Open Access Library Journal

Stroke Assessment Scales: The Dilemma Validity or Reliability

Marco Aurelio Gralha de Caneda

Neurology Department of GHC, Hospital Nossa Senhora da Conceição, Porto Alegre, Brazil
Email: mcaneda@terra.com.br

Received 26 July 2014; revised 20 September 2014; accepted 24 October 2014

Abstract

The author discusses the conceptual bases related to general and clinimetric properties of stroke scales, relating them to its application in research and clinical practice. This paper briefly presents the definitions of reliability and validity and its usage in clinical studies, and didactically demonstrates statistical approaches for obtaining agreement and correlation coefficients. Finally, it proposed to criticize the most common bias in the literature on the use of these concepts and forwards suggestions for its appropriate guidance to studies formulation.

Keywords

Reliability, Validity, Stroke Scales, Neurological Scales

Subject Areas: **Epidemiology, Evidence Based Medicine, Internal Medicine, Neurology**

1. Introduction

A stroke assessment scale is a simplified and organized selection of items found in routine neurological examination or daily activities of any individual. It is not just used as a selection criterion for certain clinical therapeutics or in the assessment of its efficacy, but also may be valuable and advantageous instruments to anticipate the outcome as quickly and accurately as possible in patients admitted for specifics treatments [1]-[3]. This facilitates the preparations of relatives about the individual's future needs after the discharge, because there's great difference among these patients, ranging from those who will need permanent care, with total dependence, or some that will develop mild to moderate repercussions, until another one who will resume their normal activities [4]-[7]. These scales can specifically to approach loss or abnormality of a psychological function, physiological or anatomical damage (*impairment scales*); the restriction and deprivation resulting in the ability to perform a task within the standards considered normal or activity limitations scales (*previously disability scales*); or the disadvantage that will harm the individual in the social context, or societal participation scales (*previously han-*

How to cite this paper: de Caneda, M.A.G. (2014) Stroke Assessment Scales: The Dilemma Validity or Reliability. *Open Access Library Journal*, **1**: e1067. http://dx.doi.org/10.4236/oalib.1101067

dicap scales) [8] [9].

2. General Properties of Scales

The clinical medicine is based on concepts and methodological bases that are directly related to biomedicine. In these properties is included the so-called *clinimetric criteria*, which are commonly used both in clinical practice as in research in general neurology [10]. Watching these *clinimetric criteria*, a strokescale should be characterized by include specificity, prognostic value, sensitivity, validity and reliability, each one with values clinically acceptable [11]-[13]. Thereunto, these scales, shall: 1) qualifying and quantifying the neurological signs and symptoms to detect abnormalities in any phase of disease, it's very important to include a selection of items based on their clinical relevance, to preserve the specificity; 2) have prognostic value in acute phase of disease; 3) have satisfactory reproducibility of repeat scoring; 4) be validated, in field of construction validity; at least, 5) have responsiveness, with capacity to fluctuate to identify clinical changes over time; 6) easy to use to patient and tester [9]-[12].

Yet, in its issues constituents, the scales shall: 1) avoid the incorporation of signs of dubious or inconclusive functional significance; 2) precisely define the maneuvers to obtain certain results; 3) avoid the inclusion of very complex or difficulties items to execute; 4) to graduate in a ranking its items in according to the parameters of the neurological examination or task difficulty and complexity; 5) into their properties and its purpose, shall be as embracing as possible, for example, many scales used in several pathologies, as in stroke, with different etiology, pathogenesis and repercussions doesn't evaluate the cognitive state, and this is an essential indicative of suitable functioning of central nervous system, but unfortunately, is often excluded from routine assessment [8] [9] [14].

3. Clinimetrics Properties of Scales

Specificity is critical to the relevance of a stroke scale. Any element just occasionally found do not to be included, so the selection of component items should be based on its diagnostic power, it means, with reasonable frequency, and on its prognostic capacity. For example, studies suggest that the level of consciousness, muscle strength, failures in the visual fields and some aspects of language are better prognostic factors in the neurological examination [12] [14].

Validity is an important feature of a scale, but difficult to assess. It can refer to the ability to determine the clinical status of an individual or predict his conditions (criterion validity); or to the property to include all the dimensions that are relevant in reference to what is being measured (content validity); or even if is possible the correlation with other scales or measuring instruments (concurrent validity) [15]. Note that there's a common bias in current literature, where some studies are classified as about validity of a method, while, in fact, are related to reliability of this instrument. Sometimes, even inadequate statistics bases, such agreement coefficients, for example, are used in the analysis of studies that intend to measure concurrent validity of certain method, whereas the reproducibility of results, despite being essential, doesn't check validity to an instrument.

Reliability is defined as the quality of a particular method to reproduce its results on repeated applications with the least possible variability [12]-[16]. The reliability cannot be conceived as a property that a method has or not, every instrument shows some degree of reliability when applied to certain populations and under certain conditions, however should respect a plausibility and clinical adequacy [2] [3] [17].

4. Ascertainment of Scales Reliability

The classification of the degree of reliability of a scale can be obtained by measuring the reproducibility of the scores found by different observers, *the inter-rater agreement*, for just an observer at different times, *the intra-rater agreement*, and also by internal consistency (homogeneity), which is the inter-correlation among the variables, or components items, of a scale [12]-[18].

In assessing the inter-rater agreement, is evaluated, ultimately, two components of possible inaccuracy of an instrument: a bias that is reflected by the differences between the marginal distributions of the variable responses of each observer; and another one, that reflects how the observers classified individuals in the same category of measurement scale [19]. The inaccuracy of inter-rater agreement possibly comes from specific characteristics of a scale, for example, of some degree of subjectivity of items components; of assessed subjects, which

may have clinical changes in the acute, subacute or chronic phase of a disease; or from examiners who may disagree by absence or insufficient prior training.

The qualitative representation of inter-rater agreement can be measured using different methods. The percentage of agreement would be the simplest way to achieve this result. There are two aspects of weakness in this kind of statistical analysis.Firstly, it doesn't qualify the examiners' agreement, which can agree in elements that don't correspond to reality, in other words, they can agree in a wrong variable. Possibly this misconception do not affect reliability, but has huge consequences on the accuracy of the method. In addition, an agreement by chance can be expected in any method and the percentage doesn't include this observation under any circumstances. Thus, percentage of agreement may be satisfactory for binary data, but do not reflect the magnitude of disagreement in ordinal or nominal data such as of stroke scales [12] [20] [21].

Another way to get the measure of agreement would be through *correlation coefficients*. These coefficients express a trend between two sets of categorical observations. It's of paramount importance to realize that measures of degree of correlation are not measures of agreement. The correlation coefficients are inappropriate for measuring concordance of results. Although it be sometimes used, correlation coefficients are inappropriate to approach of reliability. Firstly, the correlation coefficients are measures of linear association between two variables, which is not the same as a measure of agreement. Furthermore, one can obtain high degrees of correlation with poor real clinical agreement. It doesn't provide values that can be compared, not expressing a true agreement. May present an excellent result, even when there's no agreement, since that there's a systematic variation in the results obtained by the examiners. So, they are considered too much "liberal" to measure the agreement, often overestimating it. Among the most current used are the Pearson coefficient and the coefficient of Spearman [11] [17] [20].

Therefore, the inter-rater concordance is best expressed by *coefficients of agreement*, which provides the proportions of agreement that can be expected to occur by chance or by real agreement in the results obtained by different examiners [12] [14].

5. Calculation of Inter-Rater Agreement

One of the possible mathematical approach to agreement is called *the concordance kappa statistic* (k), that can be interpreted as a proportional correction to the concordance by chance, and has been widely used in several studies of neurological disease, focusing on the neurological examination [9] [21], diagnosis [1] [22]-[24] or in therapeutic efficacy evaluation [1]-[3]. It can be expressed mathematically by the following formula:

$$k = \frac{P_o - P_e}{1.0 - P_e}$$

where: "P_o" is the proportion of actual agreements and "P_e" is the proportion of agreements expected by chance [17].

A statistical weakness of k coefficient is that it doesn't provide the magnitude of the examiners disagree. Every disagreement is treated equally. So, is convenient, where there're more than two classificatory categories, to weight these degrees of variance, in order to more accurately measure the amounts of it between examiners. This is the *weighted kappa coefficient* $(k\rho)$ [16] [17] [20]. This coefficient can be mathematically expressed by:

$$k\rho = \frac{P_o\rho - P_e\rho}{1 - P_e\rho}$$

where: "$P_o\rho$" is the proportion of weighted actual agreements and "$P_e\rho$" is the proportion of weighted agreements expected by chance [20].

The most commonly problem cited in relation to the kappa statistic is that its value is very dependent on the proportion of results allocated in each category. The coefficient is directly dependent on both the number of variables for classification as the number of subjects in sample classified in each category [20]. That's quite evident when it has only two of it. The reason for this effect is that the frequencies generated by the chance vary widely and the expected value in each of these categories, by chance, definitely changes the value of k, what induce an error called *prevalence bias*.

The consequence of this property is that can be generated a confusion bias in the comparison of k values from different studies, when the prevalence of each category are not similar between these. Computer programs, provide an extra coefficient, when in such situations is the k adjusted for prevalence bias (PABAK) [20]. Despite the criticism about the kappa statistic, this is one of the most correct approach to problems in agreement [12] [14] [16].

Another statistical instrument to approach inter-rater agreement, which has been used in recent studies of reliability, is the *Intraclass Correlation Coefficient* (*ICC*). This coefficient expresses the proportion of total variance that is due to the true variance between patients. Include the variability found between examiners, between individuals and residual error.It's increased if the variance caused by differences among patients is high [19] [25].The ICC can be expressed mathematically by the following formula:

$$\mathrm{ICC} = \frac{\sigma s^2}{\sigma s^2 + \sigma o^2 + \sigma e^2}$$

where: σs^2 is the variance between individuals, σo^2 is the variance between examiners and σe^2 and is the variance of the error.

The interpretation of *ICC* is performed similarly to the k coefficient. The values of this coefficient of agreement can range from −1 to +1, indicating, respectively, to total disagreement and total agreement. The agreement by chance produces a coefficient equal to 0 [26]. A value equal or greater than 0.60 would be evidence of a clinically acceptable reliability, with satisfactory, good or excellent, inter-rater agreement [12] [16] [19] [20]. Intermediate values between these extremes are classified as according to **Table 1**.

Table 1. Classification of inter-rater agreement according to coefficients $\kappa, \kappa\rho$ and ICC values.

Coefficient Value	Classification
0 - 0.20	Insufficient
0.21 - 0.40	Low
0.41 - 0.60	Moderate
0.61 - 0.80	Good
0.81 - 1.0	Excellent

6. Conclusions

A stroke scale is useful if there's minimal variability in its results and these are comparable to other scales with same purpose or can be correlate to other assessment tools. Reliability is clinically appropriate according to the degree of agreement intra or inter-examiner; and there's homogeneity and consistency between its component items. The inter-rater agreement could be calculated using a simple percentage. This, however, doesn't consider the possible agreement by chance, what is therefore unsuitable for use in clinical trials. Also inter-rater agreement can be obtained by correlation coefficients that have a statistic tendency to overestimate the concordance without expressing your true level. Ultimately it can be calculated by agreement coefficients, which seem to be the best means of assessment. The definitive validation of scales remains difficult to obtain in the absence of a gold-standard assessment instruments for the diagnostic and, especially, prognosis, of neurological disorders, precluding assessment to criterion validity. However, construct validity can be assessed by comparing results of similar scales. In this case, the statistical analysis using the agreement coefficient is inappropriate, tending the results to extremely poor values. On the other hand, the correlation coefficients are very accurate instruments for use with this purpose.

The use of reliable, valid, sensitive, and specific scales is fundamental in stroke, as to evaluate the effectiveness of new therapies, some distant away from us a few years ago, but increasingly present today, as to describe the consequences of neurological damage and clinical changes responsible for handicaps and disabilities that will represent a better or worse quality of life for patients and their relatives.

Disclosure

The author reports no conflicts of interest in this work. the consequences of neurological damage and clinical

changes responsible for handicaps and disabilities that will represent a better or worse quality of life for patients and their relatives.

References

[1] Kasner, S. (2006) Clinical Interpretation and Use of Stroke Scales. *The Lancet Neurology*, **5**, 603-612.

[2] Lai, S.M., Duncan, P.W. and Keighley, J. (1998) Prediction of Functional Outcome after Stroke. *Stroke*, **29**, 1838-1842. http://dx.doi.org/10.1161/01.STR.29.9.1838

[3] de Caneda, M.A.G., Fernandes, J.G., Almeida, A.G. and Mugnol, F.E. (2006) Reliability of Neurological Assessment Scales in Patients with Stroke. *Arquivos de Neuropsiquiatria*, **64**, 690-697.

[4] Meyer, B.C., Hemmen, T.M., Jackson, C.M. and Lyden, P.D. (2002) Modified NIHSS for Use in Stroke Clinical Trials. Prospective Reliability and Validity. *Stroke*, **33**, 1261-1266. http://dx.doi.org/10.1161/01.STR.0000015625.87603.A7

[5] Haan, D.R., Horn, J. and Limburg, M. (1993) A Comparison of Five Stroke Scales with Measures of Disability, Handicap and Quality of Life. *Stroke*, **24**, 1178-1181. http://dx.doi.org/10.1161/01.STR.24.8.1178

[6] Kay, R., Wong, K.S. and Perez, G. (1997) Dichotomizing Stroke Outcomes Based on Self-reported Dependency. *Neurology*, **49**, 1694-1696. http://dx.doi.org/10.1212/WNL.49.6.1694

[7] Wilson, J.T., Hareendran, A. and Grant, M. (2002) Improving the Assessment of Outcome in Stroke. *Stroke*, **33**, 2243-2246. http://dx.doi.org/10.1161/01.STR.0000027437.22450.BD

[8] Orgogozo, J.M. (1994) The Concepts of Impairment, Disability and Handicap. *Cerebrovascular Disease*, **4**, 2-6. http://dx.doi.org/10.1159/000108531

[9] Orgogozo, J.M. (1998) Advantage and Disadvantage of Neurological Scales. *Cerebrovascular Disease*, **8**, 2-7. http://dx.doi.org/10.1159/000047505

[10] Asplund, K. (1987) Clinimetricsin Stroke Research. *Stroke*, **18**, 528-530.

[11] Duffy, L., Gajree, S., Langhorne, P., Stott, D. and Quinn, T.J. (2013) Reliability (Inter-Rater Agreement) of the Barthel Index for Assessment of Stroke Survivors. Systematic Review and Meta-Analysis. *Stroke*, **44**, 462-468. http://dx.doi.org/10.1161/STROKEAHA.112.678615

[12] Hantson, L. and De Keyser, J. (1994) Neurologic Scales in the Assessment of Cerebral Infarction. *Cerebrovascular Diseases*, **4**, 7-14. http://dx.doi.org/10.1159/000108532

[13] Banks, J.L. and Marotta, C.A. (2007) Outcomes Validity and Reliability of the Modified Rankin Scale: Implications for Stroke Clinical Trials. *Stroke*, **38**, 1091-1096. http://dx.doi.org/10.1161/01.STR.0000258355.23810.c6

[14] Côté, R., Batista, R.N., Wolfson, C.M. and Hachinski, V. (1988) Stroke Assessment Scales: Guidelines for Development, Validation and Reliability Assessment. *Canadian Journal of Neurological Sciences*, **15**, 261-265.

[15] Quinn, T.J., Dawson, J., Walters, M.R. and Lees, K.R. (2009) Reliability of the Rankin Scale. A Systematic Review. *Stroke*, **40**, 3393-3395. http://dx.doi.org/10.1161/STROKEAHA.109.557256

[16] Lyden, P.D. and Lau, G.T. (1991) A Critical Appraisal of Stroke Evaluation and Rating Scales. *Stroke*, **22**, 1345-1352. http://dx.doi.org/10.1161/01.STR.22.11.1345

[17] Streiner, D.L. and Norman, G.R. (1995) Health Measurements Scales: A Practical Guide to Their Development and Use. 2ª Edition, Oxford University Press, Oxford, 104-127.

[18] Koran, L.M. (1975) The Reliability of Clinical Methods, Data and Judgments. *The New England Journal of Medicine*, **293**, 642-646. http://dx.doi.org/10.1056/NEJM197509252931307

[19] Everitt, B.S. (1989) Measurements in Medicine and Statistical Methods for Medical Investigations. Oxford University Press, London, 16-28.

[20] Altmann, D.G. (1990) Practical Statistics for Medical Research. 1st Edition, CRC Press, London, 396-409.

[21] Koran, L.M. (1975) The Reliability of Clinical Method, Data and Judgments II. *The New England Journal of Medicine*, **293**, 695-701. http://dx.doi.org/10.1056/NEJM197510022931405

[22] Berger, K., Kase, C.S. and Buring, J.E. (1996) Interobserver Agreement in the Classification of Stroke in the Physician's Health Study. *Stroke*, **27**, 238-242. http://dx.doi.org/10.1161/01.STR.27.2.238

[23] Johnson, C.J., Kittner, S.J., McCarter, R.J., Sloan, M.A., Stern, B.J., Buchholz, D. and Price, T.R. (1995) Interrater Reliability of an Etiologic Classification of Ischemic Stroke. *Stroke*, **26**, 46-51. http://dx.doi.org/10.1161/01.STR.26.1.46

[24] Shinar, D., Gross, C.R., Mohr, J.P., Caplan, L.R., Price, T.R., Wolf, P.A., *et al*. (1985) Interobserver Variability in the Assessment of Neurologic History and Examination in the Stroke Data Bank. *JAMA Neurology*, **42**, 557-565. http://dx.doi.org/10.1001/archneur.1985.04060060059010

[25] Fleiss, J.L. (1981) Statistical Methods for Rates and Proportions. 1st Edition, John Wiley & Sons, London, 218.

[26] Harrison, J.K., McArthur, K.S. and Quinn, T.L. (2013) Assessment Scales in Stroke: Clinimetric and Clinical Considerations. *Clinical Interventions in Aging*, **8**, 201-211.