Scand J Rehab Med 28: 3–7, 1996

# A CRITERION FOR STABILITY OF THE MOTOR FUNCTION OF THE LOWER EXTREMITY IN STROKE PATIENTS USING THE FUGL-MEYER ASSESSMENT SCALE

H. Beckerman, MSc PT,[1] T.W. Vogelaar, PT,[1] G.J. Lankhorst, MD PhD[1] and A.L.M. Verbeek, MD PhD[2]

*From the [1]Department of Rehabilitation Medicine, Academisch Ziekenhuis Vrije Universiteit Amsterdam, and [2]Department of Epidemiology, University of Nijmegen, The Netherlands*

**ABSTRACT. A test-retest reproducibility study was performed to define a criterion for stability as opposed to change of motor function of the lower extremity in stroke patients. Forty-nine patients with stroke were examined twice by the same physiotherapist, using the Fugl-Meyer Assessment Scale. The interval between both measurements was three weeks. The mean differences between the first and the second measurement were small, with 0.04 points for the lower extremity scale and 0.92 points for the balance scale, respectively. Intraclass correlation coefficient for the lower extremity scale was 0.86, and 0.34 for the balance scale. The standard error of measurement for each scale was 1.76 and 1.17 points, respectively. The standard error of measurement can be transformed in an 'error threshold', which is a criterion to differentiate real changes from changes due to chance variation or measurement error. As the absence of real change is a parameter for stability, a change of less than 5 points for the lower extremity scale and of less than 4 points for balance confirms stability of motor function.**

*Key words:* reproducibility, cerebrovascular accident, lower extremity, measurement, reliability.

## INTRODUCTION

Stroke recovery generally begins early, with the main improvement occurring over the first three to four months after stroke. Thereafter, the recovery pattern generally reaches a plateau (7, 13, 16, 18). In randomized clinical trials it is thought necessary, although not always possible, only to include patients with a stable neurologic status, in order to prevent bias of the study results by spontaneous recovery (14). The Fugl-Meyer Assessment Scale, a disease-specific performance-based measurement instrument especially designed to assess the recovery pattern of physical functions following stroke, was considered for discrimination of the stable and unstable group (6, 8, 9). The Fugl-Meyer Assessment Scale consists of six separate scales to measure quantitatively motor function of the upper extremity and lower extremity respectively, balance, pain, range of motion, and sensation. Nevertheless, it is unknown which criterion, i.e. how many points of change during a follow-up period, has to be used in order to classify patients as neurologically stable or unstable. A quantification of the magnitude of the measurement error in stable subjects could answer the question how much difference is due to real change, and how much is due to chance variation or measurement error. In the absence of real change it can be concluded that there is a stable or stationary status.

The purpose of this study was to define a criterion for stability of the motor function of the lower extremity and balance, using a test-retest design.

## METHODS

### *Patients*

The study population consisted of stroke patients referred to the outpatient Department of Rehabilitation Medicine of the 'Academisch Ziekenhuis Vrije Universiteit', Amsterdam. Patients between the ages of 18 to 75 years, who at least one year previously had suffered an ischaemic or haemorrhagic stroke of a cerebral hemisphere resulting in hemiplegia, could participate. They were all experiencing walking problems caused by a spastic equinus or equinovarus position of the foot. Patients without sufficient communication and cognition functions, or with an unsatisfactory general condition were excluded. The participants were examined on two separate occasions with a 3-week interval. No clinically relevant changes in the motor function status of these chronic patients were expected during this interval. All patients were examined by the same experienced physiotherapist (second author TWV).

Table I. *Characteristics of the study population (n = 49)*

| Variable | Number of participants |
|---|---|
| Gender | |
| Female | 16 |
| Male | 33 |
| Age (yrs) | |
| ≤ 50 | 10 |
| 51–60 | 22 |
| 61–70 | 14 |
| > 70 | 3 |
| Months post-stroke | |
| 13–24 | 15 |
| 25–36 | 9 |
| 37–48 | 4 |
| 49–60 | 7 |
| > 60 | 14 |
| Hemiplegic side | |
| Left | 24 |
| Right | 25 |
| Type of stroke | |
| Haemorrhagic | 13 |
| Ischaemic | 36 |

Sixteen women and 33 men, with a median age of 58 years (range: 21–72 years), and with a median time since stroke of 37 months (range: 13–185 months) gave their informed consent to participate in the study (Table I).

*Measurement instrument*

In this study we used two subscales of the Fugl-Meyer Assessment Scale: the 'motor function of the lower extremity' (FM-LE, 17 items, maximum score of 34 points) and the 'balance' scale (FM-B, 7 items to test the sitting and standing balance, maximum score of 14 points). Most items consist of standardized motor activities which are to be performed independently by the patient. The scoring involves direct observation of the performance. Each performance is rated on an ordinal 3-point scale (0 = the item cannot be performed; 1 = the item can be partially performed; 2 = the item can be fully performed) (6).

*Data analysis*

Test-retest reproducibility was assessed by calculating 3 different numerical indexes: the intraclass correlation coefficient, the standard error of measurement, and the error threshold. These reproducibility indexes are all based on the analyses of variance components. Analysis of variance (ANOVA) was performed using the PC version of the program GENOVA, developed by Crick & Brennan (3).

1. *Intraclass Correlation Coefficient (ICC).* ICC is a preferred method of quantifying reproducibility (10). The ICC can be calculated as the ratio of the variance between subjects (i.e. patients) and the total variance.

2. *Standard Error of Measurement (SEM).* Quantifying the test-retest reproducibility of an assessment involves calculating the variability in measurements of the same patient. The SEM provides an interpretation of the magnitude of this within-subject variability, which is also known as error variance. The SEM is the square root of the within-subject variance (11, 17), and is expressed in the same dimension as the measurement. (Note that the standard error of measurement (SEM) is not synonymous with the standard error of the mean, also abbreviated as SEM.)

3. *Error Threshold (ET).* Assuming that the measurement errors of two measurements are independent of each other, an interval can be calculated which expresses the uncertainty about the difference between two true scores. The difference between both measurements should be at least $1.96*\sqrt{2}*SEM$ in order to be 95% confident of a real difference between the true scores. We call the quantity $1.96*\sqrt{2}*SEM$ the 'error threshold' (ET). In other words, the ET is a criterion to differentiate real changes from changes due to chance variation or measurement error.

## RESULTS

*Criterion for stability*

Descriptive statistics (the mean and standard deviation) for both parts of the Fugl-Meyer Assessment Scale are presented in Table II. On the first measurement, the FM-LE score varied between 5 and 29 points (15%–85% of the maximum possible score), whereas the FM-B score varied between 6 and 14 points (43%–100% of the maximum possible score).

Table II. *Mean and standard deviation of the Fugl-Meyer Assessment Scale on two separate occasions with a three-week interval*

SD = standard deviation

| | First measurement | | Second measurement | | Difference | |
|---|---|---|---|---|---|---|
| Fugl-Meyer Assessment Scale | Mean | SD | Mean | SD | Mean | SD |
| Lower extremity (0–34) | 17.67 | 4.50 | 17.71 | 4.89 | −0.041 | 2.483 |
| Balance (0–14) | 12.78 | 1.64 | 13.69 | 0.80 | −0.918 | 1.382 |

Table III. *Test-retest reproducibility results of the Fugl-Meyer Assessment Scale*
ICC = Intraclass Correlation Coefficient
SEM = Standard Error of Measurement
ET = Error Threshold

| Fugl-Meyer Assessment Scale | Between subject variance | Within subject variance | Total variance | ICC | SEM | ET |
|---|---|---|---|---|---|---|
| Lower extremity (0–34) | 18.988 | 3.082 | 22.070 | 0.86 | 1.756 | 4.87 |
| Balance (0–14) | 0.701 | 1.357 | 2.058 | 0.34 | 1.165 | 3.23 |

$$ICC = \frac{\text{between-subject variance}}{\text{between-subject variance}+\text{within-subject variance}}$$

$$SEM = \sqrt{\text{within-subject variance}} = \sqrt{\text{total variance}} * \sqrt{(1\text{-ICC})}$$

$$ET = 1.96^{*}\sqrt{2}^{*}SEM$$

The mean differences between the first and the second measurement were small. Nevertheless, the differences varied between −5 points and +9 points on the FM-LE scale, and between −6 and +2 points on the FM-B scale. For the FM-B score the mean difference between the first and the second measurement was significantly different from zero (two-tailed paired *t*-test: *p*-value 0.0001).

No statistically significant relationship was found between the test-retest differences and single patient characteristics (Table I; *t*-tests, Pearson's correlation coefficients).

The ICC for the FM-LE is 0.86, and 0.34 for FM-B (Table III). The SEM, which is expressed in the same unit as the Fugl-Meyer score, is 1.76 and 1.17 points for each scale, respectively (Table III). Twice the SEM value approximately encompasses 95% of the obtained scores around the true score. For example, given that one single assessment results in a FM-LE score of 18 points, the true score will lie between 14.5 and 21.5 points. The same calculation could be presented for FM-B. The ET for each scale is 4.87 and 3.23 points, respectively. This means that, to be sure that the changes are real rather than due to a measurement error, the score of the FM-LE should increase by at least 5 points, whereas the score on the balance scale should increase by at least 4 points. These differences are equal to 15% and 29% of the maximum possible score. Smaller changes are to be interpreted as measurement error; there is no indication of real change and it may be inferred that the patient's motor function is stable.

## DISCUSSION

To gain more insight into the methodological quality of instruments measuring change or, as in our case, stability of the variable within a subject over time, the standard error of measurement (SEM) as well as the error threshold (ET) are well suited indexes since the magnitude of within-subject variance is the most relevant (11, 12). The ET, which can be transformed from the SEM, is a real threshold to differentiate real changes from changes due to chance variation or measurement error. ICCs are less appropriate since the variance between subjects is considered as the variance of interest, whereas stability depends on the within-subject variance. Furthermore, in our study population, the ceiling effect of the balance scale resulted in a skewed distribution of balance scores, and a relatively small between-subject variance. This has probably caused the low ICC of 0.34. By choosing a study population with a greater variety (heterogeneity) of balance scores, the resulting ICC would be more impressive (10). Pearson's product moment correlation coefficients are even less appropriate for quantifying test-retest reproducibility, because systematic differences between measurements are neglected (1, 4). Reliability studies of the Fugl-Meyer Assessment Scale using Pearson's product moment correlation coefficient should therefore be interpreted with caution (2, 5). Moreover, reproducibility coefficients such as the ICC and Pearson's product moment correlation coefficient are expressed as a dimensionless number between 0 and 1, and do not open for a straightforward interpretation.

In some way, our results could be compared with the results recently published by Sanford et al. (15), investigating the interrater reliability of the Fugl-Meyer Assessment Scale among three experienced physiotherapists. They investigated 12 acute stroke patients (less than 6 months after the stroke), aged 49 to 86 years (mean age: 66 years). All patients were following an inpatient rehabilitation programme at the time of the reliability study. Patients were tested within one working day of the previous assessment. The interrater ICC for the lower extremity was 0.92 and for the balance score 0.93. The SEMs were 3.20 and 1.00, respectively. As compared with our results, the SEM of the lower extremity is 1.44 points larger, whereas the SEM of the balance score is nearly the same as in our study (1.00 versus 1.17). In general, intrarater (or test-retest) reliability is higher than interrater reliability because each rater brings in some variance. Nevertheless, the ICCs presented by Sanford et al. are better than ours. This could be explained by the differences between the study populations with respect to the between-subject variance. Heterogeneity of subjects may account for higher ICCs. In our chronic stroke patients, the within-subject differences were smaller, resulting in smaller SEMs.

Short-term follow-up studies of stroke patients have shown that the motor recovery usually occurs and reaches its plateau within six months. Our chronic study population was investigated at least one year after their stroke. Therefore the motor function of the lower extremity and the sitting and standing balance of these chronic stroke patients were assumed to be stable between the first and second assessments.

Because of the large standard errors of measurement, observed scores are obviously a very imprecise measure of the unknown true lower extremity and balance scores. Clinical decisions based on these imprecise scores will certainly bear a high risk of false decisions. Our study has shown that in chronic stroke patients a criterion of a 5-point change of the Fugl-Meyer lower extremity score, and a 4-point change of the balance score over a 3-week period seems scientifically valid to differentiate between stable and unstable (improved or deteriorated) stroke patients. These criteria, which are equal to 15% and 29% of the maximum possible score of each scale, are very large. Therefore, in clinical trials as well as in clinical practice, scores on the Fugl-Meyer Assessment Scale should be used with caution.

## ACKNOWLEDGEMENT

This study is part of a project that has been supported by a grant from the Netherlands Heart Foundation (project 91.060).

## REFERENCES

1. Bland, J. M. & Altman, D. G.: Statistical methods for assessing agreement between two methods of clinical measurement. Lancet *i*: 307–310, 1986.
2. Brinker, B. P. L. M. den & Kobus, M. H.: Betrouwbaarheid van bewegingsobservaties van video-opnamen van de Fugl-Meyer test. Ned T Fysiotherapie *98*: 120–122, 1988.
3. Crick, J. E. & Brennan, R.L.: Manual for GENOVA: a generalized analysis of variance system. American College Testing Program, Iowa City, 1983.
4. Deyo, R. A., Diehr, P. & Patrick, D. L.: Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Control Clin Trials *12*: 142S–158S, 1991.
5. Duncan, P. W., Propet, M. & Nelson, S. G.: Reliability of the Fugl-Meyer assessment of sensorimotor recovery following cerebrovascular accident. Phys Ther *63*: 1606–1610, 1983.
6. Duncan, P. W. & Badke, M. B.: Measurement of motor performance and functional abilities following stroke. *In* Stroke rehabilitation: the recovery of motor control. (ed. P. W. Duncan & M. B. Badke). Year Book Publishers Inc, Chicago, 1987.
7. Duncan, P. W., Goldstein, L. B., Matchar, D., Divine, G. W. & Feussner, J.: Measurement of motor recovery after stroke. Outcome assessment and sample size requirements. Stroke *23*: 1084–1089, 1992.
8. Fugl-Meyer, A. R., Jääskö, L., Leyman, J., Olsson, S. & Steglind, S.: The post-stroke hemiplegic patient. I: a method of evaluation of physical performance. Scand J Rehabil Med *7*: 13–31, 1975.
9. Fugl-Meyer, A. R.: Post-stroke hemiplegia: assessment of physical properties. Scand J Rehabil Med suppl 7: 85–93, 1980.
10. Guyatt, G., Walter, S. & Norman, G.: Measuring change over time: assessing the usefulness of evaluative instruments. J Chron Dis *40*: 171–178, 1987.
11. Guyatt, G. H., Kirschner, B. & Jaeschke, R.: Measuring health status: what are the necessary measurement properties? J Clin Epidemiol *45*: 1341–1345, 1992.
12. Kramer, M. S. & Feinstein, A. R.: Clinical biostatistics. LIV. The biostatistics of concordance. Clin Pharmacol Ther *29*: 111–123, 1981.
13. Partridge, C. J., Johnston, M. & Edwards, S.: Recovery from physical disability after stroke: normal patterns as a basis for evaluation. Lancet *i*: 373–375, 1987.
14. Pollock, C., Freemantle, N., Sheldon, T., Song, F. & Mason, J. M.: Methodological difficulties in rehabilitation research. Clin Rehabil *7*: 63–72, 1993.
15. Sanford, J., Moreland, J., Swanson, L. R., Stratford, P. W. & Gowland, C.: Reliability of the Fugl-Meyer Assessment for testing motor performance in patients following stroke. Phys Ther *73*: 447–454, 1993.

16. Skilbeck, C. E., Wade, D. T. & Langton Hewer, R.: Recovery after stroke. J Neurol Neurosurg Psychiatry *46*: 5–8, 1983.
17. Streiner, D. L. & Norman, G. R.: Health measurement scales. A practical guide to their development and use. Oxford University Press, Oxford, 1993.
18. Wade, D. T. & Langton Hewer, R.: Functional activities after stroke: measurement, natural history and prognosis. J Neurol Neurosurg Psychiatry *50*: 177–182, 1987.

*Address for offprints:*

Ms H Beckerman
Department of Rehabilitation Medicine
Academisch Ziekenhuis Vrije Universiteit
PO Box 7057
NL-1007 MB Amsterdam
The Netherlands