INTER-RATER RELIABILITY OF THE SØDRING MOTOR EVALUATION OF STROKE PATIENTS (SMES)

Karin Ek Halsaa, RPT, Karen Margrethe Sødring, RPT, Eva Bjelland, RPT, Kari Finsrud, RPT and Erik Bautz-Holter, MD, PhD

From the Clinic for Geriatrics and Rehabilitation Medicine, University Hospital, Oslo, Norway

ABSTRACT. The Sødring Motor Evaluation of Stroke patients is an instrument for physiotherapists to evaluate motor function and activities in stroke patients. The rating reflects quality as well as quantity of the patient's unassisted performance within three domains: leg, arm and gross function. The inter-rater reliability of the method was studied in a sample of 30 patients admitted to a stroke rehabilitation unit. Three therapists were involved in the study; two therapists assessed the same patient on two consecutive days in a balanced design. Cohen's weighted kappa and McNemar's test of symmetry were used as measures of item reliability, and the intraclass correlation coefficient was used to express the reliability of the sumscores. For 24 out of 32 items the weighted kappa statistic was excellent (0.75–0.98), while 7 items had a kappa statistic within the range 0.53–0.74 (fair to good). The reliability of one item was poor (0.13). The intraclass correlation coefficient for the three sumscores was 0.97, 0.91 and 0.97. We conclude that the Sødring Motor Evaluation of Stroke patients is a reliable measure of motor function in stroke patients undergoing rehabilitation.

Key words: cerebrovascular disorders; physiotherapy; reliability of the Sødring Motor Evaluation of Stroke patients.

INTRODUCTION

Almost all stroke patients develop motor and balance problems. Most receive intensive physiotherapy to promote recovery of the affected side and facilitate normal movement by using both sides of the body to balance and move (5). A systematic physiotherapeutic assessment of stroke patients includes a motor evaluation, which is important in planning treatment and assessing changes in motor function over time. Assessment methods should be relevant, valid, reliable, sensitive to change in the clinical condition, easy to use and communicable. Appropriate clinical instruments are essential in stroke research. The Sødring Motor Evaluation of Stroke patients (SMES) (10) was designed to provide relevant information for physiotherapists in their clinical work with stroke patients and to be useful in stroke research. The main characteristics of the SMES are that the rating reflects quality as well as quantity of performance, and that it measures the patient's ability to carry out activities unassisted. By contrast, assessment methods in which the patient is helped into position for a test register the patient's and the therapist's combined effort. As the amount of assistance is difficult to measure, this tends to invalidate the recordings of the patient's true motor capacity.

Using the SMES, the item scoring is ordinal with either three or five levels, depending on the function being assessed. In a previous study, factor analysis produced a three-factor pattern: "Leg", "Arm" and "Gross function". The construct, concurrent and predictive validity of the SMES has been found to be good (10, 11). During rehabilitation, a stroke patient may be assessed by more than one physiotherapist and high reliability between scorings made by different raters is therefore essential. The purpose of this study was to examine the inter-rater reliability of the SMES.

MATERIAL AND METHODS

To examine whether SMES is useable on all levels, patients considered for recruitment were categorized crudely by a physiotherapist into one of three groups according to global motor function (low, medium or high). A convenience sample of 30 individuals (10 low, 12 medium and 8 high) who had suffered an acute stroke was recruited from the Rehabilitation Stroke Unit, Ullevål University Hospital. Patients with impairment causing difficulties in understanding verbal/non-verbal communication (such as a cognitive deficit or aphasia) and those who for other medical reasons could not be tested (such as amputees) were excluded. The stroke was a recurrent one for 4 patients and a first ever one for 26 patients. The mean age of patients was 77 years (range 54–92 years) and 60% were women. Of this population, 16 had a right-hemisphere lesion, 12 a left-hemisphere lesion and 2 a brainstem stroke. Three therapists

Table I. *Assessment plan*

No. of patients	Physiotherapist at assessment 1	Physiotherapist at assessment 2
5	A	B
5	A	C
5	B	A
5	B	C
5	C	B
5	C	A

were involved in the study. They were all experienced in the assessment and treatment of stroke patients. Prior to commencing the study they had two weeks practical training with the SMES. The patients were tested on two consecutive days. Each therapist tested 10 patients on day one and 10 on day two. The first assessment was made 5–103 days after the stroke (mean 20 days). The test conditions were made as similar as possible; at the same time of day and in the same environment.

The design was balanced (Table I). According to the assessment plan a therapist could occur in both assessment group one and two. In this case the kappa values will measure a mix of the level of agreement between raters and of agreement between the first assessment and the second assessment. This will tend to decrease the estimated reliability. The design with three raters was chosen because of the anticipated workload on the raters. As a measure of item reliability, Cohen weighted kappa statistics was used. The weights for a 3 × 3 table are 0.25 and 1 and for a 5 × 5 table 0.06, 0.25, 0.56 and 1. McNemar's test of symmetry was used to test for bias (both using exact statistical methods (6)). The interpretation of a weighted kappa is not straightforward (3). Fleiss (2) has modified the arbitrary benchmarks given by Landies and Koch for kappa as well as for weighted kappa. He recommends that kappa values $\leq$0.40 signify poor agreement, values between 0.40 and 0.75 fair to good agreement and values $\geq$0.75 excellent agreement (8). This categorization is frequently used in the literature (3). The SMES instrument has previously been factor analysed (10). The sumscores of the items (unweighted) belonging to each factor were used as indices of global functioning within each domain.

In an investigation in which scores from different raters are pooled, one has to choose a coefficient of reliability that is sensitive to both random and systematic difference among raters (9). The intraclass correlation coefficient ICC (2) is such a coefficient. The ICC was used to express reliability of the sumscores and was calculated using BMDP program 5V (1). As mentioned, SMES differs from other assessment methods in the sense that it evaluates unassisted movements only. In a previous study, this design seemed to give the method a higher predictive validity than the reference method (11). However, in the present study it created some methodological problems: the SMES assessment starts with the patient in a supine position. All the four items of leg function and the three first items of arm function are carried out supine. For subjects unable to get into the sitting position unassisted, the SMES comes to a halt when these initial eight items have been carried out. The remaining items will then be rated with the lowest possible score: 1 (cannot perform the activity). For patients able to sit up by themselves, the assessment continues in the sitting position, but if he or she is unable to stand up, the scoring stops for all items in standing position. For nine patients in this study, the assessment halted after item 8, while for one additional subject it halted for item 20 and 22–32. As a consequence of this, the data for item 9 onwards were omitted from the statistical analysis for nine patients, and for ten patients in all for item 20 and after item 21, in order to produce unbiased estimates.

RESULTS

Inter-rater agreement was calculated both for each item and for the sumscores. Table II shows that all items concerning leg function had excellent reliability according to Cohen's kappa statistics. For arm function (Table III), 10 out of 16 items had excellent reliability, 5 items had fair to good reliability and 1 poor reliability. For the factor "Gross function" (Table IV), 10 out of 12 items had an excellent reliability, whereas 2 showed fair to good reliability. When the subgroup with low functioning, i.e. unable to carry out item 8 and higher, was analysed separately, the kappa values were of the same magnitude for the first 7 items. McNemar's test of symmetry indicated no significant bias on any item. The intraclass correlation coefficient demonstrated excellent reliability for all three sumscores (leg 0.97, arm 0.91 and gross function 0.97).

DISCUSSION

This study shows that the SMES is a reliable method for assessing motor function in stroke patients. Reliability is the extent of agreement between repeated measurements.

Table II. *Inter-rater reliability of items of factor "Leg function"*

Item	Weighted kappa (95%CI)	Level of agreement	McNemar's test of symmetry, *p*-value
Supine			
1 Flex hip/knee	0.98 (0.96–1.00)	Excellent	1.00
2 Place feet on plinth	0.91 (0.83–0.99)	Excellent	0.39
3 Dorsiflex ankle, leg straight	0.86 (0.71–1.00)	Excellent	0.58
4 Bridging	0.94 (0.89–0.98)	Excellent	0.34

Table III. *Inter-rater reliability of items of factor "Arm function"*

Item	Weighted kappa (95%CI)	Level of agreement	McNemar's test of symmetry, *p*-value
Supine			
5 Hand towards opposite shoulder	0.95 (0.95–0.99)	Excellent	0.15
6 Lift straight arm up/down	0.90 (0.82–0.99)	Excellent	0.21
7 Arm up: flex./ext. elbow	0.94 (0.87–1.00)	Excellent	0.35
8 From supine to sitting	0.79 (0.65–0.93)	Excellent	0.61
Sitting			
9 Hand towards opposite shoulder	0.81 (0.66–0.96)	Excellent	0.38
10 Support on straight arm	0.75 (0.60–0.91)	Excellent	0.50
11 Lift straight arm up/down	0.78 (0.63–0.94)	Excellent	0.27
12 Stretch arm forward	0.80 (0.67–0.93)	Excellent	0.50
13 As #10, flex/ext. elbow	0.53 (0.27–0.78)	Fair to good	0.27
14 Flex./ext. fingers	0.74 (0.53–0.96)	Fair to good	0.34
15 Opposition of fingers	0.73 (0.53–0.93)	Fair to good	0.36
16 Bring fork/spoon to mouth	0.87 (0.73–1.00)	Excellent	0.69
17 Hold/cut meaty object	0.93 (0.82–1.00)	Excellent	0.75
18 Tip to affected side	0.13 (–0.25–0.52)	Poor	0.60
19 Tip to sound side	0.63 (0.29–0.97)	Fair to good	0.50
Standing			
20 Protective ext. of hands forward	0.59 (0.24–0.94)	Fair to good	0.53

Table IV. *Inter-rater reliability of items of factor "Gross function"*

Item	Weighted kappa (95%CI)	Level of agreement	McNemar's test of symmetry, *p* value
Sitting			
21 From sitting to standing	0.76 (0.58–0.95)	Excellent	0.46
Standing			
22 Step forward with sound leg	0.78 (0.59–0.97)	Excellent	0.26
23 Step forward with affected leg	0.78 (0.59–0.97)	Excellent	0.17
24 Walk 10 m forward	0.93 (0.87–0.99)	Excellent	0.34
25 Walk 2 m backwards	0.93 (0.87–0.99)	Excellent	0.11
26 Walk 10 steps upstairs	0.81 (0.64–0.99	Excellent	0.39
27 Walk 10 steps downstairs	0.93 (0.87–1.00)	Excellent	0.69
28 Tip backwards	0.55 (0.09–0.99)	Fair to good	0.63
29 Tip, standing on sound leg	0.66 (0.37–0.95)	Fair to good	0.38
30 Tip, standing on affected leg	0.75 (0.53–0.96)	Excellent	0.50
31 Standing to lying on floor	0.88 (0.76–0.99)	Excellent	0.69
32 Lying on floor to standing	0.93 (0.83–1.00)	Excellent	0.25

There will always be random errors because of unsystematic chance factors that confound the measurement. Non-random error is systematic bias; for example, if one of the observers systematically rates with a lower or a higher score . The kappa statistics is a coefficient of agreement corrected for chance agreement and has been advocated as the only acceptable method of assessing intrapatient/intraobserver as well as interobserver variability of scale items (8). A problem with the weighted kappa is that false high values may result if the scorings cluster in some region of the item scale. However, this was not the case in our study, where the patients covered a wide range of functions.

According to Fleiss's criteria, only one out of 32 items (item 18, "sitting, tip to affected side") reached a low level of agreement.

Random errors in assessment methods can arise from different sources: the assessment method itself, the observer or the patient. The following examples from our study confirm this: In the SMES, items testing balance reactions and gross functions are rated on a three-level scale. This is believed to give more reliable data than the more fine-meshed five-level scale used for the arm and leg items (4). Despite this, we found a lower level of agreement on the items scored on three levels, just as we had for the items testing balance reactions. The low kappa for item 18 appears to be due to the effect of non-random error, as one of the therapists rated that item on a lower score than the others two-third of the time. It is our experience that the evaluation of the quality of motor function comprising greater parts of the body, such as balance reactions, is complicated to operationalize and so it is more difficult to get agreement on ratings. Furthermore, the items on balance are the only SMES items in which the therapist may touch the patient to elicit the reactions and incorrect handling may affect the results. We did consider omitting item 18 from the SMES because of its poor reliability. The argument for not doing so, is that we feel that this test is a valuable contribution to the assessment of the patient. A way of dealing with this problem could be to improve the written instructions in the manual, as well as stressing the evaluation of this item in the practical training before the use of the SMES.

It might be thought that assessment on two consecutive days could have an impact on the results from one day to the next. In one case this was observed: the patient was unable to sit up from supine the first time he was tested, 17 days after onset. The next day the test results were different, in that the patient obtained better scores in the arm as well as in gross function, and then he did manage to sit up. The ratings for the leg function were the same. The reason why the patient improved is difficult to ascertain. It might have been that he was in bad form the first day, which in our experience often has an impact on motor capacity. Alternatively, a spontaneous functional improvement may have taken place, since the patient was still in quite an early phase of recuperation.

We conclude that the SMES has high inter-rater reliability provided that the raters are familiar with stroke and movement analysis and have been trained in the use of the instrument. As it is the inter-rater reliability that has been examined, a generalization of the results is justified to studies where different raters are involved, as well as in test-retest situations.

ACKNOWLEDGEMENT

We thank Professor Knut Laake for his contribution to the statistical analyses and constructive criticism during the project.

REFERENCES

1. Dixon, W. J.: BMDP Statistical Software. University of California Press, Berkeley, 1992.
2. Fleiss, J. L.: The design and analysis of clinical experiments, pp. 1–32, John Wiley & Sons, New York, 1986.
3. Jelles, F., Bennekom, C. A. M. V., Lankhorst, G. J., Sibbel, C. J. P. & Bouter, L. M.: Inter- and intrarater agreement of the rehabilitation active profile. J Clin Epidemiol *48:* 407–416, 1995.
4. Keith, R. K.: Functional assessment measures in medical rehabilitation: current status. Arch Phys Med Rehabil *65:* 74–78, 1984.
5. Lennon, S.: Key physiotherapy indicators for quality of stroke care. Physiotherapy *82:* 655–665, 1996.
6. Metha, C. & Patel, N.: StatXact 3 for Windows. Cytel Software Corporation. Cambridge, MA, 1995.
7. Richman, J.: Research methodology and applied statistics: 3. Measurement procedures in research. Physiother Canada *32:* 253–257, 1980.
8. Sheik, K.: Disability scales: assessment of reliability. Arch Phys Med Rehabil *67:* 245–249, 1986.
9. Shrout, P. R. & Fleiss, J. L.: Intraclass correlation: uses in assessing rater reliability. Psychol Bull *86:* 420–428, 1979.
10. Sødring, K. M., Bautz-Holter, E., Ljunggren, A. E. & Wyller, T. B.: Description and validation of motor function and activities in stroke patients. The Sødring Motor Evaluation of Stroke patients. Scand J Rehab Med *27:* 211–217, 1995.
11. Wyller, T. B., Sødring, K. M., Sveen, U., Ljunggren, A. E. & Bautz-Holter E.: Predictive validity of the Sødring Motor Evaluation of Stroke patients (SMES). Scand J Rehab Med *28:* 211–216, 1996.

Accepted March 17, 1999

Address for offprints:

Karin Ek Halsaa, Physiotherapist
Clinic for Geriatrics and Rehabilitation Medicine
Ullevål University Hospital
NO-0407 Oslo
Norway