Clinical Medicine 2014 Vol 14, No 3: 264–7

Inter-rater agreement of observable and elicitable neurological signs

Authors: Mark Thaller[A] and Thomas Hughes[B]

ABSTRACT

This paper reports on a study that aimed to assess the inter-rater agreement of observable neurological signs in the upper and lower limbs (eg inspection, gait, cerebellar tests and coordination) and elicitable signs (eg tone, strength, reflexes and sensation). Thirty patients were examined by two neurology doctors, at least one of whom was a consultant. The doctors' findings were recorded on a standardised pro forma. Inter-rater agreement was assessed using the kappa (κ) statistic, which is chance corrected. There was significantly better agreement between the two doctors for observable than for elicitable signs (mean ± standard deviation [SD] κ, 0.70 ± 0.17 vs 0.41 ± 0.22, p=0.002). Almost perfect agreement was seen for cerebellar signs and inspection (a combination of speed of movement, muscle bulk, wasting and tremor); substantial agreement for strength, gait and coordination; moderate agreement for tone and reflexes; and only fair agreement for sensation. The inter-rater agreement is therefore better for observable neurological signs than for elicitable signs, which may be explained by the additional skill and cooperation required to elicit rather than just observe clinical signs. These findings have implications for clinical practice, particularly in telemedicine, and highlight the need for standardisation of the neurological examination.

KEYWORDS: Elicitable signs, inter-rater reliability, neurological examination, observable signs, telemedicine

Introduction

A neurological consultation comprises a verifiable history, a reliable examination, appropriate investigations and their subsequent interpretation. When a specialist opinion is sought using telemedicine, the remote clinician relies on another doctor's neurological examination. Some neurological signs have to be elicited by the examining physician, eg tone, strength and sensory deficits, but some valuable signs can be seen and heard by the remote and examining physician, eg walking, speed of finger movements and maintaining the outstretched arm in a particular posture.

The inter-rater reliability of the National Institutes of Health Stroke Scale (NIHSS, Table 1), which splits motor aspects into five groups – no drift (0), drift before 10 seconds (1), falls before 10 seconds (2), no effort against gravity (3) and no movement (4) – and the traditional neurological examination has been investigated before (Table 2), but these investigations did not include a comparison of signs categorised as elicitable or observable. These studies have not analysed their data according to whether the clinical signs could be observed from the end of the bed.

Telemedicine has been used to provide an out-of-hours stroke thrombolysis service to hospitals in south-east Wales since April 2012. We therefore investigated the inter-rater agreement of some elicitable and observable neurological signs in the upper and lower limbs to inform an assessment of their utility in the clinical examination performed using telemedicine.

Methods

Thirty patients (mean ± standard deviation [SD] age 55 ± 15 years) recruited over a 4-week period in a routine neurology outpatient clinic gave written consent to be examined by a consultant and, in the same clinic session, one other neurology doctor (foundation year 2, core medical trainee, specialty registrar or consultant). The second examiner, blinded to the findings of the first, repeated the examination of the upper and lower limbs. Examiners were asked to record their findings immediately on a standardised pro forma (Table 3) by selecting from binary options (eg present/absent for clonus) and categorical options (eg absent, depressed, normal or brisk for reflexes and Medical Research Council grades 0–5 for strength). Clinicians did not undertake any special training or instruction in clinical examination as part of this study and were asked to examine patients in accordance with their usual clinical practice, with appropriate equipment provided.

Inter-rater agreement was assessed using the κ statistic, which makes no assumptions about which doctor is correct – only whether they agree. The κ benchmarks used in this paper were that of Landis and Koch: <0 represents poor agreement, 0–0.20 slight agreement, 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement, 0.61–0.80 substantial agreement and 0.81–1.00 almost perfect agreement.[15] A significant difference in agreement is present if there is no overlap in the 95% confidence intervals for the κ value. The mean κ and *t*-test results were used to assess the significance of the difference between grouped data. The analysis was performed using Microsoft Excel 2007 spreadsheet software.

The study was part of a medical student placement and was approved by the North Wales Research Ethics (Central & East) Proportionate Review Sub-Committee (11-WA-0311) and the Cardiff and Vale Research and Development Department (11/CMC/5212).

Authors: [A]foundation year 1 doctor, Glangwili Hospital, Carmarthen, UK; [B]consultant neurologist, Department of Neurology, University Hospital of Wales, Cardiff, UK

Results

The results are summarised in Fig 1 and Table 4. The inter-rater reliability for observable signs was better than for elicitable signs (mean ± SD κ value 0.70 ± 0.17 vs 0.41 ± 0.22, p=0.002). We considered whether the difference between observable and elicitable signs was a consequence of the variable number of available options – for example, reflexes could be normal, brisk, reduced or absent but speed of movement could only be normal or slow. We therefore recalculated the inter-rater agreement for all data using a binary grouping – for example, reflexes could be abnormal (brisk, reduced or absent) or normal and strength could be abnormal (any grading ≤4) or normal (grade 5). The difference in the inter-rater agreement between observable and elicitable signs was still significant (mean ± SD κ value 0.76 ± 0.09 vs 0.46 ± 0.21, p=0.014).

Discussion

Signs that have to be elicited involve skill on the part of the examiner, the cooperation of the patient and then interpretation – for example, to test tone, the patient must be relaxed and comfortable and the examining doctor must have an understanding of the actions required to elicit the clinical features of spasticity and rigidity. Informal observation of the techniques used by different doctors in this study suggested marked variations in technique and interpretation, which may explain the poor inter-rater agreement. By comparison, it is more straightforward to observe patients at rest or when performing actions such as tapping movements of the finger and thumb to assess speed of movement or walking, which may explain the better agreement seen for these observable signs. Miller and Johnston[16] found foot tapping (κ=0.73) to be more reliable (sensitivity 86% and specificity 84%) for upper motor neurone weakness than Babinski testing (plantar reflex) (κ=0.30) (sensitivity 35% and specificity 77%).

The previous literature (see Tables 1 and 2) shows a wide variation in the elicitable signs, with the κ statistic value ranging from 0.29 to 1.00 (mean 0.65) for strength and from 0.15 to 1.00 (0.46) for sensation. The variation in reliability of the peripheral neurological examination in the literature, as well as with the results of this study, highlights that relying on another doctor's assessment may affect diagnosis and management.

One of the concerns of clinicians providing opinions about patients they are not able to examine in person is that their clinical examination is impoverished by the lack of direct patient contact. However, this study suggests that those signs that require elicitation have poorer inter-rater reliability than 'end-of-the-bed' signs, which can be observed by both the attending physician and the remote physician using telemedicine equipment. The importance of being a good noticer[17] is as relevant today as it ever

Table 1. Inter-rater reliability of NIHSS.

	Anderson (2013)[1] (n=83)	Demaershalk (2012)[2] (n=20)	Gonzalez (2011)[3] (n=40)	Meyer (2005)[4] (n=25)	Handschu (2003)[5] (n=41)	Meyer (2002)[6] (n=45)	Shafqat (1999)[7] (n=20)	Brott (1989)[8] (n=24)	Goldstein (1989)[9] (n=20)
Motor arm	1.00	0.79 (right)	0.74	0.82 (right)	0.90	0.96 (right)	0.82	0.85	0.77
	–	0.83 (left)	–	0.88 (left)	–	0.97 (left)	–	–	–
Motor leg	0.97	0.79 (right)	0.62	0.80 (right)	0.92	0.96 (right)	0.83	0.83	0.78
	–	0.79 (left)	–	0.74 (left)	–	0.95 (left)	–	–	–
Sensation	1.00	0.64	Not done	0.80	0.91	0.91	0.48	0.60	0.50
Limb ataxia	0.35	0.03	0.98	0.80	0.95	0.69	–0.07	0.57	–0.16

NIHSS = National Institutes of Health Stroke Scale.

Table 2. Inter-rater reliability of components of the traditional neurological examination.

Component	Carswell (2012)[10]	Hand (2006)[11]	Jepsen (2006)[12]	Lindley (1993)[13]	Shinar (1985)[14]
Tone	0.63	–	–	–	–
Strength	0.29	0.65 (arm)	0.54	0.77 (arm)	0.49 (right hand)
	–	0.72 (hand	–	0.68 (hand)	0.58 (left hand)
	–	0.57 (leg)	–	0.64 (leg)	–
Reflexes	0.18	–	–	–	–
Sensation	–	0.49	0.48–0.69	0.15 (arm)	0.50 (right hand)
	–	–	–	0.19 (hand)	0.32 (left hand)
Gait	0.63	0.62	–	–	–
Cerebellar	–	–	–	0.46	0.45
Coordination	0.60	–	–	–	–

Table 3. Proforma for examination of the limbs (replicated for each side).

Speed of movement	Normal	Slow						
Muscle bulk	Normal	Decreased						
Muscle wasting	No	Yes						
Tremor	No	Yes						
Chorea	No	Yes						
Fasciculations	No	Yes						
Pronator drift	No	Yes						
Tone	Normal	Decreased	Increased					
MRC strength score	5	4	3	2	1	0		
Biceps reflex	Normal	Depressed	Brisk	Absent				
Triceps reflex	Normal	Depressed	Brisk	Absent				
Supinator reflex	Normal	Depressed	Brisk	Absent				
Patellar reflex	Normal	Depressed	Brisk	Absent				
Achilles reflex	Normal	Depressed	Brisk	Absent				
Plantar reflex	Flexor	Mute	Extensor					
Clonus	Present	Absent						
Light touch	Normal	Decreased	Absent					
Pin prick	Normal	Decreased	Absent					
Vibration	Normal	Decreased	Absent					
Temperature	Normal	Decreased	Absent					
Proprioception	Normal	Decreased	Absent					
Finger nose coordination	Normal	Abnormal						
Heel shin coordination	Normal	Ataxic	Abnormal not ataxic					
Dysdiadokinesia	Normal	Abnormal						
Gait	Normal	Apraxic	Waddling	Ataxic	Festinant	Antalgic	High stepping	Spastic
Walk on heels	Normal	Abnormal						
Walk on toes	Normal	Abnormal						
Romberg's test	Normal	Abnormal						

MRC = Medical Research Council.

Table 4. Main components of the neurological examination of the limbs with combined data for each aspect.

Component	Grouped K value	Standard error	95% confidence interval	Number of sets of data points
Inspection	0.83	0.09	0.74 to 0.92	240
Tone	0.51	0.35	0.16 to 0.86	60
Power	0.63	0.21	0.42 to 0.84	60
Reflexes	0.57	0.12	0.45 to 0.68	150
Sensation	0.33	0.15	0.17 to 0.48	293
Gait	0.62	0.24	0.38 to 0.86	43
Coordination	0.78	0.21	0.57 to 0.99	60
Cerebellar	0.77	0.30	0.47 to 1.08	43

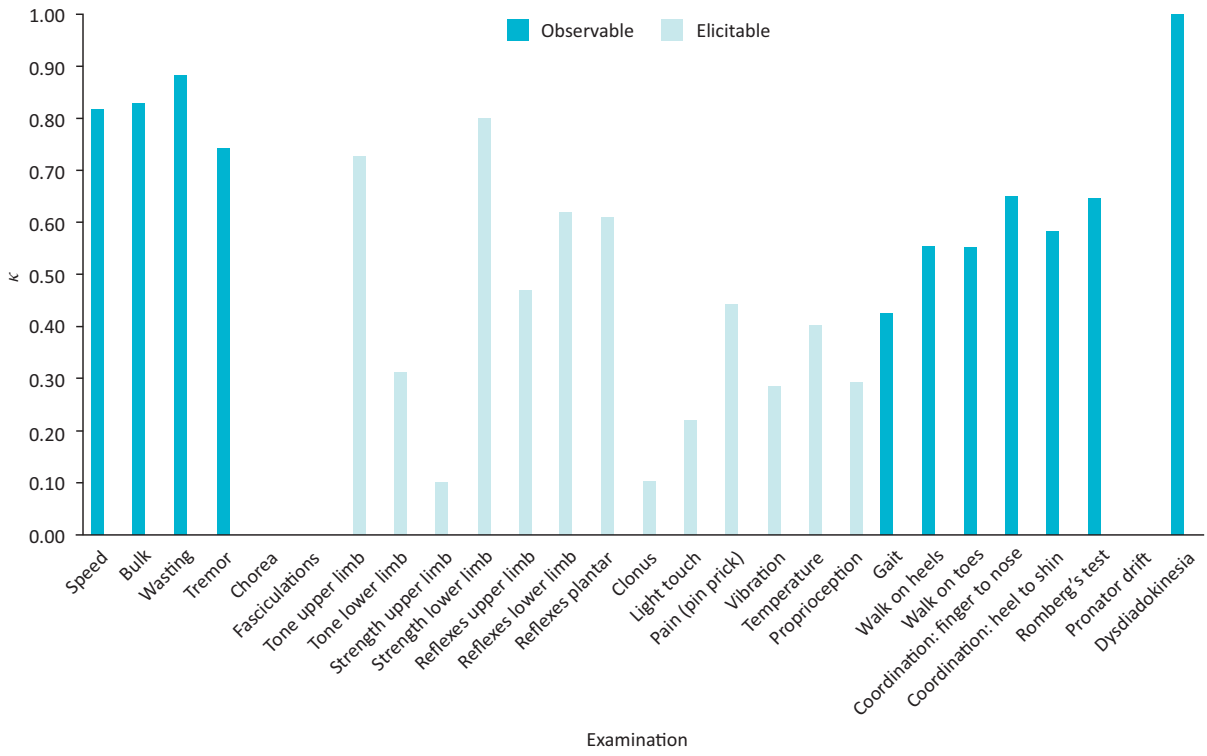

Fig 1. Agreement of neurological signs. Note: κ results for chorea, fasciculation and pronator drift are zero, because they were not observed by any of the clinicians.

was, and rather than compromising clinical skills, the technology of telemedicine may demand of clinicians a review of the parts of the clinical examination that are most reliable.

Conclusion

Observable neurological signs have significantly better inter-rater agreement than elicitable signs. These findings have implications for clinical practice, including telemedicine. ■

Acknowledgements

We would like to thank clinical colleagues in the department of neurology for their help and support. This work was first presented at the video conference All Wales Stroke Meeting (AWSM).

References

1 Anderson ER, Smith B, Ido M, Frankel M. Remote assessment of stroke using iPhone 4. *J Stroke Cerebrovascular Dis* 2013;22:340–4.
2 Demaerschalk BM, Vegunta S, Vargas BB *et al.* Reliability of real-time video smartphone for assessing National Institutes of Health Stroke Scale scores in acute stroke patients. *Stroke* 2012;43:3271–7.
3 Gonzalez MA, Hanna N, Rodrigo ME *et al.* Reliability of pre-hospital real-time cellular video phone in assessing the simplified National Institutes of Health Stroke Scale in patients with acute stroke: a novel telemedicine technology. *Stroke* 2011;42:1522–7.
4 Meyer BC, Lyden PD, Al-Khoury L *et al.* Prospective reliability of the STRokE DOC wireless/site independent telemedicine system. *Neurology* 2005;64:1058–60.
5 Handschu R, Littmann R, Reulbach U *et al.* Telemedicine in emergency evaluation of acute stroke: interrater agreement in remote video examination with a novel multimedia system. *Stroke* 2003;34:2842–6.
6 Meyer BC, Hemmen TM, Jackson CM, Lyden PD. Modified National Institutes of Health Stroke Scale for use in stroke clinical trials: prospective reliability and validity. *Stroke* 2002;33:1261–6.
7 Shafqat S, Kvedar JC, Guanci MM *et al.* Role for telemedicine in acute stroke: feasibility and reliability of remote administration of the NIH stroke scale. *Stroke* 1999;30:2141–5.
8 Brott T , Adams HP Jr, Olinger CP *et al.* Measurements of acute cerebral infarction: a clinical examination scale. *Stroke* 1989;20:864–70.
9 Goldstein L, Bertels C, Davis J. Interrater reliability of the NIH stroke scale. *Arch Neurol* 1989;46:660–2.
10 Carswell C, Rañopa M, Pal S *et al.* Video rating in neurodegenerative disease clinical trials: the experience of PRION-1. *Dement Geriatr Cogn Dis Extra* 2012;2:286–97.
11 Hand P, Haisma JA, Kwan J *et al.* Interobserver agreement for the bedside clinical assessment of suspected stroke. *Stroke* 2006;37:776–80.
12 Jepsen J, Laursen LH, Hagert CG *et al.* Diagnostic accuracy of the neurological upper limb examination I: Inter-rater reproducibility of selected findings and patterns. *BMC Neurology* 2006;6:8.
13 Lindley RI, Warlow CP, Wardlaw JM *et al.* Interobserver reliability of a clinical classification of acute cerebral infarction. *Stroke* 1993;24:1801–4.
14 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;42:557–65.
15 Landis J, Koch G. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–74.
16 Miller T, Johnston SC. Should the Babinski sign be part of the routine neurologic examination? *Neurology* 2005;65:1165–8.
17 Asher R. Clinical sense. *BMJ* 1960;1:985–93.

Address for correspondence: Dr M Thaller, Glangwili Hospital, Dolgwilli Road, Carmarthen, Carmarthenshire SA31 2AF. Email: mark.thaller@doctors.org.uk