [average and maximum] and corresponding NNT values. Note that NNT min is always smaller than the average NNT. A table with comparisons of RD and NNT numbers based on different proposed procedures is given in table IV.
Data from Bluhmki et al. (2009), Stroke treatment with alteplase given 3.0–4.5 h after onset of acute ischaemic stroke (ECASS III): additional outcomes and subgroup analysis of a randomised controlled trial

Scores
Absolute Values; Data as Available

Wilcoxon-Mann-Whitney-U Test

<table>
<thead>
<tr>
<th>Scores</th>
<th>Sc 0</th>
<th>Sc 1</th>
<th>Sc 2</th>
<th>Sc 3</th>
<th>Sc 4</th>
<th>Sc 5</th>
<th>Sc 6</th>
<th>Valid No.</th>
<th>Mean Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>78.0</td>
<td>245.5</td>
<td>397.0</td>
<td>503.5</td>
<td>605.0</td>
<td>719.0</td>
<td>799.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alteplase (N = 418)</td>
<td>94</td>
<td>94</td>
<td>57</td>
<td>45</td>
<td>49</td>
<td>56</td>
<td>23</td>
<td>418</td>
<td>392.300</td>
</tr>
<tr>
<td>Placebo (N = 403)</td>
<td>61</td>
<td>86</td>
<td>66</td>
<td>45</td>
<td>64</td>
<td>59</td>
<td>22</td>
<td>403</td>
<td>430.396</td>
</tr>
</tbody>
</table>

%: row percent

Direction of Location: (T): Alteplase < (R): Placebo
Rank Sum (T) 163981.5; Exp. Rank Sum 171798.0

Test for Difference (two-sided)
Asymptotic
uncorrected P = 0.0196 (X-sq= 5.4444, DF = 1)
cont. corr. P = 0.0196 (X-sq= 5.4437, DF = 1)
t approx. P = 0.0199 (t = 2.3333; DF = 818)

Effect Size Measures and Confidence Intervals
P(X<Y) 0.4676 P(Y<X) 0.3748 P(X=Y) 0.1576
S.E. (Wei-Lachin) 0.0198

Mann Whitney Statistic 95.0% CI
P(X<Y) + 0.5 P(X=Y) 0.5464 0.5076 to 0.5852
Mann Whitney Difference P(X<Y) - P(Y<X) 0.0928 0.0152 to 0.1704
Odds Ratio P(X<Y) / P(Y<X) 1.2476

Table I

Computer Output of program TESTIMATE from idv based on data from Bluhmki et al. 2009, fig. 31
Data from Bluhmki et al. (2009), Stroke treatment with alteplase given 3.0-4.5h after onset of acute ischaemic stroke (ECASS III): additional outcomes and subgroup analysis of a randomised controlled trial

Scores
Absolute Values; Data as Available
Cochran-Mantel-Haenszel Pooling
Peto-Wilcoxon Test

<table>
<thead>
<tr>
<th>SCORES GROUPS</th>
<th>Sc 0</th>
<th>Sc 1</th>
<th>Sc 2</th>
<th>Sc 3</th>
<th>Sc 4</th>
<th>Sc 5</th>
<th>Sc 6</th>
<th>Valid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alteplase (N = 418)</td>
<td>94</td>
<td>94</td>
<td>57</td>
<td>45</td>
<td>49</td>
<td>56</td>
<td>23</td>
<td>418</td>
</tr>
<tr>
<td>Placebo (N = 403)</td>
<td>61</td>
<td>86</td>
<td>66</td>
<td>45</td>
<td>64</td>
<td>59</td>
<td>22</td>
<td>403</td>
</tr>
<tr>
<td>P =0.0196 (Xsq= 5.4444, DF= 1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Proportional Odds Ratio and Confidence Interval
(Variance Estimate: Peto-Wilcoxon)
Reference group: Placebo

<table>
<thead>
<tr>
<th>Statistic</th>
<th>95.0% CI (LB, UB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.3310</td>
<td>( 1.0468, 1.6923)</td>
</tr>
</tbody>
</table>

Mann-Whitney Statistics and Confidence Interval
(Variance Estimate: Peto-Wilcoxon)
Reference group: Placebo

<table>
<thead>
<tr>
<th>Statistic</th>
<th>95.0% CI (LB, UB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5475</td>
<td>( 0.5076, 0.5869)</td>
</tr>
</tbody>
</table>

Table II

Computer Output of program TESTIMATE from idv based on data from Bluhmki et al. 2009, fig. 2
### How large is large?

<table>
<thead>
<tr>
<th>Interpretation of results</th>
<th>Measures of Relevance with Relations</th>
<th>MEASURES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standardized Difference MW POR ARD MRD NNT average NNTmin</td>
<td></td>
</tr>
<tr>
<td>Large superiority</td>
<td>0.8 0.71 3.798 0.21 0.322 4.76 3.11</td>
<td></td>
</tr>
<tr>
<td>Medium superiority</td>
<td>0.5 0.64 2.365 0.14 0.212 7.14 4.77</td>
<td></td>
</tr>
<tr>
<td>Small superiority</td>
<td>0.2 0.56 1.436 0.06 0.090 16.67 11.11</td>
<td></td>
</tr>
<tr>
<td>Equality</td>
<td>0.0 0.50 1.000 0.00 inf. inf.</td>
<td></td>
</tr>
<tr>
<td>Small inferiority</td>
<td>−0.2 0.44 0.697 −0.06 −0.090 (−16.67) (−11.11)</td>
<td></td>
</tr>
<tr>
<td>Medium inferiority</td>
<td>−0.5 0.36 0.423 −0.14 −0.212 (−7.14) (−4.72)</td>
<td></td>
</tr>
<tr>
<td>Large inferiority</td>
<td>−0.8 0.29 0.263 −0.21 −0.322 (−4.76) (−3.11)</td>
<td></td>
</tr>
</tbody>
</table>

inf.= infinity; numbers in brackets: harm instead benefit

MW = Mann-Whitney measure, POR = Proportional Odds Ratio, ARD = Average Risk Difference, MRD = Maximum Risk Difference, NNT = Number-Needed-to-Treat

Table III
## Comparison of several RR and NNT calculations

<table>
<thead>
<tr>
<th>Method Description</th>
<th>RD</th>
<th>NNT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Method proposed by Saver, ECASS III results cited in Saver et al. 2009&lt;sup&gt;14&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Dichotomisation 0-1 vs 2-6</td>
<td>0.073</td>
<td>13.7 (14.3†)</td>
</tr>
<tr>
<td>• Global, expert joint outcome table</td>
<td>0.164</td>
<td>6.1</td>
</tr>
<tr>
<td>• Global, other method</td>
<td>0.139</td>
<td>7.2</td>
</tr>
<tr>
<td>2. OAST-method (Bath et al. 2011)&lt;sup&gt;11&lt;/sup&gt;, global</td>
<td>0.093*</td>
<td>10.8*</td>
</tr>
<tr>
<td>3. Our proposed method, global</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Expected/generalised</td>
<td>0.046</td>
<td>21.7</td>
</tr>
<tr>
<td>• Maximum/minimum</td>
<td>0.072</td>
<td>14.0</td>
</tr>
</tbody>
</table>

*Our calculation based on OAST proposal. Note that OAST RD is always larger than our average RD by a factor of 2. †The value 14.3 was given by Bluhmki et al. 2009<sup>31</sup> using the same calculation procedure.

Table IV
References:


14. Saver JL, Gornbein J, Grotta J Liebeskind D, Lutsep H, Schwamm L. Number needed to treat to benefit and to harm for intravenous tissue plasminogen activator therapy in the 3-
to 4.5-hour window: Joint outcome table analysis of the ECASS 3 trial. Stroke. 2009; 40: 2433-2437

15. Lansberg MG, Schrooten M, Bluhmki E, Thijs VN, Saver JL. Treatment time-specific number needed to treat estimates for tissue plasminogen activator therapy in acute stroke based on shifts over the entire range of the modified ranking scale. Stroke. 2009; 40: 2079-2084


30. Bamber DC. The area above the ordinal dominance graph and the area below the receiver operating characteristics curve graph. J Math Psychol. 1975; 12: 387-415


33. Noether GE. Sample size determination for some common nonparametric tests. J Am Statist Assoc. 1987; 82: 645-647


