Stroke Assessment Scales: Guidelines for Development, Validation, and Reliability Assessment

Robert Côté, Renaldo N. Battista, Christina M. Wolfson and Vladimir Hachinski

ABSTRACT: The validity and reliability of clinical instruments, including clinical scales, need to be determined. This paper presents guidelines for development, validation, and reliability assessment of stroke assessment scales.

RÉSUMÉ: Des échelles pour evaluer les infarctus cérébraux: des conseils pour le développement, la validité, et la précision. La validité et la fiabilité d'instruments cliniques, incluant des échelles de mesure cliniques, doivent être établies. Dans cet article, nous présentons des critères de développement, de validation, et d'établissement de la fiabilité d'échelles de mesure utilisées avec des patients atteints d'un accident cérébrovasculaire.

Can. J. Neurol. Sci. 1988; 15:261-265

Stroke ranks third among the causes of mortality in North America and is one of the leading causes of physical disability in the adult population.[1] A significant number of patients admitted to hospital with an acute stroke subsequently deteriorate.[2-5] The clinical evaluation of stroke patients by medical observers is still the most sensitive and practical means of assessing neurological function and any changes in it, but the development and use of stroke assessment scales could improve the measurement and recording of such clinical variables.

Quantitative measurement of clinical phenomena is increasingly recognized, not only as an essential part of clinical research but as an important step in improving clinical decision-making and management. The measurement of neurological deficit in stroke patients is no exception.

1. The Neurological Examination And Current Scoring Systems

Formal neurological examination is still the standard and most sensitive means of assessing neurological function. How it is performed, however, may vary from examiner to examiner and give rise to different degrees of inter-observer variability.[6] Some of the other shortcomings of repeated neurological examination, as routinely employed during the acute stroke period, are: 1) it is time-consuming and impractical, 2) its use is restricted to physicians and the exchange of valuable information with other health professionals is consequently limited, and 3) it does not lend itself easily to statistical analysis. Although most scales in current use embody attempts to achieve objectivity and standardization by quantifying and simplifying specific aspects of the neurological examination, they fall short of succeeding for one or more of the following reasons: 1) absence of a precise definition of the manoeuver performed to elicit signs; 2) incorporation of signs of doubtful functional significance, e.g. pupillary responses and changes in reflexes, 3) no or faulty weighting of the different items on a scale, e.g. since impairment of consciousness is the single most important index of cerebral dysfunction, the scale should reflect this fact; 4) inclusion of observations that are too complex or too unreliable to give consistent results, e.g. different varieties of agnosia and fine points about the sensory examination; and 5) lack of validity and reliability. The last shortcoming is the greatest and most prevalent of all; few have attempted to assess inter-observer reliability or to correlate the data recorded on the scales with outcomes or other measures of cerebral dysfunction.

2. The Need To Develop Stroke Assessment Scales

A) Use in Patient Care

1) Predicting immediate outcome and guiding decision-making for patient management Although the degree of neurological impairment correlates with outcome, such variables as age and incontinence may be significant and should be taken into consideration.[7] A prognostic score in the acute stroke period must take account of the clinical features that have the strongest value in predicting the patient's survival or level of function after the acute phase has passed. Attempts at designing such scoring systems have been made in different selected pop-

From the Department of Neurology and Neurosurgery, McGill University and the Divisions of Clinical Epidemiology and Neurology, Montreal General Hospital (Dr. Côté); the Department of Epidemiology and Biostatistics, McGill University and the Division of Clinical Epidemiology, Montreal General Hospital (Dr. Battista); the Department of Epidemiology and Biostatistics, McGill University (Dr. Wolfson); the Investigative Stroke Unit, University Hospital, London (Dr. Hachinski)

Received November 12, 1987. Accepted in final form April 5, 1988

Reprint requests to: Dr. Robert Côté, Division of Clinical Epidemiology, Montreal General Hospital, 1650 Cedar Avenue, Montreal, Quebec, Canada H3G 1A4

https://doi.org/10.1017/S0317167100027712 Published online by Cambridge University Press

ulations. Allen[8] recently proposed a simple prognostic score that predicts the functional outcome of stroke patients and is easy to use at the bedside. The features found to have good predictive value include 1) age, 2) limb paralysis, and 3) level of consciousness. The ability to determine the likelihood of survival and recovery early in the acute period of stroke and to identify patients with the best prognosis may not only affect the selection of patients for admission to certain more intensive care units but also influence the choice of initial therapy and the planning of rehabilitative programs.

2) Ongoing clinical monitoring Stroke scales could also be used by physicians and other medical personnel in different institutions to evaluate the initial level of neurological function and to monitor its subsequent evolution. Such an objective clinical indicator of neurological change could facilitate the elaboration and ongoing assessment of different management schemes in individual patients. Many patients with acute strokes deteriorate after admission to hospital for diverse reasons ranging from extension of the thrombo-embolic process to cerebral edema and systemic factors. Some are potentially treatable and a sensitive and objective scoring system might permit earlier therapeutic intervention and better patient care. One of the most successful examples of a simple reliable scale for assessing the status and evolution of brain-injured patients is the Glasgow Coma Scale.[9] Unfortunately, it is mainly useful when there is a decreased level of consciousness that is not observed in most stroke patients; in patients who are alert or only drowsy, it lacks sensitivity. A complementary neurological scale suitable for stroke patients has recently been proposed[10] and is being validated. Other simple scoring systems are also widely used in the management of stroke patients, but many of them suffer from observer variability.[11]

B) Use in Research

1) Guide for diagnostic purposes Scoring systems have potential value in diagnosis. Sandercock and colleagues[12] have recently reported on a stroke diagnostic score which enabled them to accurately differentiate between hemorrhage and infarction as the underlying cause. Although less accurate than computed tomography, this type of diagnostic score could be used, according to the authors, as a screening device in epidemiological studies. Whether it will prove to have any practical value in such studies is unclear at this time.

2) Study of the natural history of acute stroke The information derived from a standardized scale applied to a large number of patients from different centres would improve our knowledge of the natural history of stroke and permit correlation of different patient characteristics and risk factors with the clinical course during the acute period. Such numerical scoring would also permit statistical analyses that would otherwise be incomplete or crude. In general, published studies dealing with the natural history of acute stroke have not used a standardized scoring procedure to evaluate the neurological course of patients.[2,4,5]

3) Evaluation of therapeutic modalities Objective assessment of therapeutic measures in stroke patients is required whether the treatment is medical or surgical.[13] Numerous scoring systems have been used to evaluate the efficacy of certain types of therapy,[14-17] but, most have been developed in the investigators' own institutions and very few have ever been formally tested for their validity and reliability.[18]

Since therapeutic trials in stroke prevention usually require large numbers of patients, a multicentre approach is almost always necessary and an objective, valid and reproducible clinical assessment system for standardizing the outcome results in the various participating institutions is crucial. We firmly believe that only assessment systems whose reliability and validity have been clearly established should be used in any clinical evaluation of patients submitted to a therapeutic trial.

A scoring system for evaluating the benefits of certain medical or surgical interventions during the acute stroke period should preferentially include the major and common deficits that not only reflect the patient's neurological status, most faithfully, but also predict patient morbidity and mortality during the acute phase, most accurately, e.g. level of consciousness and motor function.[19-21] Although the ideal scoring system has yet to be devised, some of the few proposed systems[10,22,23] may provide a basis for the elaboration of more reliable and valid scoring instruments.

Evaluation of the effectiveness of rehabilitative measures for stroke patients after the acute phase will require a system that is heavily weighted towards functional outcome and the activities of daily living.[24]

3. Guidelines For Developing A Workable System

The initial neurological status and changes in function in acute stroke require documentation in an objective and semi-quantitative manner if the assessments of medical observers (physicians, nurses or other health professionals), at the same or different centres and at repeat examinations by the same observer are to be comparable. We have already spelled out certain guidelines that should be taken into consideration during the elaboration of a standardized neurological assessment system,[10] such as a) simple and non-ambiguous definitions for each modality tested, b) a minimum number of grades per modality to minimize inter-observer variability, c) the selection of relevant modalities that are most commonly affected in acute stroke, d) ease of use and interpretation by observers with different medical backgrounds, and e) brevity and simplicity.

Although the aims of a stroke scoring system can be multiple, different systems may be better suited to specific purposes since one system may not meet all needs.

4. Validation And Reliability Assessment

The development of a measurement scale is a step-wise process aimed at establishing its validity and reliability. The validity of a measurement instrument is the extent to which it measures what it purports to measure, while its reliability is an index of the reproducibility of the results it obtains on repeated trials. Since reliability subsumes the absence of random error (due to chance) in a series of measurements, it is a measure of precision. Validity is always a question of degree and a measurement is only valid with respect to a specified measurement object. A valid measure is free of both systematic and random error. Thus, reliability is a necessary but not sufficient condition for validity.[25,26] A third property of measurement instruments is responsiveness, i.e. their capacity to identify changes in the

https://doi.org/10.1017/S0317167100027712 Published online by Cambridge University Press

severity of specific health conditions.[27]

A) Validation

The validation of any stroke assessment scale is an ongoing process; as evidence of validity accumulates, confidence in declaring the instrument valid increases. The first step in a validation exercise is to define the measurement instrument with respect to neurological deficits found in stroke patients.

1) Content Validity

Content validity is an index of how well an instrument reflects the components of what is being measured — in this case, neurological deficit. Content validation relies heavily on judgement and different sources of judgement can be used, such as experts' opinions, review of the literature, surveys of providers and/or patients. Multiplying sources of information and consultation and putting the scale to the test of multiple iterations will increase the likelihood of properly defining the content of a stroke scale. Differential weighting of specific items in the scale may pose some difficulties. If there is no compelling clinical reason to overweight some items in a scale, give equal weights to all the items retained. If good clinical sense requires differential weighting, do it and reassess it in the predictive validation exercise described later in this paper.

Content validity was achieved during the development of the Canadian Neurological Scale (CNS).[10,18] A collection of items produced by a panel of neurologists measured what they all agreed upon as being neurological deficits. The choice of the items and their differential weights were based on the clinical judgement of the participating neurologists and the most recent information from the literature.[10,19-21,28-30] As an example, more weight was given to the item level of consciousness to reflect its prognostic value. At present, the system focuses on two main areas: mentation (level of consciousness, orientation, speech) and motor function (face, arm, leg).

2) Criterion-related Validity

A measurement is said to have criterion-related validity when the results it obtains compare closely with those achieved using a criterion or a gold standard for which validity has already been established. In the case of stroke scales, a standard neurological evaluation performed by a physician could be viewed as the gold standard. A simple correlation coefficient between the new measure and the gold standard becomes the validity coefficient of the new measure. Although a gold standard might exist, several reasons could justify the development of a new measurement instrument, e.g. greater ease of use, enlargement of the spectrum of potential users, and/or cost considerations.

a) CONCURRENT VALIDATION: If the new instrument and the gold standard method are applied simultaneously, concurrent validation could be established. A correlation coefficient could also be computed between the scores generated by the two approaches and become the validation coefficient.

Concurrent validation of the CNS has been accomplished using data collected on admission to a Neurological Intensive Care Unit (NICU). Standard neurological examinations were performed on all patients admitted to the NICU by two physicians working independently but using a specially designed assessment form. Following these independent assessments, the neurologists were asked to reach consensual decisions on classifying the severity of strokes in these patients in order to establish a gold standard for the validation exercise. The CNS was administered independently by two nurses to all eligible patients upon admission to the NICU. The assessments by the physicians and nurses were all done within approximately 2 hours of each other.

To assess its concurrent validity, the CNS scores (nurses' assessments) will be correlated with the gold standard (physicians' assessments).

b) PREDICTIVE VALIDITY: Predictive validity relates the scores on a measurement instrument to the future health status of patients as measured by mortality, morbidity, and quality of life indicators. Multivariate techniques can be used to model the data and further validation can be achieved by applying the predictive regression equation to a validation set. This approach enables us to determine the proper weighting to be given to specific items in the stroke scale.

Mortality, morbidity and activities of daily living (ADL)[31] were assessed at regular follow-ups at 2 weeks, 1 month, 3 months and 6 months for all patients participating in the CNS validation exercise. Various multivariate techniques will be used to explore its predictive power.[32,33,34]

3) Construct Validity

The construct validity of a stroke scale is a measure of its ability to behave in a predetermined hypothesized fashion that is compatible with a theoretical framework. There are two forms of construct validity: discriminant and convergent validity.

a) DISCRIMINANT VALIDITY: A stroke scale has discriminant validity if the type of neurological deficit it measures is clearly differentiated from a second type of deficit as measured by another scale. Thus, a correlation coefficient can be computed between the scale under study and the second scale alleged to measure a different type of deficit. Discriminant validity is established if the correlation coefficient between the two scales is low.

In the CNS study, discriminant validity will be assessed by comparing the CNS scores and the Glasgow Coma Scores obtained on patients on their admission to the NICU. The limitations of the Glasgow Coma Scale in monitoring stroke patients with no or only a minor decrease in their level of consciousness are the main justification for developing the CNS scale. We hypothesize that the scores obtained on the two scales on the same panel of patients will correlate poorly. A small correlation between the two would indicate good discrimination and add to the construct validity of the CNS.

b) CONVERGENT VALIDITY: A scale has convergent validity if its results correlate highly with other measures of the same construct or attribute.

An important feature that the CNS should exhibit is the ability to reflect changes in patient status over time. For example, if a clinical assessment of a patient is poor on admission but is improved two days later, the score obtained by applying the CNS should show a similar change. This validation exercise will be performed on a subgroup of patients. Only those patients exhibiting an unstable condition during their stay in the NICU

https://doi.org/10.1017/S0317167100027712 Published online by Cambridge University Press

will be retained for this analysis. They will be given a CNS assessment (by a nurse) and a standard neurological examination (by a physician). High correlation between the two sets of scores would indicate that the CNS scale is able to monitor changes in the clinical status of stroke patients. Convergent validity would also be established and would provide further support for the construct validity of the scale. This validation approach is equivalent to establishing the responsiveness of the measurement scale.[27]

Validation is a cumulative and evolving rather than an all or none process. The different approaches we describe can be used alone or in combination.

B) Reliability Assessment

Random error in a measurement can arise from different sources: the measurement itself, the person administering the instrument, or the person to whom it is being administered. Different approaches exist to assess the reliability of a measurement instrument.

1) Measures of internal consistency

The internal consistency of a measurement instrument is a measure of its internal cohesiveness. It is a function of two factors: the number of items in the scale and the mean correlation between them. To increase the reliability of a measurement scale, one must not only increase the number of items in the scale but also increase the inter-item correlation.

Computation of Cronbach's alpha coefficient for continuous data or the Kuder Richardson 20 coefficient for dichotomous data are two methods developed in psychometric theory to assess the internal consistency of measurement scales.[25,26,35]

The internal consistency of the CNS will be measured by computing Cronbach's alpha.

2) Inter- intra-observer Agreement

Reliability assessment of measurement scales establishes the degree of reproducibility of measurements made by different observers and by one observer at different times. Inter-intra-observer agreement can be assessed by using Kappa and weighted Kappa statistics. These statistics assess the level of agreement between observers or within a specific observer while correcting for chance agreement.[36-38]

Inter-observer agreement will be assessed from the CNS scores gathered initially on all patients by calculating the agreement between the results obtained by the two nurses who independently scored them.

Observer agreement will be analyzed using Kappa for the two level items in the CNS and weighted Kappa statistics for items involving more than two ordinal categories and for the total score.

Although other approaches to reliability assessment of a measurement exist,[25,26] they have not been used to assess the reliability of the CNS.

Conclusion

Recent technology has opened new fields of research in cerebrovascular disease and prompted the generation of novel hypotheses that may, after appropriate testing, lead to new therapeutic approaches. The application of modern methodology to clinical research on stroke demands objective and reproducible means of assessing neurological function to permit valid comparison of patients and objective assessment of ultimate outcomes. Validated stroke assessment scales are one way of advancing the application of the scientific method to clinical neurology.

Acknowledgements

The authors are grateful to Dr. N.J.B. Wiggin for his professional assistance and Ms. Diane Telmosse for her technical support.

References

1. Kurtzke JF. Epidemiology of cerebrovascular disease Cerebrovascular Survey Report for Joint Council Subcommittee on Cerebrovascular Disease. National Institute of Neurological and Communicative Disorders and Stroke and National Heart and Lung Institute. January 1980: 135-176.
2. Britton M, Roden A. Progression of stroke after arrival at hospital. Stroke 1985; 16: 629-632.
3. Hachinski V, Norris JW. The deteriorating stroke. *In*: Meyer JS, Lechner H, Reivich M, Ott EO, Aranibar A, eds. Proceedings of the 10th International Salzburz Conference, 1980.
4. Jones HR, Millikan CH. Temporal profile (clinical course) of acute carotid system cerebral infarction. Stroke 1976; 7: 64-71.
5. Jones HR, Millikan CH, Sandok BA. Temporal profile (clinical course) of acute vertebrobasilar system cerebral infarction. Stroke 1980; 11: 173-177.
6. Shinar D, Gross CR, Mohr JP, et al. Interobserver variability in the assessment of neurologic history and examination in the stroke data bank. Arch Neurol 1985; 42: 557-565.
7. Jongbloed L. Prediction of function after stroke: A critical review. Stroke 1986; 17: 765-776.
8. Allen CMC. Predicting the outcome of acute stroke: A prognostic score. J Neurol Neurosurg Psychiatry 1984; 47: 475-480.
9. Teasdale G, Jennett B. Assessment of coma and impaired consciousness. A practical scale. Lancet 1974; 2: 81-83.
10. Côté R, Hachinski V, Shurvell BL, et al. The Canadian neurological scale: a preliminary study in acute stroke. Stroke 1986; 17: 731-737.
11. Lindsay KW, Teasdale GM, Knill-Jones RP. Observer variability in assessing the clinical features of subarachnoid hemorrhage. J Neurosurg 1983; 58: 57-62.
12. Sandercock PAG, Allen CMC, Cornston RN, et al. Clinical diagnosis of intracranial hemorrhage using Guy's hospital score. Br Med J 1985; 291: 1675-1677.
13. Spence JD, Donner AP. Problems in design of stroke treatment trials. Stroke 1982; 13: 94-99.
14. Patten BM, Mendell J, Bruun B, et al. Double-blind study of the effects of dexamethasone on acute stroke. Neurology 1972; 22: 377-383.
15. Mulley G, Wilcox RG, Mitchell JRA. Dexamethasone in acute stroke. Br Med J 1978; 2: 994-996.
16. Fawer R, Justafré JC, Berger JP, et al. Intravenous glycerol in cerebral infarction: A controlled 4-month trial. Stroke 1978; 9: 484-486.
17. Duke RJ, Bloch RF, Turpie AGG, et al. Intravenous heparin for the prevention of stroke progression in acute partial stable stroke: A randomized controlled trial. Ann Intern Med 1986; 105: 825-828.
18. Côté R, Battista RN, Wolfson C, et al. Validation and reliability assessment of the Canadian neurological scale in acute stroke. (In preparation)
19. Oxbury JM, Greenball RCD, Grainger KMR. Predicting the outcome of stroke: Acute stage after cerebral infarction. Br Med J 1975: 125-127.
20. Caronna JJ, Levy DE. Clinical predictors of outcome in ischemic stroke. Neurologic Clinics 1983; 1: 103-117.
21. Chambers BR, Norris JW, Shurvell BL, et al. Prognosis of acute

https://doi.org/10.1017/S0317167100027712 Published online by Cambridge University Press

stroke. Neurol 1987; 37: 221-225.

22. Norris JW, Hachinski VC. Comment on "Study design of stroke treatments". Stroke 1983; 15: 527.
23. Tuthill JE, Pozen TJ, Kennedy FB. A neurological grading system for acute strokes. Am Heart J 1969;78:53-57.
24. Dombovy ML, Sandok BA, Basford JR. Rehabilitation for stroke: A review. Stroke 1986; 17: 363-369.
25. Nunnally JC. Psychometric theory. New York: McGraw-Hill, 1978.
26. Carmines EG, Zeller RA. Reliability and validity assessment. A Sage University Paper, 1978.
27. Guyatt G, Walter S, Norman G. Measuring change over time: Assessing the usefulness of evaluative instruments. J Chronic Dis 1987;40: 171-178.
28. Britton M, de Faire V, Helmers C, et al. Prognostication in acute cerebrovascular disease. Acta Med Scand 1980; 207: 37-42.
29. Prescott RJ, Garraway WM, Akhtar AJ. Predicting functional outcome following acute stroke using a standard clinical examination. Stroke 1982; 13: 641-647.
30. Frithz G, Werner I. Studies on cerebrovascular strokes. Clinical findings and short term prognosis in a stroke material. Acta Med Scand 1976; 199: 130-133.
31. Katz S, Akpom CA. A measure of primary sociobiological functions. Int J Health Serv 1976; 6: 493-507.
32. Breslow NE, Day NE. Statistical Methods in Cancer Research: The Analysis of Case Control Studies. IARC 1, 1980.
33. Kalbfleish JD, Prentice RP. The statistical analysis of failure time data. John Wiley and Sons, New York, 1980.
34. Wagner DP, Knaus WA, Draper EA. Statistical validation of a severity of illness measure. Am J Public Health 1983; 73: 878-884.
35. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951; 16: 297-334.
36. Cohen J. A coefficient of agreement for nominal scales. Educ Psych Meas 1960; 20: 37-46.
37. Cohen J. Weighted Kappa: Nominal scale agreement with provision for scales disagreement or partial credit. Psychol Bull 1968; 70: 213-220.
38. Wolfson C. Measures of agreement for qualitative data. Unpublished M.Sc. Thesis. Department of Mathematics, McGill University, 1978.

https://doi.org/10.1017/S0317167100027712 Published online by Cambridge University Press