Boosting The Chances To Improve Stroke Treatment

Erik Cobo, PhD; Julio J. Secades, MD; Francesc Miras, BSc; José Antonio González, PhD;
Jeffrey L. Saver, MD; Cristina Corchero, BSc; Roser Rius, PhD; Antoni Dàvalos, MD

Background and Purpose—There is a lack of agreement regarding measuring the effects of stroke treatment in clinical trials, which often relies on the dichotomized value of 1 outcome scale. Alternative analyses consist mainly of 2 strategies: use all the information from an ordinal scale and combine information from several outcome scales in a single estimate.

Methods—We reanalyzed 3 outcome scales that assessed patient recovery (modified Rankin Scale, National Institutes of Health Stroke Scale, and Barthel Index). With data collected from the 1652 patients in the Citicoline pooling data analysis, we used 2 standard techniques of exploratory multivariate analysis to analyze the distances among ranks and to isolate the common and the unique information provided by each of the 3 scales.

Results—The different scale values correspond to gradually different patient status, confirming that information is lost when a scale is collapsed to just 2 values, whether recovered or not. The scales shared 90.7% (95% CI, 84.5–96.9) of their information, with no individual scale contributing unique information.

Conclusions—Salient stroke outcome information is lost when an ordinal scale is collapsed into fewer categories. In contrast, the full scales provide a comprehensive patient outcome estimate. Furthermore, in the context of stroke clinical trials, those scales are highly correlated, providing the rationale to pool them into a single estimate. These insights may be used to optimize the analysis of stroke trials to increase study power to detect efficacious interventions. **(*Stroke.* 2010;41:e143-e150.)**

Key Words: analysis ■ biostatistics ■ clinical trials ■ scales ■ stroke management ■ stroke recovery

Stroke is the second most common cause of death and a major cause of disability worldwide.[1] Although at least 178 randomized clinical trials enrolling >73 000 patients were conducted for 75 promising agents over the 20th century, only 3 trials reported positive findings and only 1 agent has been approved by the Food and Drug Administration for use in acute cerebral ischemia.[2]

The pivotal phase III clinical trial is used to guarantee that only a small α-proportion of nonefficacious interventions are approved for application to patients. Statisticians have developed methods to maximize study power; that is, the proportion of interventions with true biological efficacy that are approved. In addition to the magnitude of the intervention effect and the sample size, power is mainly related to measurement reliability and statistical analysis. Although patient recovery after stroke is routinely measured by ordinal outcome scales, the most popular statistical analysis[3] has been comparing the proportion of recovered patients, defined as those with scores above a single prespecified cut-point on a single outcome scale. As Wardlaw et al[4] have shown, "a change of a single point on the Ranking scale can make the difference between 'success' and 'failure.' of the trial.

In this article, we address 2 increasingly used though controversial alternatives to the traditional, dichotomized, single-scale analysis: (1) the use of the entire range of information captured in an ordinal scale; and (2) the use of more than just 1 single scale. For the former, opinions are clearly divided between those in favor of using all the ordinal information to maximize power[5,6] and those in favor of dichotomization because of its readability.[7] If the scale is truly ordinal—that is, if differences in scale categories are related to some progressive degree of biological or clinical difference in patient outcome— then some information is lost when different categories are merged. The second controversy confronts the classical univariate effect estimate based on just 1 scale with an effect estimate based on information pooled from several—often 3 to 4—scales ("global test").[8,9] The rationale is that if more information is used from every patient, fewer patients would be needed. One regulatory agency has argued that this global estimate causes difficulties in its clinical interpretation[10] and its explanation to patients. An additional concern is how to interpret a single "global" effect if discordant treatment effects are observed on each component scale. The key point is whether those scales are

Received September 10, 2009; accepted November 4, 2009.

From Department of Statistics and Operations Research (E.C., F.M., J.A.G., C.C., R.R.), Technical University of Catalonia, Barcelona, Spain; Medical Department (J.J.S.), Ferrer International, Barcelona, Spain; Stroke Center and Department of Neurology (J.L.S.), David Geffen School of Medicine at the University of California, Los Angeles, Calif; Department of Neurosciences (A.D.), Hospital Germans Trias i Pujol, Badalona, Spain.

Correspondence to Erik Cobo, DEIO, UPC, c/ Jordi Girona 1, C5-213, 08034 Barcelona, Spain. E-mail erik.cobo@upc.edu

***Stroke* is available at http://stroke.ahajournals.org** **DOI: 10.1161/STROKEAHA.109.567404**

measuring different patient characteristics or just 1 single dimension of patient recovery.

We use 2 well-established methods of explanatory multivariate analysis to study 2 assumptions underlying alternative methods. First is the existence of some progressive order among the different outcome values of ordinal scales. Do patients belong to just 2 alternative categories (success and failure), or may they be ranked within the full scope? To put it more simply, does an order make sense? Second is the amount of common vs unique information provided by the 3 most usual scales. Does a combination of outcome scores make sense, or does this pooling imply that some specific information is lost? In addition, the correspondence among values of different scales is provided.

Materials and Methods

Patient Population

We analyze the individual data pooled in a meta-analysis of Citicoline,[11] which merged 4 clinical trials conducted in the United States between 1997 and 2001. They included 1652 patients with a 24-hour previous stroke. Main clinical and demographic characteristics can be found elsewhere.[11]

Recovery Measures

Patients were assessed 12 weeks after stroke using the Modified Rankin Score (mRS), the National Institute of Health Stroke Score (NIHSS), and the Barthel Index (BI). The worst values on the NIHSS (42), BI (0), and mRS (6) were imputed to the 275 patients who died before 12 weeks had passed.

Bivariate Analysis of Ordering

To explore the existence of a genuine rank ordering of the levels of each scale, we analyze if there is a trend in the central values (means and medians) of 1 scale given the values of the other scales (eg, we compare the mean NIHSS scores for patients rated 0–6 on mRs).

Simultaneous 3-Dimensional Analysis of Ordering

Multiple correspondences analysis (MCA) explores the relationship between the categories of the different scales. It makes no assumption about their order. For all cases in each scale category (eg, mRS=1), it computes its profile, that is, the percentage of patients in each category of the remaining scales. Closer categories are expected to have more similar profiles. The degree of dissimilarity between the profiles of 2 categories is quantified by the χ^2 distance. Based on algebra analysis, MCA defines a new multidimensional space and locates each category in this new space to reproduce their original distances with the minimum possible number of independent dimensions. As 2 analogies, the distances among 50 European cities should be able to be plotted in just 2 dimensions: north-to-south and east-to-west, but the distances between 50 railway stations of the same line should be able to be plotted in only 1 dimension. The French term "analyses des correspondances" was used to denote a "system of associations" between the categories in a data set. Hence, there is an agreement between the positions (in the new dimensions defined by MCA) of the categories in terms of their association in the data matrix. It has been used in the medical field to study the interrelationships among risk factors[12] and to reduce highly correlated categorical variables to fewer but more significant dimensions. For example, Briand[13] reduced 5 exposure variables to 1 aggregated exposure indicator. Islami[14] constructed a composite wealth score using multiple socioeconomic status assessments. Here, we apply MCA to analyze the distances among the successive ranks.

Common and Specific Information Behind the 3 Scales

The amount of common information shared by the 3 scales has been quantify with principal component analysis (PCA). MCA and PCA are very similar, with the difference being that the latter assumes a measurement unit and then it is able to use means and variances. PCA considers variability as information that allows differentiation of patients. It defines a new dimension (component or factor) that—as much as possible—retains the variability contained in the original variables. The information not included in the first component is then used to define a second component that, being independent to the former, retains, again, as much information as possible. PCA also provides an R^2-like measure of the amount of original information retained in the new space.

We also assessed the implications of unit measure and dichotomization by comparing the values of the parametric (Pearson), the rank (Spearman), and the dichotomized (Phi) correlation coefficients between pairs of variables.

MCA and PCA have been implemented using version 5.6.0 of SPAD (Système Portable pour l'Analyze de Données). Confidence intervals (CI) have been computed by standard intensive jackknife resampling methods using version 2.6.0 of the free software R.

Results

Description

In the 4 studies, 1652 patients were included and they provided values for the 3 outcome scales. Figure 1 and Table 1 show the shape and descriptive statistics of the 3 outcome scales.

Analysis of Ordering

Analyses of the means from 1 scale along the values of other scales suggest the existence of a natural ordering within the categories of 3 scales, with almost no ties and only some minor discordant results (Figure 2).

Figure 3 shows the position, with CI, of each category within the first MCA dimension. The BI value 100 is plotted close to NIHSS and mRS values 0, 1, and 2. On the other side of the new dimension, the 3 worst values of each scale are represented together. Only 4 inversions have been observed: 2 on NIHSS (value 12 fell into position before 11; and 10 fell into position before 8 and 9) and 2 inversions on BI (categories 25 and 30, as well as 5 and 10). In the center we found mRS value of 4, BI values of 25 and 30, and NIHSS values of 8, 9, and 10.

Uniform Unit of Measurement

Spearman-Rank correlations among the 3 scales (Table 2) ranged from 0.89 to 0.94, being slightly superior to the Pearson unit-measure correlations of 0.84 to 0.90 but clearly superior to the Phi correlation between the dichotomized scales of 0.58 to 0.68.

Taking into account the numeric values of the 3 scales, the first PCA component retained 90.7% (95% CI, 84.5%–96.9%) of the overall information, quantified by the sum of the 3 variances of the original 3 scales (Figure 4). The remaining 2 components were quantitatively negligible (6.0% and 3.3%) and qualitatively showed no pattern of information specifically related to any scale.

Discussion

Scale Ordinality

Our results provide evidence that for the leading outcome scales measured 12 weeks after stroke, the different values of

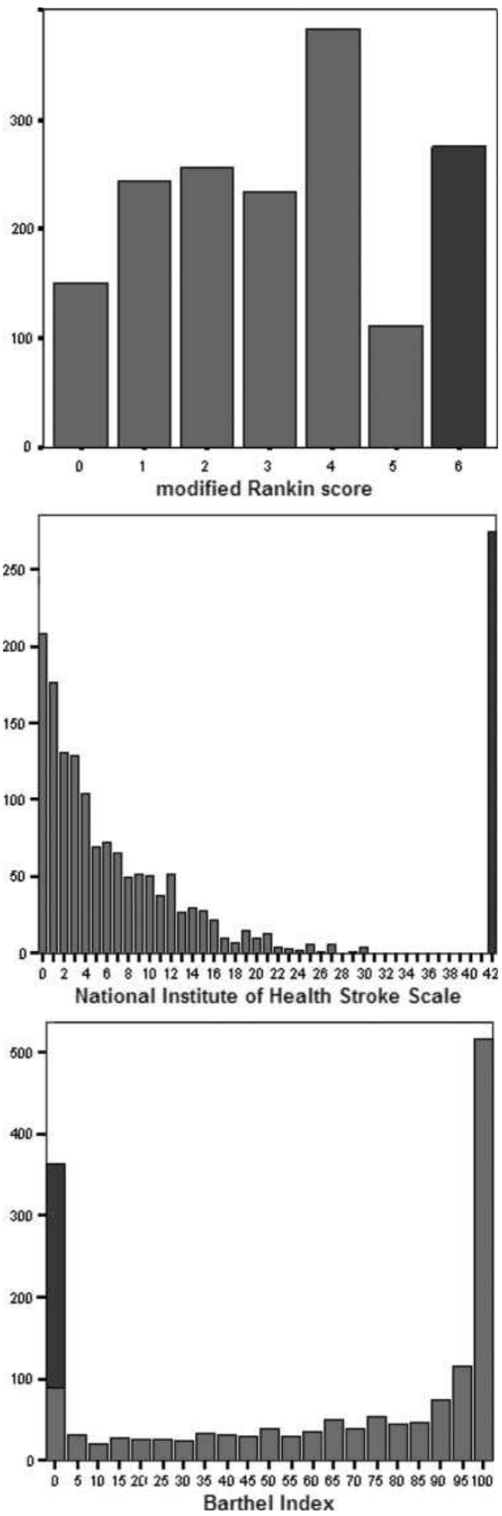

Figure 1. Univariate distributions for the outcome neurological scales on day 90. The mRs is a global measure of disability and handicap ranging from 0 (no symptoms) to 6 (death). NIHSS is a neurological impairment scale ranging from 0 (no deficits) to 42, rated by an expert neurologist. BI is a functional activities of daily living scale ranging from 0 (completely disabled) to 100 (fully independent), with only 21 possible values ending in 0 or 5 (eg, 0, 5, 10, . . . 95, 100). The mRS is more uniform than the other scales, with a mode on the value of 4. NIHSS is skewed to the lowest values, although the worst value, 42, is strongly influenced by its imputation to dead patients. BI is U-shaped, with values concentrated at both extremes. Imputed values to the 275 patients who died are highlighted.

Table 1. Descriptive Statistics for the Neurological Scales on Day 90

	Median (IQR)	Mean (SD)
mRS	3 (2)	3.1 (1.9)
NIHSS	6 (12)	11.9 (14.5)
BI	75 (90)	59.7 (40.8)

IQR indicates interquartile range.

each correspond to different patient status and those values are gradually ordered. This finding supports the recommendation of the European Agency for the Evaluation of Medicinal Products that "dichotomization of outcome (positive–negative) is not recommended for neurological scales as patients in the same category may be clinically distinct and important information might be missed."[15]

However, from a medical point of view, the objective could be "whether subjects finish the study with minimal or no disability" and, then, the main analysis may rely on the proportion of patients reaching this goal. This medical objective should be weighted with the statistical objective of maximizing the trial power, as well as the need to provide regulators, clinicians, and patients with a readable treatment effect measure.

The binary correlations in Table 2 are lower than those reported by Tilley et al[8] for the NINDS tissue plasminogen activator stroke trial (0.67 vs 0.75 for mRS–NIHSS and 0.58 vs 0.67 for BI–NIHSS). In addition to random sample variability, evaluator reliability, a different data imputation strategy, or a more heterogeneous patient population (ie, wider eligibility criteria, longer treatment delay, and so on) can also explain this result. In any case, the benefit of combining the univariate estimators in a single multivariate one (as the Generalized Estimating Equations [GEE] does) is higher for lower correlations."[16]

Uniform Unit

Our results show that the scales are not linear (Figure 2), with some clustering for the healthier scale values that may suggest smaller information loss when pooling mRS of 0 with mRS of 1 than when pooling from mRS of 2 to 6. However, the assumption of a uniform unit has a limited impact in correlation measures. In addition to the standard assumptions of independent and identical randomly distributed variables, the Pearson correlation coefficient also assumes a normal bivariate distribution and a linear relationship among the variables, but the Spearman correlation coefficient only assumes a linear relationship among the ranks calculated considering ties.

Any mathematical model is a simplification of nature, but our data agree much more with the oversimplification of unit measure (ie, that the difference between mRS=1 and mRS=2 equals the difference between mRS=4 and mRS=5) than the oversimplification of dichotomization (ie, that patients with mRS=2 are equal to patients with mRS=5). So, if in the quest for simplification researchers agree to pay the smaller penalty of unit measure, then parametric statistics may be used to quantify treatment effect size, probably leading to more powerful and readable analysis.

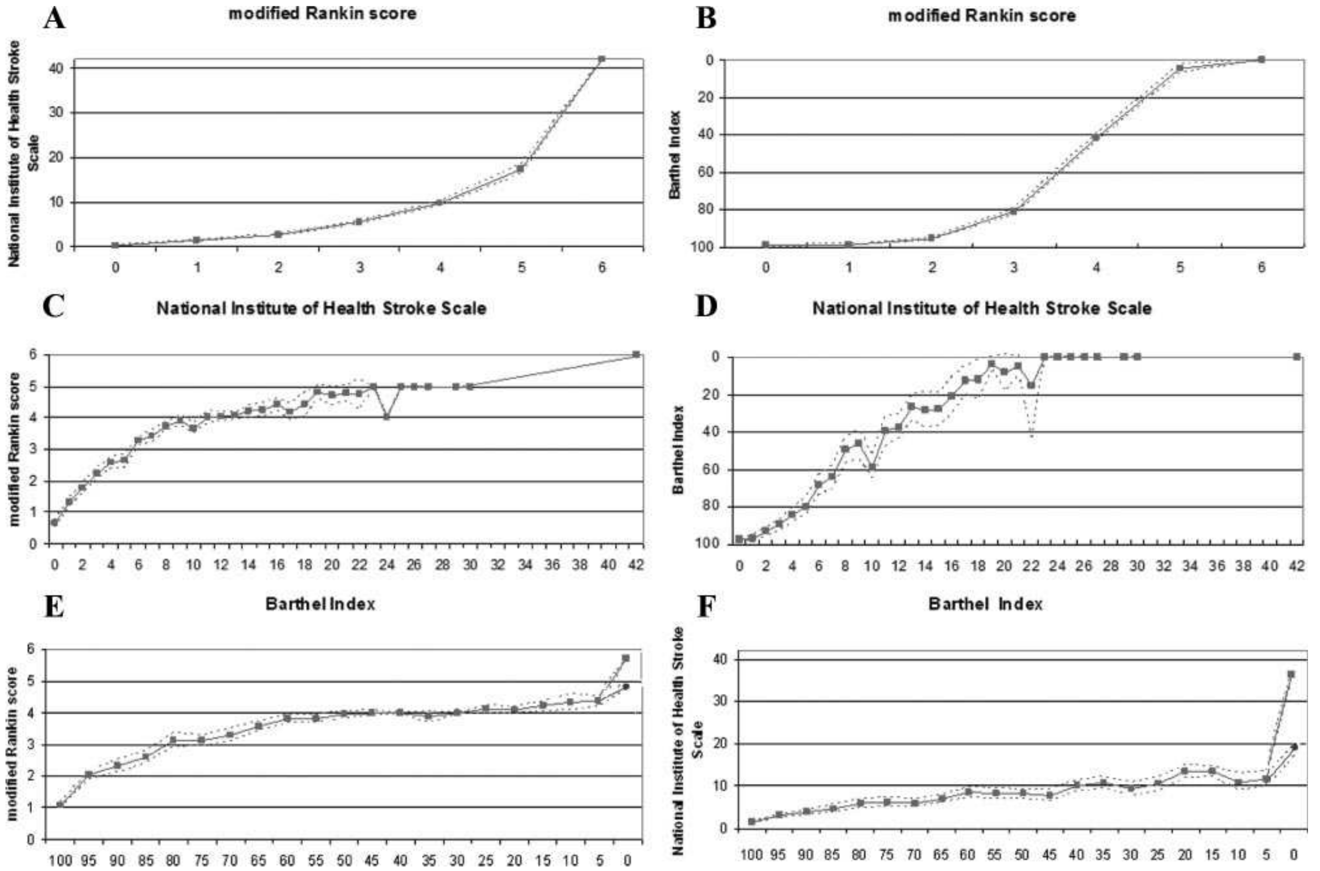

Figure 2. Progression of means over the scale values. Solid (broken) lines represent the means (95% CI) of the scale plotted on the vertical axis depending on the values of the scale on the horizontal axis. Light gray lines represent the results for all patients, and the dark gray lines represent the survivors subset (89 patients with the lowest BI value). The mRS (A, B) shows no inconsistencies in the progression of means from the other 2 scales, although this "progression" is higher from 3 to 6 than 0 to 3. The pattern of NIHSS also shows a progression in the means from mRS (C) and BI (D), with minor deviations in less-observed categories (eg, 10). BI shows a natural ordering when analyzing mRS values (E) but some oscillations or minor inconsistencies in the less-observed categories (15–40) of NIHSS (F).

The assumption of uniform unit has been widely used with those neurological scales. For example, scale means have been provided in the description of one-fifth of stroke clinical trials.[17] In a review of their clinical interpretation,[18] the NIHSS and BI reliability are quantified by the unit-measure intraclass correlation coefficient. In a more sophisticated way, the VISTA[19] and the SAINT investigators[20] used baseline NIHSS as a quantitative predictor variable of outcome in 2 similar logistic regression. Joint analysis of the 2 SAINT trials estimated that the odds of a favorable outcome is expected to be systematically multiplied 0.814-times for each 1-point increment on the NIHSS baseline, either in an increment from 1 to 2 or in an increment from 21 to 22.

Common vs Unique Information

The scales have traditionally been used to assess the different dimensions of health recognized by the World Health Organization: physical deficits (NIHSS), functional disability (BI and mRS), and social handicap (mRS). However, they are highly correlated, and all of them can also be conceived as measuring stroke recovery. Our results show that 90% of the value differences on each of the 3 scales reflect variation along a single common stroke recovery dimension. Because the remaining variability cannot be attributed to specific information indexed uniquely by an individual scale, controversial clinical trials results among scales are not expected. Furthermore, our results support interpreting the 3 scales as repetitive measures of the same parameter: patient recovery. As an analogy, cardiologists average 3 repetitions of independent determinations of blood pressure values to improve reliability. For example, the CONSORT statement,[21] in item 6b, requires specifying "any methods used to enhance the quality of measurements (eg, multiple observations)." Our estimated value of 90.7% of common variance can be directly compared with the intraclass correlation coefficient obtained on blood pressure data. Montes et al[22] estimated the concordance between a family doctor and a semiautomatic device to be 0.84 (95% CI, 0.78–0.90). In other settings, such as health-related quality-of-life measures, usual intraclass correlation coefficient values can range[23] between 0.55 and 0.79. The assumption of a common correlation without specific contributions from any scale has been used in simulations to explain the extra power provided by the global statistic.[16]

Readability and Transparency

In choosing a statistical method, in addition to the medical question and its statistical properties, its readability should

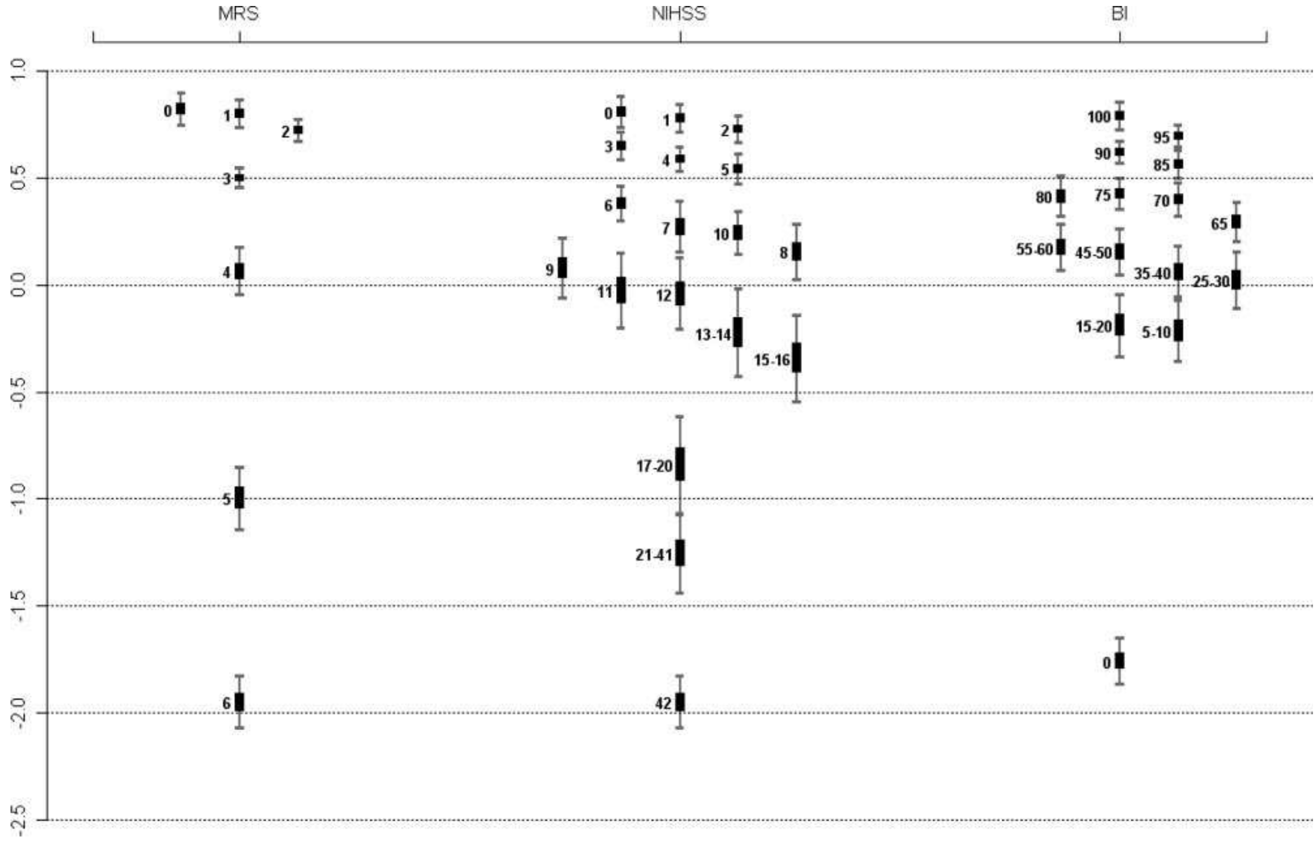

Figure 3. MCA: first dimension.# MCA first computes the χ^2 distance between each pair of categories by comparing their profiles (that is, the percentages of all remaining values). Second, it looks for the smallest dimension that better-reproduces those distances. The χ^2 distance between categories is a measure of dissimilarity between their profiles. #Other dimensions were negligible. The vertical axis represents the first new dimension that could be interpreted as patient recovery. It has been rescaled to have a mean of 0 and standard deviation of 1. The bar length is the 95% CI and the center square is the 50% CI. The natural rank is highly preserved, but consecutive scale values related to healthier status (top) are plotted closer than those related to worse health conditions (bottom). As an example of correspondences among values from different scales, mRS value 4 corresponds to values 9 and 35–40 on NIHSS and BI scales.

also be considered to achieve the desired transparency to enhance physician advice and patient choice. Our results suggest that in the setting of clinical trials, the 3 outcome scales can be seen as a repeated measure of the same clinical feature: patient recovery. Provided that this assumption is not violated in the actual clinical trial data (ie, that different scales do not show true different estimations of effect size), it facilitates the interpretation. For example, the GEE odds ratio estimates a single treatment effect on "patient recovery" with "minimal or no disability."

To check the transparency and readability of the different alternative statistical analyses, random surveys of the general population of potential stroke patients, of patients who have had a stroke, and of physicians who treat stroke patients should be conducted to ask the various groups what type of treatment effect size measure they find most interpretable.

Table 2. Correlations Between Pairs of Scales

Correlations	Unit Measure (Pearson)	Ranks (Spearman)	Dichotomized† (Phi)
mRS–NIHSS	0.84	0.91	0.68
mRS–BI	0.90	0.94	0.67
NIHSS–BI	0.84	0.89	0.58

†mRS≤1, NIHSS≤1, BI≥95.

Rank correlations are slightly higher than unit measure correlations but much higher than their dichotomized version.

Limitations

External validity is maybe the most important limitation. To reach its main objective (to get an unbiased estimate of the treatment effect size), clinical trials rely on the random allocation of treatments to patients fulfilling the eligibility criteria. Because they are not based on random samples of patients, the extrapolation of their results needs the additional assumption that the treatment effect is still the same on the external, more general, new target population. Because our results rely on clinical trial data, we do not pretend that our conclusions apply to other conditions. Specifically, in contrast to our results, it is highly plausible that the 3 outcome scales should have a completely different meaning in settings other than stroke trials. So, we do not propose that the 3 scales should be interpreted as just measuring only 1 patient characteristic in the general population. Furthermore, the 3 outcome scales we used were recorded 12 weeks after stroke by evaluators blinded to treatment assignment but aware of the other scale values, violating a key design feature for

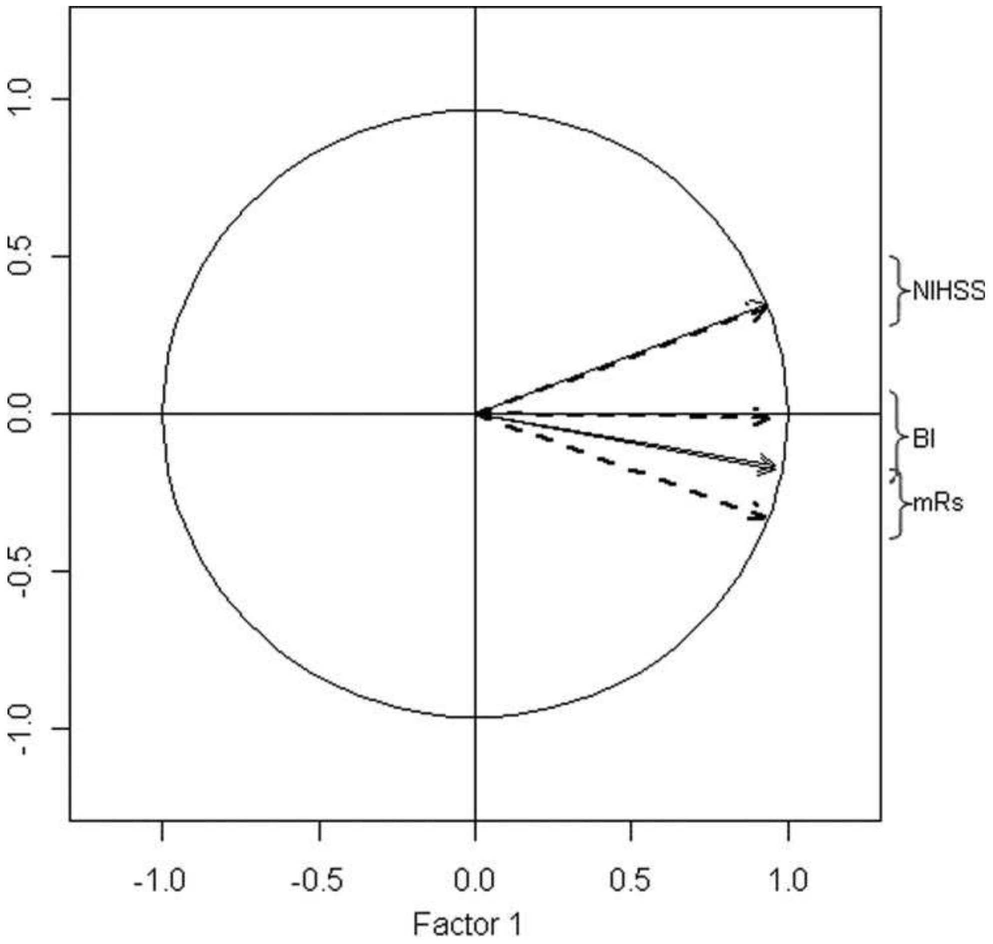

Figure 4. Representation of the scales in the first 2 PCA dimensions. NIHSS, mRS, and inverted BI all extend in the same direction along the first component, which can be regarded as patient recovery and accounts for 90.7% of the original variability. The exclusion of patients who died resulted in a very similar pattern (dotted lines), with minor changes only in the second dimension, which confronts the neurological information from NIHSS with the more general information from BI and mRS, but it accounts only for 6.0% of the overall information provided by the 3 scales. If the 3 variables had been fully independent, each new dimension would have retained one-third of the original information provided by the 3 variables. In the other extreme situation when the 3 variables are perfectly correlated, only 1 new dimension would retain 100% of the information. In such a situation, the 3 variables will order the individuals in exactly the same manner, with exactly the same relative distance between consecutive cases.

measuring concordance and diagnostic performance.[24] To apply our results to other clinical settings, adequately collected data should be analyzed.

In all of our analyses we have assumed that the equivalence premise holds for all the scale values, that is, that patients sharing the same scale value are identical among each other but different from any other patient, at least in the measured characteristic. This premise is intrinsic to any measurement, irrespective of other measure properties such as rank or unit measure. However, this assumption is highly questionable in the case of the NIHSS scale. For example, a NIHSS of 3 could represent a major deficit in 1 area or a collection of 3 very minor deficits. Our analysis does not provide any evidence about the correspondence between these 2 different kinds of NIHSS of 3, but maybe a future MCA performed on the different components of this scale could add some new information. Taking into account the idiosyncrasies of stroke patients, it should also be studied to see if they can help in assessing the equivalence between those different NIHSS=3 values.

In the interest of parsimony, our reported analyses have not considered 3 variables that may potentially affect the distribution of the outcome scales: study, allocated treatment, and Treatment delay. When those 3 variables have been included in our analyses, the results have been almost identical.

Implications for Phase III Clinical Trials

Our results point out that the frequent dichotomization of ordinal scales merges scale values that correspond to different patient status, suggesting the loss of power previously raised by several authors, either by using theoretical formulas of sample size[3] or by reanalyzing previous stroke clinical trials.[25] However, Savitz et al[26] found that the Cochran-Mantel-Haenszel ordinal shift test failed to outperform the dichotomized Rankin outcome of 0 to 2 in the post hoc analysis of the NINDS and ECASS trials.

We have shown that the assumption of unit measurement has a modest impact on results, allowing parametric statistics, such as means and standard deviations, to display patient evolution on any of those scales and using design features to improve statistical efficiency (such as averaging repeated determination of the same scale to control either inter-rater or intrapatient variability). Researchers can avoid the assumption of a constant scale unit by using an ordinal analysis, which retains much of the power and is less sensible to extreme values. In ordinal analyses, treatment effect estimates can take the form of odds ratios[3] or the Mann-Whitney-Wilcoxon c-statistic; that is, the proportion of patients having better outcomes with the study intervention than with the control,[27] which is a single and interpretable measure of stroke treatment effect.[5,28]

Our results also provide a foundation for merging estimates from different scales. Because the scales do not statistically measure unique dimensions, results in different directions should not be encountered, and a single combined estimate will increase study power. The O'Brien test may combine quantitative variables from several outcomes.[29] If the unit-measure assumption does not hold, the adjusted O'Brien rank–sum-type test can be applied to the ordinal outcome scales.[30]

By interpreting the scales as repetitions of the measurement of patient recovery, we pose the question of "adding patients or adding measures,"[31] because extra information can also be provided by other scales, such as the Glasgow Outcome Scale.

Conclusion

Our results suggest that the common practice of using a single dichotomized scale in primary end point analysis is not optimal, and it may result in reduced study power, making trials needlessly larger or increasing the chance that studies will fail to detect the benefit of a genuinely efficacious agent. Our results indicate that more information can be used in the statistical analysis of stroke trials if dichotomization is avoided and several scales are joined to estimate a single efficacy measure.

However, we have not studied the power of the suggested methods under different alternative treatment effects. To definitively boost the chances of improving stroke treatment, statistical simulation studies should be performed to consider the type of treatment effect of the new compound over the comparator.

Acknowledgments

The authors thank Matthew Elmore for improving the English style, J.L. Carrasco for raising the equivalence between the intraclass correlation coefficient reliability index and the first PCA dimension, and Belchin Adriyanov and Jordi Cortés for computing assistance. The authors also thank Ferrer International for providing the Citicoline data and funding this study. This article has been greatly improved by the reviewers' comments. All authors have contributed and approved the final manuscript. E.C., J.A.G., and J.J.S. had the original idea. F.M. and C.C. performed and validated the analysis. R.R., J.L.S., and A.D. contributed to data interpretation and analysis. E.C. is the guarantor.

Sources of Funding

Funding received from Ferrer International.

Disclosures

J.J.S. is an employee of Ferrer International. E.C., J.L.S, and A.D. are scientific consultants for and have received lecture honoraria from Ferrer International.

References

1. Donnan GA, Fisher M, Macleod M, Davis SM. *Stroke Lancet*. 2008;371: 1612–1623.
2. Kidwell CS, Liebeskind DS, Starkman S, Saver JL. Trends in acute ischemic stroke trials through the 20th century. *Stroke*. 2001;32: 1349–1359.
3. Bolland K, Sooriyarachchi MR, Whitehead J. Sample size review in a head injury trial with ordered categorical responses. *Stat Med*. 1998;17: 2835–2847.
4. Wardlaw JM, Sandercock PAG, Warlow CP, Lindley RI. Trials of thrombolysis in acute ischemic stroke. Does the choice of primary outcome measure really matter? *Stroke*. 2000;31:1133–1135.
5. 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.
6. Young FB, Lees KR, Weir CJ; for the GAIN trial Steering Committee and Investigators. Strengthening acute stroke trials though optimal use of disability end points. *Stroke*. 2003;34:2676–2680.
7. 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.
8. Tilley BC, Marler J, Geller NL, Lu M, Legler J, Brott T, Lyden P, Grotta J; for the National Institute of Neurological disorders and stroke t-PA stroke trial study group. Use of a global test for multiple outcomes in stroke trials with application to the National Institute of Neurological disorders and stroke t-PA stroke trial. *Stroke*. 1996;27: 2136–2142.
9. Lees KR, Zivin JA, Askwood T, Dávalos A, Davis SM, Diener HC, Grott J, Lyden P, Shuaib A, Hardemark HG, Wasiewski WW. NXY-059 for acute ischemic stroke. *N Engl J Med*. 2006;354:588–600.
10. Fisher M, Hanley DF, Howard G, Jauch EC, Warach S; for the STAIR group. Recommendations from the STAIR V meeting on acute stroke trials, technology and outcomes. *Stroke*. 2007;38:245–248.
11. Davalos A, Castillo J, Alvarez-Sabin J, Secades JJ, Mercadal J, Lopez S, Cobo E, Warach S, Sherman D, Clark WM, Lozano R. Oral citicoline in acute ischemic stroke:an individual patient data pooling analysis of clinical trials. *Stroke*. 2002;33:2850–2857.
12. Lefevre-Colau MM, Fayad F, Rannou F, Fermanina J, Coriat F, Mace Y, Revel M, Poiradeau S. Frequency and interrelations of risk factors for chronic low back pain in a primary care setting. *Plos One* 2009;4:e4874.
13. Briand S, Beresniak A, Nguyen T, Yonli T, Duru G, Kambire Ch, Perea W; for the Yellow fever risk assessment group (YF-RAG). PLOS neglected tropical diseases 2009;3:e483.
14. Islami F, Kamangar F, Nasrollahzadeh D, Aghcheli K, Sotoudeh M, Abedi-Ardekani B, Merat S, Naseri-Moghaddam S, Semnani S, Sepehr A, Wakefield J, Moller H, Abnet CC, Dawsey SM, Boffetta P, Malekzadeh R. Socio-economic status and oesophageal cancer: results from a population-based case-control study in a high-risk area. *Int J Epidemiol*. 2009;38:978–988.
15. Committee for proprietary medicinal products of the European Agency for the evaluation of Medicinal Products. Points to consider on clinical investigation of Medicinal products for the treatment of acute stroke. 2001;CPMP/EWP/560/98.
16. Bolland K, Whitehead J, Cobo E, Secades J. Evaluation of a sequential global test of improved recovery following stroke as applied to the ICTUS trial of Citicoline. *Pharm Stat*. 2009;8:136–149.
17. Song F, Jerosch-Herold C, Holland R, Drachier M, Mares K, Harvey I. Statistical methods for analysing Barthel scores in trials of poststroke interventions: a review and computer simulations. *Clin Rehabil*. 2006;20: 347–356.
18. Kasner SE. Clinical interpretation and use of stroke scales. *Lancet Neurol*. 2006;5:603–612.
19. König IR, Ziegler A, Bluhmki E, Hacke W, Bath PM, Sacco RL, Diener HC, Weimar C; Virtual International Stroke Trials Archive (VISTA) Investigators. Predicting long-term outcome after acute ischemic stroke: a simple index works in patients from controlled clinical trials. *Stroke*. 2008;39:1665–1666.
20. Diener HC, Lees KR, Lyden P, Grotta J, Davalos A, Davis SM, Shuaib A, Ashwood T, Wasiewski W, Alderfer V, Hårdemark HG, Rodichok L; for the SAINT I and II Investigators Stroke. NXY-059 for the treatment of acute stroke: pooled analysis of the SAINT I and II Trials. *Stroke*. 2008;39:1751–1758.
21. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, Gøtzsche PC, Lang T; for the CONSORT Group. The revised CONSORT

statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med*. 2001;134:663–694.

22. Montes G, Fernandez JA, Prada A, Polonio R, Rodriguez D, Perula LA. Fiabilidad en la medición de la presión arterial:paciente frente a profesionales de atención primaria. *Atención Primaria*. 2000;25: 27–35.
23. Tebe C, Berra S, Herdman M, Aymerich M, Alonso J, Rajmil L. Fiabilidad y validez de la versión española del KIDSCREEN-52 para población infantil y adolescente. *Med Clin (Barc)*. 2008;130: 650–654.
24. Bossuyt P, Reitsma J, Bruns D, Gatsonis C, Glasziou P, Irwig L, Lijmer J, Moher D, Rennie D, De Vet H. Towards complete and accurate reporting of studies of diagnostic accuracy:The STARD initiative. *Clin Chem*. 2003;49:1–6.
25. The Optimising Analysis of Stroke Trials (OAST) Collaboration. Calculation of sample size for stroke trials assessing functional outcome: comparison of binary and ordinal approaches. *Int J Stroke*. 2008;3: 78–84.
26. 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.
27. Newcombe RG. Confidence intervals for an effect size measure based on the Mann–Whitney statistic. Part 1: General issues and tail-area-based methods. *Statist Med*. 2006;25:543–557.
28. Acion L, Peterson JJ, Temple S, Arndt S. Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects. *Statist Med*. 2006;25:591–602.
29. O'Brien PC. Procedures for comparing samples with multiple endpoints. *Biometrics*. 1984;40:1079–1087.
30. Huang P, Tilley BC, Woolson RF, Lipsitz S. Adjusting O'Brien's test to control type I error for the generalized nonparametric Behrens-Fisher problem. *Biometrics*. 2005;61:532–539.
31. Ahn C, Jung S. Efficiency of general estimating equations estimators of slopes in repeated measurements: Adding subjects or adding measurements? *Drug Inform J*. 2003;37:309–316.