Spinal Cord (2007) 45, 444–451

www.nature.com/sc

Original Article

Inter-rater reliability of motor and sensory examinations performed according to American Spinal Injury Association standards

G Savic*[,1], EMK Bergström[1], HL Frankel[1], MA Jamous[1] and PW Jones[2]

[1]*National Spinal Injuries Centre, Stoke Mandeville Hospital, Buckinghamshire Hospitals NHS Trust, Aylesbury, Bucks, UK;* [2]*Department of Mathematics, School of Computing and Mathematics, Keele University, Keele, UK*

Study design: Prospective observational.
Aim: To examine inter-rater reliability of motor and sensory examinations performed according to American Spinal Injury Association (ASIA) standards.
Setting: National Spinal Injuries Centre, Stoke Mandeville Hospital, Buckinghamshire Hospitals NHS Trust, UK.
Material and method: Results of ASIA motor and sensory examinations performed by two experienced examiners on 45 patients with spinal cord injury (SCI) were compared.
Results: Total ASIA scores showed very strong correlation between the two examiners, with Pearson correlation coefficients and intraclass correlation coefficients exceeding 0.96, $P<0.01$ for total motor, light touch and pin prick scores. The agreement for individual muscle testing of the 10 ASIA key muscles showed substantial agreement for majority of muscles, with the weighted Kappa coefficient range 0.649–0.993, $P<0.05$. The overall agreement in assignment of manual muscle testing grades (0–5) was 82% on the right and 84% on the left, with the strongest agreement for grade '0' and the weakest for grade '3'. The unweighted Kappa coefficient for agreement in motor and sensory levels ranged from 0.68 to 0.78 ($P<0.01$). There was no difference in ASIA impairment grades derived from the two examiners' results.
Conclusions: Our study results showed very good levels of agreement in ASIA clinical examinations between two experienced examiners. The established degree of variability due to inter-rater differences should be taken into account in study design of clinical trials with more than one assessor.
Sponsorship: Supported by the International Spinal Research Trust, UK, grant CLI001.
Spinal Cord (2007) **45,** 444–451; doi:10.1038/sj.sc.3102044; published online 27 March 2007

Keywords: spinal cord injury; ASIA standards; reliability; manual muscle testing

Introduction

This study was part of the International Spinal Research Trust (ISRT) Clinical Initiative study.[1] The aim of the Clinical Initiative was to develop a battery of clinical, functional and neurophysiological tests which could be used for monitoring efficacy of new therapeutic interventions in patients with spinal cord injury (SCI). This particular study examined the inter-rater reliably of the clinical neurological examination performed according to International Standards for Neurological Classification of Spinal Cord Injury[2] in view of future multicentre clinical trials with multiple assessors.

The sixth edition (Revision 2000) of the International Standards for Neurological Classification of Spinal Cord Injury is currently in use.[2] The Standards were developed by the American Spinal Injury Association (ASIA) for assessing the neurological deficit in patients with SCI and for classifying the injury. They are endorsed by the International Spinal Cord Society (ISCoS) and are used worldwide both in everyday clinical practice and in clinical research. The standards are accompanied by a reference manual, which gives detailed explanation on how to perform motor and sensory neurological examination and how to classify the SCI based on the results of the examination.[3]

Several studies examined inter-rater reliability of the previous versions of the ASIA Standards.[4–8] This led to clarification of identified problem areas and to improvement of each subsequent version of the standards.

The standards distinguish between the examination and classification as two separate skills. Most of the studies in the past examined variations in classification

*Correspondence: G Savic, National Spinal Injuries Centre, Stoke Mandeville Hospital, Aylesbury, Bucks HP21 8AL, UK

results between different examiners resulting from differences in classification skills.[4–7] Fewer looked at examinations skills and how they affect the final examination and classification results.[8–10]

The aim of this study was to test only examination skills in order to establish what level of agreement could be expected between the results of examinations carried out by two experienced examiners and to determine how differences in these results affect the final classification of injury. We also discuss the implications that inter-rater differences might have on clinical trials which use different components of the standards as outcome measures and in which more than one assessor perform serial neurological examinations according to ASIA standards.

Materials and methods

The study was approved by Aylesbury Vale Local Research Ethics Committee. All volunteers were given a written information sheet, a verbal explanation of the procedure and a chance to ask any questions before deciding whether to participate in the study. Those who volunteered to take part signed a consent form.

Sample

A total of 45 patients with SCI were assessed by two examiners. If both examiners were not available to perform a full ASIA assessment, the second examiner performed either motor or sensory part of the examination only. At the end of the study, of the 45 patients, 43 had a motor examination and 30 had a sensory examination performed by both examiners.

Procedure

Two examiners, a clinical scientist with medical background and a senior research physiotherapist, performed a motor and/or sensory examination according to the ASIA standards within 5 days of each other. Both examiners were experienced in the ASIA assessment before the study and additionally met several times at the beginning of the study in order to standardise their examination technique. They both had read Version 2000 of the ASIA standards and watched ASIA instruction videos together. The only departure from the ASIA instructions was the use of Neurotips for pinprick sensory testing, rather than safety pins, which are not used in clinical practice in the UK. Neurotips were specifically designed for clinical use and, similarly to safety pins, have a sharp and a blunt end and are disposable. For ethical reasons, only one examiner (GS) performed rectal examinations.

The aim of this study was to assess and compare only examination skills of the two examiners and see how they affected the final examination and classification results. To eliminate the inter-rater differences in classification skills, the classification of injury for all examinations was carried out by one examiner (GS), based on the results of every examination.

Analysis

Sample characteristics were presented using descriptive statistics.

Total ASIA motor and sensory scores were analysed using Bland and Altman's level of agreement (mean difference ± 2 SD of the mean difference),[11] Pearson correlation coefficient and intraclass correlation coefficient (ICC) with its confidence interval (CI).[12] The two-way mixed effects model ICC was used, where the subjects effect is random and the assessor effect is fixed and it is assumed that there is no interaction effect. The scale for interpretation of ICC values, according to Shrout,[13] defines the agreement as:

0–0.1 = virtually none
0.1–0.4 = slight
0.41–0.6 = fair
0.61–0.8 = moderate
0.81–1 = substantial

Kappa-statistics (percentage agreement corrected for chance) were used for calculating agreement for manual muscle testing (MMT) of individual muscles.[12] Both weighted Kappa coefficient and unweighted Kappa coefficient were used; the first one because it is the appropriate measure of agreement for an ordinal scale such as the 0–5 scale for MMT and the second to make our results comparable with those of a published study.[9] The scale for determining the level of agreement by Kappa-values according to Landis and Koch (1977, p 37), quoted in Dunn (1989), states:[12]

0 = poor
0.01–0.2 = slight
0.21–0.4 = fair
0.41–0.6 = moderate
0.61–0.8 = substantial
0.81–1 = almost perfect

The agreement in assigning a manual muscle-testing grade (0–5) for all tested muscles was expressed as percentage agreement between the examiners.

The agreement in motor and sensory level of injury and in ASIA impairment grade derived from the examination results was expressed as percentage agreement between the two examiners and, being a measure of agreement for nominal data, as unweighted Kappa coefficient.

Statistical programmes SPSS Version 13 and StatXact Version 4 were used for statistical analysis.

The initial motor results analysis – the agreement and correlation of total ASIA motor scores, included all the 43 patients who had motor examination performed by both examiners. To eliminate the influence of cases in which examiners would be expected to agree perfectly, the 21 patients with motor complete thoracic injury (motor score 50 by both examiners) were excluded from all subsequent motor results analyses. The remaining 22 patients had the total motor scores analyses repeated and also had the analysis of agreement in individual key muscles scores and MMT grades. The agreement in the motor level was only calculated for those patients whose

motor level could be derived from their motor examination; therefore, the patients with the level of injury above C4 and between T2 and L1 (whose motor level could only be derived from their sensory level) were excluded from this analysis. This left 15 patients for the motor level agreement analysis. All sensory results analyses were carried out on all the 30 patients who had sensory examination performed by both examiners.

The initial sample had so many patients with complete thoracic injury, because the wider Clinical Initiative study targeted mainly patients with thoracic injury.[1] Many of the patients from this sample also took part in other components of the Clinical Initiative.

Results

Sample characteristics

The sample consisted of 45 patients with SCI. The mean age was 40.3 years (range 18–72), 38 were men and seven women. The SCI was complete (ASIA grade A) in 24 patients, sensory incomplete (ASIA B) in four, ASIA C in four and ASIA D in 13. In 15 patients the injury was at the cervical level, in 29 thoracic and in one patient lumbar. The time since injury ranged from 3 months to 43 years.

Of the 45 patients, 43 had motor examination and 30 sensory examinations carried out by both examiners. Figure 1 shows the level and completeness of SCI for the motor and sensory groups and for the second motor group (motor 2) – the 22 patients left after the exclusion of cases with motor complete thoracic injury.

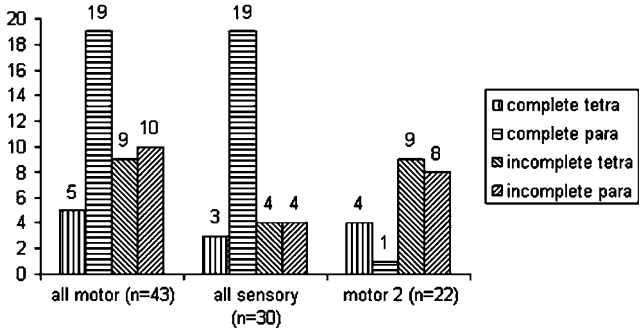

Figure 1 Level (tetra, para) and completeness (complete, incomplete) of injury in the 43 patients with motor examination and the 30 patients with sensory examination carried out by both examiners. The third group (motor 2), used in subsequent motor analyses, had 22 patients left after the exclusion of cases with motor complete thoracic injury. n = number of patients

Total ASIA scores

Table 1 shows the mean motor, light touch and pin prick scores by the two examiners, the score ranges and Bland and Altman's level of agreement (mean difference ± 2 SD of the mean difference).

Table 1 Mean ASIA motor, LT and PP scores and score ranges by the two examiners and Bland and Altman's level of aggreement between the two examiners

	Mean ASIA score (range)		*Bland and Altman's level of agreement (Mdiff ±2SD Mdiff)*
	GS	*EB*	
Motor ($n=43$)[a]	56.04 (7–90)	56.02 (10–90)	$0.02 \pm 2 \times 1.22$
Motor 2 ($n=22$)[b]	61.82 (7–90)	61.77 (10–90)	$0.45 \pm 2 \times 1.73$
LT ($n=30$)[c]	58.83 (23–94)	58.13 (22–96)	$0.7 \pm 2 \times 1.76$
PP ($n=30$)[c]	56.3 (21–94)	57.3 (21–96)	$1.0 \pm 2 \times 3.7$

Abbreviations: ASIA, American Spinal Injuries Association; EB, examiner 2; GS, examiner 1; LT, light touch; Mdiff, mean difference; *n*, number of patients; PP, pin prick; SD, standard deviation
[a]43 patients who had motor examination performed by both examiners
[b]22 patients remaining after exclusion of cases with complete paraplegia
[c]30 patients who had sensory examination performed by both examiners

Table 2 Total motor scores correlation between the two examiners

	r	P-*value*	*ICC*	*CI*	P-*value*
$n=43$[a]	0.999	<0.01	0.999	0.996–0.999	<0.001
$n=22$[b]	0.990	<0.01	0.998	0.994–0.999	<0.001
5/5 excluded[c]	0.987	<0.01	0.994	0.985–0.997	<0.001

Abbreviations: CI, confidence interval; ICC, intraclass correlation coefficient; *n*, number of patients; *r*, Pearson correlation coefficient
[a]43 patients who had motor examination performed by both examiners
[b]22 patients remaining after exclusion of cases with complete paraplegia
[c]muscles above the level of injury scored 5/5 by both examiners excluded from analysis

The total ASIA scores showed very strong correlation between the two examiners (Tables 2 and 3), with Pearson correlation coefficients (r) and ICC exceeding 0.99, $P<0.01$ for total motor and light touch scores and 0.97, $P<0.01$ for pin-prick scores. To eliminate the effect of dermatomes and myotomes with normal function on intra-rater agreement, the analysis was repeated after all the myotomes above the level of injury scored '5' by both examiners and all the dermatomes above the level of injury scored '2' by both examiners were excluded. The coefficients remained in the 'substantial' range even after this exclusion.

When the analysis was carried out by level and grade of injury, the agreement was better for thoracolumbar than for cervical level and for complete than for incomplete lesions, but still very strong for all subgroups, with all ICC > 0.9, $P<0.01$ and no statistically significant difference, determined by noting that the confidence intervals for the ICCs overlapped.

Analysis by myotomes

This analysis was carried out on the 22 patients left after exclusion of cases with motor complete thoracic injury. In the primary analysis, which included all tested myotomes (Table 4a), the agreement for individual muscle testing of the 10 ASIA key muscles showed substantial to almost perfect agreement for all the muscles (weighted Kappa coefficient 0.649–0.993, $P<0.01$, depending on the muscle tested). For the

Table 3 Total light touch and pin prick scores correlation between the two examiners

	r	P-*value*	*ICC*	*CI*	P-*value*
LT (n = 30)					
All dermatomes	0.994	<0.01	0.997	0.993–0.99	<0.001
2/2 excluded[a]	0.992	<0.01	0.996	0.992–0.998	<0.001
PP (n = 30)					
All dermatomes	0.978	<0.01	0.988	0.975–0.994	<0.001
2/2 excluded[a]	0.962	<0.01	0.980	0.958–0.990	<0.001

Abbreviations: LT, light touch; CI, confidence interval; ICC, intraclass correlation coefficient; *n*, number of patients; PP, pin prick; *r*, Pearson correlation coefficient
[a]dermatomes above the level of injury scored 2/2 by both examiners excluded from analysis

Table 4a Percentage agreement, unweighted and weighted Kappa coefficients for manual muscle testing of individual key muscles by the two examiners – primary analysis

Key muscle (myotome)	*Side*	N	*Inter-rater agreement measure*				
			%	*Kappa*	P-*value*	*WK*	P-*value*
Biceps brachii and brachialis (C5)	R	22	91	0.694	0.001	0.694	0.010
	L	22	91	0.730	0.000	0.649	0.000
Extensor carpi radialis (C6)	R	22	86	0.752	0.000	0.932	0.000
	L	22	91	0.823	0.000	0.973	0.000
Triceps brachii (C7)	R	22	86	0.765	0.000	0.969	0.000
	L	22	86	0.781	0.000	0.972	0.000
Flexor digitorum profundus (C8)	R	22	82	0.728	0.000	0.975	0.000
	L	22	73	0.629	0.000	0.946	0.000
Abductor digiti minimi (T1)	R	21	71	0.604	0.000	0.965	0.000
	L	22	77	0.752	0.000	0.963	0.000
Iliopsoas (L2)	R	22	73	0.593	0.000	0.963	0.000
	L	22	91	0.857	0.000	0.987	0.000
Quardiceps femoris (L3)	R	22	95	0.936	0.000	0.993	0.000
	L	22	91	0.882	0.000	0.987	0.000
Tibialis anterior (L4)	R	22	77	0.670	0.000	0.972	0.000
	L	22	82	0.742	0.000	0.948	0.000
Extensor hallucis longus (L5)	R	22	77	0.692	0.000	0.954	0.000
	L	22	77	0.688	0.000	0.961	0.000
Gastrocnemius and soleus (S1)	R	22	77	0.707	0.000	0.943	0.000
	L	22	73	0.664	0.000	0.951	0.000

Abbreviations: Kappa, unweighted Kappa coefficient; L, left; *N*, number of observations; R, right; WK, weighted Kappa coefficient; %, percentage agreement
All tested muscles were included in the primary analysis

Table 4b Percentage agreement, unweighted and weighted Kappa coefficients for manual muscle testing of individual key muscles by the two examiners – secondary analysis

Key muscle (myotome)	*Side*	N	*Inter-rater agreement measure*				
			%	*Kappa*	P-*value*	*WK*	P-*value*
Biceps brachii and brachialis (C5)	R	5	60	—	ns	—	ns
	L	5	60	—	ns	—	ns
Extensor carpi radialis (C6)	R	9	67	0.400	0.008	0.854	0.027
	L	7	71	0.417	0.034	—	ns
Triceps brachii (C7)	R	8	63	—	ns	—	ns
	L	8	63	0.415	0.026	0.785	0.017
Flexor digitorum profundus (C8)	R	9	56	—	ns	0.902	0.023
	L	9	33	—	ns	—	ns
Abductor digiti minimi (T1)	R	8	25	—	ns	0.823	0.011
	L	9	56	0.455	0.002	0.832	0.000
Iliopsoas (L2)	R	13	54	—	ns	—	ns
	L	16	88	0.770	0.000	0.972	0.000
Quardiceps femoris (L3)	R	16	94	0.895	0.000	0.981	0.000
	L	16	91	0.823	0.000	0.952	0.000
Tibialis anterior (L4)	R	13	62	0.444	0.002	0.908	0.000
	L	16	75	0.642	0.000	0.949	0.000
Extensor hallucis longus (L5)	R	15	67	0.490	0.003	0.879	0.000
	L	16	69	0.503	0.001	0.888	0.000
Gastrocnemius and soleus (S1)	R	15	67	0.545	0.000	0.821	0.000
	L	16	63	0.522	0.000	0.869	0.000

Abbreviations: Kappa, unweighted Kappa coefficient; L, left; *N*, number of observations; R, right; WK, weighted Kappa coefficient; %, percentage agreement
Muscles above the level of injury scored 5/5 by both examiners and muscles below the zone of partial preservation scored 0/5 by both examiners were excluded from the secondary analysis

secondary analysis all the myotomes above the level of injury scored '5' by both examiners and all the myotomes below the zone of partial preservation in complete SCI scored '0' by both examiners were excluded. In the secondary analysis (Table 4b), Kappa did not indicate statistically significant agreement in several myotomes because of the small number of observations. Where it did, the agreement was again substantial to almost perfect (weighted Kappa coefficient 0.785–0.981, $P<0.05$, depending on the muscle tested). Table 4a and b show percentage agreement, unweighted Kappa coefficient (Kappa) and weighted Kappa coefficient (WK) by myotomes for primary (4a) and secondary (4b) analysis.

Analysis by MMT grades

The overall agreement in assignment of MMT grades (0–5) between the two examiners was 82% on the right and 84% on the left side. The number of assignments and agreements for each MMT grade for the left and the right side are presented in Table 5. The strongest agreement was for grades '0' and '5' and the weakest for grades '2' and '3'. The secondary analysis of remaining muscles with grade '5' and '0' (after exclusion of myotomes above the level of injury scored '5' by both examiners and myotomes below the zone of partial preservation scored '0' by both examiners) showed weaker agreement for grade '5', but still very strong for grade '0'.

Level of injury and ASIA impairment grade

As mentioned in the methodology section, the classification of injury for all assessments was carried out by one examiner (GS) based on the written results of the two examiners, in order to eliminate inter-rater differences in classification skills.

The agreement in the motor level was only calculated for the patients whose motor level could be derived from their motor examination, that is, with level of injury C5-T1 and L2-S5. As there were no patients below the level of L1 in the whole sample, this left only 15 patients, with level of injury between C5 and T1, for motor level analysis. The agreement in sensory level was calculated for all the 30 patients who underwent sensory examination by both examiners.

Table 6 gives the percentage agreement and unweighted Kappa coefficient for motor and sensory level agreement on the right and on the left. The agreements ranged between 73 and 80% and all Kappa values were within the substantial agreement range.

In cases where the neurological levels were different between the examiners, the motor levels differed only by one level; in three cases on the right and in four cases on the left. The sensory levels differed by one segment in 11 cases (four on the right and 11 on the left) and by two segments in three cases (two on the right and one on the left).

The ASIA impairment grades based on the examination results of the two examiners were the same for every subject.

Table 5 Overall assignment of the manual muscle testing grades (0–5) by the two examiners and agreement between them

	MMT grade (0–5)							
	Analysis 1						*Analysis 2*	
	'0'	*'1'*	*'2'*	*'3'*	*'4'*	*'5'*	*0'*	*5'*
Right (R)								
Assignment 'GS' R (N)	49	13	9	11	63	74	1	11
Assignment 'EB' R (N)	50	12	11	7	64	75	1	14
Agreement R (N)	48	10	5	2	49	66	1	5
Agreement R (%)	97	80	50	22	77	89	100	40
Left (L)								
Assignment 'GS' L (N)	46	14	17	18	59	66	5	5
Assignment 'EB' L (N)	47	12	21	12	57	71	6	10
Agreement L (N)	46	8	10	8	48	64	5	5
Agreement L (%)	99	62	55	53	83	93	91	67

Abbreviations: EB, examiner 2; GS, examiner 1; L, left; MMT, manual muscle testing; *N*, number of observations; R, right; %, percentage agreement
Analysis 1: the first six columns (0–5) show agreement for all the tested myotomes in 22 patients
Analysis 2: the last two columns (0′ and 5′) show agreement for grades '0' and '5' in myotomes at and below the level of injury (after exclusion of myotomes above the level of injury scored '5' and myotomes below the zone of partial preservation scored '0' by both examiners)

Table 6 Percentage agreement and unweighted Kappa coefficient for motor ($n=15$) and sensory ($n=30$) level agreement on the right and on the left

Side	*Agreement in neurological level*			
	Number	*Percentage*	*Kappa*	P-*value*
Motor level				
Right	12/15	80%	0.76	$P<0.01$
Left	11/15	73%	0.68	$P<0.01$
Sensory level				
Right	24/30	80%	0.78	$P<0.01$
Left	22/30	73%	0.70	$P<0.01$

Discussion

The purpose of this study was to examine inter-rater reliability of the ASIA neurological examination between two well trained, experienced examiners and its implications in clinical trials with serial neurological examinations and more than one assessor.

We did not test the differences in classification skills between the examiners, as those can be eliminated by having all the classifications carried out by one person from properly completed ASIA neurological forms. What we did study was how results of examinations affected classification of injury, as changes in level and grade of injury are often used as outcome measures in clinical therapeutic trials.

Overall, our study showed a very strong agreement for both motor and sensory components of the neurological examination, even after exclusion of myotomes scored '0' and '5' and dermatomes scored '0' and '2' by both examiners.

For total ASIA scores, the agreement was slightly better for motor than for sensory scores, and better for light touch than for pin-prick scores, but still well in the 'substantial' range for all three scores (all ICCs >0.96, $P<0.01$). The examiners tended to display closer agreement when testing subjects with complete than incomplete injuries and subjects with thoracic than cervical level of injury, but none of the differences were statistically significant.

It is difficult to determine which of the 10 ASIA key muscles generated the best agreement, as the number of observations differed from one muscle to the next after the exclusion of muscles with grades '0' and '5' by both examiners, and Kappa coefficients did not reach statistical significance for all the myotomes because of the small number of observations. Keeping these limitations in mind, the *quardiceps femoris* muscle showed the strongest level of agreement both before and after the above-mentioned exclusion.

As expected, the strongest agreement was for MMT grades '0' and '5'; hence secondary analyses were performed after myotomes above the level of injury scored '5' and myotomes below the zone of partial preservation scored '0' by both examiners were excluded. The weakest agreement was found for MMT grade '3', followed by grade '2'. Differences in assigning those two particular grades have implications on ASIA impairment grade classification and could, in some cases, result in classifying the same injury as ASIA grade 'C' by one examiner's results and as ASIA grade 'D' by another's. Even though it made no difference to the ASIA impairment grade classification in our study, this should be kept in mind in clinical trials where change of one grade on the ASIA impairment scale is the main outcome measure.

Very few studies in the past addressed examinations skills of the ASIA Standards and are not fully comparable with ours.[8–10] They usually had more examiners, but fewer patients than our study, the ratio of patients with complete and incomplete injury was different between the studies, as was the level of injury and the statistical methods used.

Cohen and Bartko (1994)[8] examined reliability of the 1992 version of the Standards on 29 examiners from 19 centres (all sites for Fidia Farmaceutical Corporation's clinical trials). In this reliability study 18 patients were examined by three raters and 14 patients by two raters. The agreement for total ASIA scores was very strong, with ICC values of 0.96 for both light touch and pin-prick scores and ICC of 0.98 for the motor score. To prove that the high level of agreement was not due to testing mainly muscles with easily scored MMT grades, when all the muscles with grades '0' and '5' were dropped from the analysis, the motor score agreement was recalculated and it remained high (ICC = 0.95).

Marino *et al*[9] carried out a reliability study with 16 examiners and 16 patients in preparation for the Proneuron Phase II autologous incubated macrophage study for the treatment of acute SCI. They concluded that the inter-rater reliability of the total ASIA scores (motor ICC = 0.97, light touch ICC = 0.96 and pin prick ICC = 0.88) exceeded recommended values and that the measures were appropriately reliable for use in clinical trials involving serial neurological examinations with multiple examiners.

Jonsson *et al*[10] used unweighted Kappa coefficients to calculate agreement by individual myotomes and dermatomes in 23 patients assessed by four examiners. The majority of ASIA key muscles showed moderate to substantial agreement after the mid-study training procedure, whereas the agreement for pin prick and light touch by dermatomes was mostly in fair–moderate–substantial range.

The levels of agreement in our study were higher than in the above studies, but this would be expected in a study with only two examiners, both of whom were very experienced in ASIA neurological assessment. Our study is probably closest in study design to Cohen and Bartko[8] in the number of raters examining the same patient and in the exclusion of muscles with grade '0' and '5' from secondary analysis. However, we did not exclude all muscles with grades '0' and '5', just those below the zone of partial preservation scored '0' by both examiners and those above the injury level scored '5' by both examiners. This left in the analysis the muscles with grades '0' and '5' below the level of incomplete SCI and in the zone of partial preservation of complete injury, in which the examiners could be expected to disagree. The ICC coefficients of Cohen *et al*'s for total ASIA scores are close to ours, especially for examiners with more than 2 years experience. The ICC coefficients of Marino *et al*'s[9] for the total ASIA scores were also well within the 'substantial' agreement range, as were Cohen's and ours. From the results of these three studies it can be concluded that total ASIA scores are reliable outcome measures in clinical trials with more than one examiner. However, the established differences in total ASIA scores between examiners should be taken into account in clinical study design, as they give the range of measurement error (acceptable or not) within which it would not be possible to assert that there was a difference between two or more treatment groups due to the treatment effect.

The only study we found that had analysed agreement by individual myotomes and dermatomes was Jonsson *et al*,[10] who used unweighted Kappa coefficient as their agreement measure. We used the same unweighted Kappa for our analysis by myotomes (Table 4a and b) for comparison reasons, however the weighted Kappa is a more appropriate measure for MMT, which is carried out on an ordinal, six-point scale. Weighted Kappa takes into account not just the ratio of actual and possible agreement corrected for chance, but also the magnitude of disagreement, by weighting larger disagreements more and smaller disagreements less. It has been used in the past for measuring agreement of ordinal MMT scales,[14] including the Medical Research Council scale,[15] a modification of which is used in the ASIA motor testing. For illustration, we gave both weighted and unweighed Kappa coefficient values together with the percentage agreement in Table 4a and b. Compared with unweighted Kappa, the weighted Kappa values were higher for all myotomes except C5, reflecting the fact that most of our disagreements were of the magnitude of one MMT grade only.

One of the aims of our study was to establish how differences in examination results affect final classification of injury, hence we eliminated inter-rater classification differences by having all classifications carried out by one examiner. For this reason, our results of the neurological level and ASIA grade agreement are not comparable with the previous studies, which all examined either classification skills only[4–7] or a combination of examination and classification skills.[8,10]

On the basis of the results of the examinations by our two examiners, the final motor and sensory level classifications both showed strong agreement. Where different, levels of injury differed mainly by one segment and only in few cases of sensory level by two segments. These results suggest that, if using changes in motor and sensory level as outcome measures in clinical trials with more than one examiner, changes of this magnitude cannot be attributed to the treatment effect, as they may be due to inter-rater variability.

The differences in examination results between our two examiners were not large enough to affect the ASIA grade classification and there was a full agreement for ASIA impairment grade in all the patients. However, the number of patients within adjacent ASIA impairment grades in this study was too small to demonstrate that a change by a single ASIA grade is a reliable outcome measure in clinical trails with more than one examiner.

It should be emphasised once more that the levels of agreement presented in this study were between two very

experienced examiners who had additional pre-data collection meetings and discussions in order to minimise differences in their examination techniques. Before using different components of the ASIA standards as outcome measures in clinical trails with more than one examiner, it would be prudent for each research team to organise additional training and discussion sessions for the assessors and to establish their own degree of inter-rater variability.

Conclusions

Our study results showed very good levels of agreement between two experienced examiners in all components of the ASIA neurological examination. The results confirm that changes in total ASIA scores and in neurological levels of injury are reliable outcome measures in clinical trials with more than one examiner. The established degree of variability between examiners should be allowed for in study design of such trials, when determining clinically significant differences between groups in order to carry out a power calculation.

Acknowledgements

The study was supported by the International Spinal Research Trust (ISRT), UK, Grant CLI001.
We thank all the patient volunteers for their participation in the study.

References

1 Ellaway PH *et al*. Towards improved clinical and physiological assessments of recovery in spinal cord injury: a clinical initiative. *Spinal Cord* 2004; **42:** 325–337.
2 Marino RJ *et al*. ASIA Neurological Standards Committee 2002. International standards for neurological classification of spinal cord injury. *J Spinal Cord Med* 2003; **26**(Suppl 1)**:** S50–S56.
3 American Spinal Injury Association and International Medical Society of Paraplegia. (eds). *Reference Manual for the International Standards for Neurological Classification of Spinal Cord Injury*. American Spinal Injury Association: Chicago 2003.
4 Donovan WH, Wilkerson MA, Rossi D, Mechoulam F, Frankowski RF. A test of the ASIA guidelines for classification of spinal cord injuries. *J Neurol Rehabil* 1990; **4:** 39–53.
5 Priebe MM, Waring WP. The interobserver reliability of the revised American Spinal Injury Association standards for neurological classification of spinal injury patients. *Am J Phys Med Rehabil* 1991; **70:** 268–270.
6 Donovan WH, Brown DJ, Ditunno Jr JF, Dollfus P, Frankel HL. Neurological issues. *Spinal Cord* 1997; **35:** 275–281.
7 Cohen ME, Ditunno Jr JF, Donovan WH, Maynard Jr FM. A test of the 1992 International Standards for Neurological and Functional Classification of Spinal Cord Injury. *Spinal Cord* 1998; **36:** 554–560.
8 Cohen ME, Bartko JJ. Reliability of ISCSCI-92 for neurological classification of spinal cord injury. In: American Spinal Injury Association and International Medical Society of Paraplegia, (ed). *Reference Manual for the Standards for Neurological and Functional Classification of Spinal Cord Injury*. American Spinal Injuries Association: Chicago 1994, pp 59–65.
9 Marino RJ, Jones L, Kirshblum S, Tal J. Reliability of the ASIA motor and sensory examination. *J Spinal Cord Med* 2004; **27:** 194.
10 Jonsson M, Tollback A, Gonzales H, Borg J. Inter-rater reliability of the 1992 international standards for neurological and functional classification of incomplete spinal cord injury. *Spinal Cord* 2000; **38:** 675–679.
11 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; **1:** 307–310.
12 Dunn G. *Design and Analysis of Reliability Studies*. Edward Arnold: London 1989, pp 30–37.
13 Shrout PE. Measurement reliability and agreement in psychiatry. *Stat Meth Med Res* 1998; **7:** 301–317.
14 Frese E, Brown M, Norton B. Clinical reliability of manual muscle testing. Middle trapezius and gluteus medius muscles. *Phys Ther* 1987; **67:** 1072–1076.
15 Florence JM *et al*. Intrarater reliability of manual muscle test (Medical Research Council scale) grades in Duchenne's muscular dystrophy. *Phys Ther* 1992; **72:** 115–122; discussion 122–126.