SHORT REPORT

Mayo and NINDS scales for assessment of tendon reflexes: between observer agreement and implications for communication

S Manschot, L van Passel, E Buskens, A Algra, J van Gijn

Department of Neurology

S Manschot
L van Passel
A Algra
J van Gijn

Department of Clinical Epidemiology, Utrecht University Hospital, Utrecht, The Netherlands
E Buskens
A Algra

Correspondence to:
Professor J van Gijn,
University Department of
Neurology, Academic
Hospital, Heidelberglaan
100, PO box 85500, 3508
GA Utrecht, The
Netherlands.
Telephone 0031 30 2507975;
fax 0031 30 2542100.

Received 2 December 1996
and in revised form
23 May 1997
Accepted 22 August 1997

Abstract

The purpose of this study was to assess the between observer reliability of two standard notation scales for grading tendon reflexes, The Mayo Clinic scale and the NINDS scale. In a university department of neurology two or three physicians judged the biceps, triceps, knee, and ankle tendon reflexes in two groups of 50 patients using either scale. The interobserver agreement was assessed by means of κ statistics. The agreement among doctors was never better than "fair" for both scales (highest κ value 0.35). A verbal description rather than a codified scale may improve communication among doctors.

(*J Neurol Neurosurg Psychiatry* 1998;64:253-255)

Keywords: notation scales; tendon reflexes; between observer agreement

During physical examination testing of the tendon reflexes is often important. There is, however, no uniformity in the notation of these reflexes. In 1989 a brief survey among neurologists and residents in our institution showed that as many as 20 different scales were used, recognised or idiosyncratic. It is clear that such variation may cause problems in oral and written communication among neurologists and other healthcare professionals. Misunderstandings of this kind could affect the management of patients. Assuming that a generally accepted standard scale for the examination of tendon reflexes might improve this situation, the National Institute of Neurological Disorders and Stroke scale (NINDS) was proposed by Hallett in 1993.¹ Recently, a report appeared suggesting sufficient reliability of the NINDS reflex notation scale.² However, the reliability of this scale has not been established or compared with other notation scales in a realistic between observer study, performed in a routine clinical setting. Accordingly, we set out to establish between observer reliability of the NINDS as well as that of the older Mayo Clinic scale for the assessment of tendon reflexes, among a group of physicians with diverse experience in neurology.^{1,3}

Table 2 NINDS scale for tendon reflex assessment¹

Description	Score
Reflex absent	0
Reflex slight, less than normal: includes a trace response or a response brought out only by reinforcement	+1
Reflex in lower half of normal range	+2
Reflex in upper half of normal range	+3
Reflex enhanced, more than normal: includes a clonus if present which optionally can be noted in an added verbal description of the reflex	+4

Patients and methods

PATIENTS

Tendon reflexes were assessed once by each physician in a randomly chosen group of 100 patients who attended the department of neurology of Utrecht University Hospital. Both outpatients (38) and inpatients (62) were enrolled to create as representative a sample of neurological patients as possible. A group of 50 patients evaluated by means of the Mayo scale consisted of 20 outpatients and 30 inpatients. The other group of 50 patients was evaluated by means of the NINDS scale and consisted of 18 outpatients and 32 inpatients. All patients visiting the outpatient department or admitted to hospital in the neurology ward were considered eligible, except those unable to give oral consent or with amputated limbs. Outpatients were examined by two observers, and patients staying in hospital by three observers in between regular clinical tasks. The physicians were provided with a brief outline of the history, because this reflects regular practice. They were asked to use their own patella hammer and apply their habitual techniques for exciting tendon reflexes. The following reflexes were recorded separately: biceps, triceps, knee, and ankle, both left and right. Also, asymmetry was recorded if corresponding left and right tendon reflexes were judged to be different.

REFLEX ASSESSMENT

The Mayo Clinic scale (table 1) was used in the first 50 patients, and the NINDS scale (table 2) in a second group of 50 patients. The use of the scales was separated in time to prevent confusion of reflex assessment. For each patient, tendon reflexes were assessed by two or three physicians out of a group of 37; neurologists (n=10) or trainees in neurology (n=27).

Table 1 Mayo Clinic scale for tendon reflex assessment¹

Description	Score
Absent	-4
Just elicitable	-3
Low	-2
Moderately low	-1
Normal	0
Brisk	+1
Very brisk	+2
Exhaustible clonus	+3
Continuous clonus	+4

Table 3 Agreement on tendon reflexes observed in left and right limbs (κ (Z_a *SD))

	Left		Right	
	Two judges†	Three judges‡	Two judges†	Three judges‡
Mayo clinic reflex scale:				
Biceps	0.19-0.25	0.25(0.10)	0.12-0.21	0.17(0.10)
Triceps	0.16-0.21	0.16(0.10)	0.05-0.26	0.17(0.10)
Knee	0.04-0.29	0.22(0.09)	0.13-0.32	0.24(0.09)
Ankle	0.05-0.29	0.22(0.09)	0.14-0.19	0.18(0.09)
NINDS reflex scale:				
Biceps	0.04-0.26	0.14(0.14)	0.07-0.28	0.22(0.13)
Triceps	0.11-0.31	0.25(0.13)	0.28-0.32	0.35(0.13)
Knee	0.12-0.43	0.29(0.12)	-0.06-0.28	0.18(0.14)
Ankle	0.12-0.42	0.22(0.12)	-0.02-0.32	0.11(0.12)

† Range of κ observed for two judges.‡ Group κ for three judges; Mayo clinic scale 30 patients, NINDS scale 32 patients.
 κ (Z_a *SD) yields 95% CI.Table 4 Agreement on asymmetry observed between left and right limb reflexes (κ (Z_a *SD))

	Two judges†	Three judges‡
Mayo Clinic reflex scale:		
Biceps	0.15-0.37	0.29(0.15)
Triceps	-0.14-0.26	0.01(0.16)
Knee	-0.17-0.37	0.02(0.15)
Ankle	0.16-0.48	0.28(0.15)
NINDS reflex scale:		
Biceps	0.13-0.32	0.24(0.14)
Triceps	0.00-0.23	0.12(0.16)
Knee	0.21-0.60	0.48(0.16)
Ankle	0.16-0.18	0.18(0.15)

† Range of κ observed for two judges.‡ κ for three judges; Mayo clinic scale 30 patients, NINDS scale 31 patients.
 κ (Z_a *SD) yields 95% CI.

We aimed at an equal number of turns for all 37 physicians, but because of limited availability this was not always possible. On examination of the patients the data were recorded on a standardised form.

DATA ANALYSIS

We compared the absolute scores obtained in individual patients for each reflex by the various observers and also whether asymmetry was found. We used the κ statistic to assess agreement among observers.⁴ This is a measure of agreement among observers corrected for chance agreement. For positive agreement κ assumes values between 0 (chance agreement only) and 1 (perfect agreement). For example, if there are four alternatives (A,B,C,D), two observers, and 20 patients, and the observers agree on 15 of the 20 occasions, the crude agreement rate is 75%, but the κ value is $(0.75-0.25 \text{ (chance agreement)})/(1-0.25 \text{ (maximal agreement after correction for chance agreement)})=0.67$. We should emphasise that κ expresses only the degree of agreement between different findings, not their accuracy. The following appraisal for positive agreement by κ values has been proposed: lower than 0.00 poor agreement; between 0.00 and 0.20 slight agreement; between 0.21 and 0.40 fair agreement; between 0.41 and 0.60 moderate agreement; between 0.61 and 0.80 substantial agreement; between 0.81 and 1.00 almost perfect agreement.⁵ For clinical findings κ values in the order of 0.40–0.70 are usual; κ values between 0.60 and 0.80 are rarely obtained.⁶ We also calculated weighted κ , which takes into account the degree of scale disagreement and allows for partial agreement.⁷

Grouped κ values were calculated for each reflex (Agree, rec proGAMMA, Groningen, the Netherlands). Left and right reflexes were analysed separately. No distinction was made between findings in inpatients or outpatients. Finally, κ values were also calculated for each pair of reflexes for any asymmetry found (in case left and right reflexes were judged to be different).

Results

The average age of our patient group was 51 (range 19–90 years); 54 were men.

For three of the 100 patients examined some assessments could not be interpreted or carried out; one physician could not rate some (three) reflexes in one patient with the NINDS scale because it was considered impossible to choose between low-normal and high-normal, one patient (NINDS scale) showed paradoxical flexion of the arm when the triceps reflex was tested, and the third patient indicated pain so that part of the examination was not possible (Mayo scale; three reflexes).

Table 3 shows the κ values (agreement for absolute scores) and table 4 shows asymmetry. Most κ values indicated only slight agreement, some moderate, and others poor. This applied to both scales. Weighted κ values were somewhat better (up to slight agreement, results not shown).

Discussion

We had expected the NINDS scale to show higher between observer agreement, because fewer options for grading are available and accordingly there is less margin for disagreement. Our study, however, showed low agreement between observers for both scales. Neither the use of the NINDS nor the Mayo Clinic scale yielded acceptable reliability in the performance of different observers. Even calculating weighted κ values resulted in κ values too low to justify routine use of either of these scales. Our findings are not in agreement with a recent study on the NINDS reflex scale in which moderate to substantial between observer reliability and substantial to near perfect within observer reliability was reported.² However, that study was performed under more or less optimal conditions—that is, all examinations were performed by only four clinical neurologists with similar backgrounds, and within a short period of time out of the context of a routine clinical examination. In day to day practice many physicians of different specialties and training, using various techniques and reflex hammers, have to communicate about their findings obtained during routine physical examination. Moreover, only weighted κ values were presented in the other study. Weighted κ values allow for partial agreement—that is, depending on the amount of discrepancy in score credit is given. In practical terms analysis using weights is similar to condensing the scale to a smaller number of steps. From this it might be inferred that a condensed version of the NINDS scale could be used with several adjacent categories actually having similar clinical implications.

With regard to our own study, one factor that, in theory, may have adversely affected agreement is within patient variation of reflex intensity between findings. Reflexes of the same patient may vary because of progression of the disease, difference in posture or muscle tone at the moment of testing, or mental activity. For outpatients the time between the reflex measurements by the two observers was half an hour at most. Such a short period can hardly cause a distortion of the agreement found.⁸ For inpatients the average interval was one day for the Mayo Clinic scale and half a day for the NINDS scale. However, if within patient variability would have been an important factor, κ values for inpatients would systematically have appeared lower than those for outpatients. No such difference in κ values between inpatients and outpatients was found (only for asymmetry they were somewhat lower).

An additional factor that may have influenced agreement is the description of the scales themselves. The Mayo Clinic reflex scale has a separate category for "normal", implying that the remaining categories are abnormal. Accordingly, the physician might be discouraged from using these "abnormal" categories, thus increasing agreement. On the other hand the NINDS scale does not have a separate category for "normal", and it is necessary to choose between low-normal or high-normal, which often proves difficult. Nevertheless, both scales appeared equally (un)reliable.

Finally, some disagreement may be related to the experience of physicians. For example, eliciting ankle or knee clonus usually requires separate manoeuvres. It seemed that a clonus was sometimes not discovered because neither the history nor the briskness of the tendon jerks prompted the specific test. Similarly, some reflexes were recorded as absent by one physician whereas another had no problem eliciting them. Varying clinical background and experience may have been an underlying cause

for lack of agreement. Yet the current selection of physicians reflects the reality of clinical practice. If anything, still a rather optimistic estimate has been obtained as only neurologists and trainees in neurology were involved.

Numerical codes have been assigned to the steps in both scales. These imply a degree of precision which, as we have shown, is unrealistic. We conclude that probably no reflex notation scale that requires coded rating in the form of numbers will reach sufficient between observer agreement for general use. Perhaps a condensed classification (as is created by analysing the results with a weighted κ) by means of a plain verbal description of the observed tendon reflexes would be most satisfactory. Terms that are understood and used by everyone include absent, low, average, brisk, a few beats of clonus, and permanent clonus. Descriptions by means of other terms would still be understood by others as long as codifications are avoided. Plain words may prevent misunderstanding between physicians and thus possible harm to patients. Asymmetry often indicates a pathological condition in itself and should be recorded separately. Obviously, a formal classification in words would again require assessment of its reliability.

We are grateful to Professor J Stam (Amsterdam) for his advice.

- 1 Hallett M. NINDS myotatic reflex scale. *Neurology* 1993;43:2723.
- 2 Litvan I, Mangone CA, Werden W, Bueri JA, Estol CJ, Garcea DO, et al. Reliability of the NINDS myotatic reflex scale. *Neurology* 1996;47:969-72.
- 3 Bastron JA, Bickford RG, Brown JR, et al. *Clinical examinations in neurology*. Mayo Clinic Staff, 1956:188.
- 4 Cohen J. A coefficient for agreement for nominal scales. *Educational and Psychological Measurement* 1960;20:37-46.
- 5 Landis RJ, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159-74.
- 6 Veldhuyzen van Zanten SJO, Hijdra A. The use of kappa in the study of variability between observers. *Ned Tijdschr Geneesk* 1988;132:199-202.
- 7 Cohen J. Weighted kappa: nominal scale agreement with provision for scale disagreement or partial credit. *Psychol Bull* 1968;70:213.
- 8 Crayton JW, King S. Inter-individual variability of the H-reflex in normal subjects. *Electromyogr Clin Neurophysiol* 1981;21:183-200.