#4

ORIGINAL CONTRIBUTION

Reliability and Repeatability of the Motor and Sensory Examination of the International Standards for Neurological Classification of Spinal Cord Injury

Ralph J. Marino, MD[1]; Linda Jones, PT, MS[2]; Steven Kirshblum, MD[3]; Joseph Tal, PhD[4]; Abhiijit Dasgupta, PhD[1]

[1]Thomas Jefferson University Hospital, Philadelphia, Pennsylvania; [2]Proneuron Biotechnologies, Los Angeles, California; [3]Kessler Institute for Rehabilitation, West Orange, New Jersey; [4]TechnoStat Ltd, Kfar Saba, Israel

Received February 7, 2007; accepted September 18, 2007

Abstract

Objective: To determine the reliability and repeatability of the motor and sensory examination of the International Standards for Classification of Spinal Cord Injury (SCI) in trained examiners.

Participants/Methods: Sixteen examiners (8 physicians, 8 physical therapists) with clinical SCI experience and 16 patients participated in a reliability study in preparation for a clinical trial involving individuals with acute SCI. After a training session on the standards, each examiner evaluated 3 patients for motor, light touch (LT), and pin prick (PP). The following day, 15 examiners reevaluated one patient. Inter-rater reliability was determined using intraclass correlation coefficients (1-way, random effects model). Intra-rater reliability was determined using a 2-way random effects model. Repeatability was determined using the method of Bland and Altman.

Results: Patients were classified as complete tetraplegia ($n = 5$), incomplete tetraplegia ($n = 5$), complete paraplegia ($n = 5$), and incomplete paraplegia ($n = 1$). Overall, inter-rater reliability was high: motor $= 0.97$, LT $= 0.96$, PP $= 0.88$. Repeatability values were small in patients with complete SCI (motor < 2 points, sensory < 7 points) but large for patients with incomplete SCI. Intra-rater reliability values were ≥ 0.98 for patients with complete SCI.

Conclusions: The summed scores for motor, LT, and PP in subjects with complete SCI have high inter-rater reliability and small repeatability values. These measures are appropriately reliable for use in clinical trials involving serial neurological examinations with multiple examiners. Further research in subjects with incomplete SCI is needed to determine whether repeatability is acceptably small.

J Spinal Cord Med. 2008;31:166–170

Key Words: Spinal cord injuries; Paraplegia; Tetraplegia; Neurological examination, motor, sensory; Outcomes; Reliability

INTRODUCTION

The International Standards for Classification of Spinal Cord Injury (the Standards) (1) are widely used for classifying and assessing patients with spinal cord injury (SCI) in both research and clinical settings. Previous studies on reliability of the motor and sensory examination in the Standards are limited and demonstrate variable results (2,3). Past versions of the Standards were revised to improve reliability of the assessment and classification. A major revision to the Standards was made in 1992, when the current key muscles were selected and light touch (LT) and pin prick (PP) sensory scores were added (4). Although the Standards were revised in 1996 and in 2000 (5,6) and the second edition of the reference manual (7) was published in 2003, the majority of changes involved the classification procedures rather than motor and sensory examination procedures. Therefore, studies of reliability of the motor and sensory examination using the Standards from 1992 or after would apply to the current version.

The requirements of an instrument depend on how it will be used. For a discriminative instrument, one meant to detect differences among individuals, reliability is important. For an evaluative instrument, one meant to detect change in function within individuals, responsive-

Please address correspondence to Ralph Marino, MD, Rehabilitation Medicine, Thomas Jefferson University Hospital, 132 S. 10th Street, 375 Main Bldg, Philadelphia, PA 19107; phone: 215.955.5756; fax: 215.955.5152 (e-mail: Ralph.marino@jefferson.edu).

This study was supported by Proneuron Biotechnologies.

ness is important (8). Reliability is generally determined using a correlation coefficient, such as the Pearson, Spearman, or intraclass correlation coefficient (ICC). These coefficients reflect the degree to which the scale is able to distinguish persons with different levels of the attribute being measured. The ICC is preferred over the Pearson coefficient because the latter does not detect systematic differences (all patients scoring 10% higher on the second evaluation) and therefore can overestimate reliability (9).

Agreement is related to but different from reliability. Agreement reflects the degree to which the same result is obtained by different raters or upon repeat testing in persons who have not changed. For individual items, agreement is generally evaluated using the kappa coefficient, which is a chance-corrected measure of agreement When there are several possible values for an item, for example the American Spinal Injury Association (ASIA) Impairment Scale (AIS) grades, a weighted kappa can be used. The weighted kappa gives partial credit for responses that are close to each other, rather than requiring the exact same response.

There are several proposed methods to evaluate responsiveness (sensitivity to change). One method is to calculate the effect size, which is defined as the difference in mean scores from baseline to follow up for subjects who have changed (based on another measure) divided by the standard deviation of baseline scores (9). This creates a standard unit of measurement that can be compared with change in other instruments. Guyatt proposed a "responsiveness statistic," which is based on the variability in scores of stable patients (9). This statistic is calculated by dividing the mean change in subjects who changed by the standard deviation of the change score in stable patients. Alternatively, if the smallest score change that was clinically significant is known, this value could be used as the numerator for the responsiveness statistic. Bland and Altman (10) described the repeatability of an instrument based on the within-subject standard deviation (SDw). The SDw is the square root of the residual mean square in a 1-way analysis of variance. The repeatability is $\sqrt{2} \times 1.96 \times$ SDw. The difference in 2 scores in a stable subject is expected to be less than 2.77 × SDw for 95% of pairs of observations. Beckerman et al (11) called this statistic the smallest real difference (SRD), which they defined as "the smallest measurement change that can be interpreted as a real difference."

Prior investigations of the reliability of the motor and sensory examination have focused on inter-rater reliability or agreement rather than on responsiveness or repeatability. Cohen et al (2) found high reliability of the LT, PP, and motor examinations; inter-rater reliability values ranged from 0.96 to 0.98 and intra-rater reliability values were 0.98 to 0.99 for the 3 scales. Jonsson et al (3) evaluated inter-rater agreement for individual sensory dermatome scores and muscle test scores but did not look at reliability of the entire scales. The purpose of the present study was to evaluate inter-rater reliability and repeatability of the LT, PP, and motor scores of the International Standards for Classification of Spinal Cord Injury.

METHODS

This was an inter-rater and intra-rater reliability study of the sensory and motor examination of the standards, conducted as part of a training session for a clinical trial of activated macrophages (Procord) in acute SCI. The training consisted of a half-day interactive teaching session by one of the authors (R.J.M.), followed by an inter-rater and intra-rater reliability assessment using inpatients and outpatients from the Kessler Institute for Rehabilitation over a 2-day period. The study was approved by the Institutional Review Board of the Kessler Medical Rehabilitation Research and Education Corporation.

Sixteen individuals (patients) with SCI from the inpatient ($n = 2$) and outpatient ($n = 14$) population at the Kessler Institute in West Orange, NJ, volunteered to participate. Patients consisted of 10 men and 6 women ranging in age from 18 to 65 years with the following injuries: complete tetraplegia ($n = 5$), motor incomplete tetraplegia ($n = 5$), complete paraplegia ($n = 5$), and motor incomplete paraplegia ($n = 1$).

Sixteen examiners (8 physicians and 8 physical therapists) with more than 2 years of experience in the field of SCI participated in the study. Each examined 3 patients on day 1 according to the 2002 Standards and the 2003 reference manual. Examiners rotated rooms after each examination, so that each patient was examined 3 times. Examiners and patients were arranged so that every examiner evaluated at least 1 patient with a neurologically complete injury and 1 with an incomplete injury. Rectal examinations were not performed. On day 2, 15 examiners evaluated 1 of the patients they had evaluated on day 1. Only 6 patients were available, 4 with complete injuries and 2 with incomplete injuries. No patient was examined more than 3 times on day 2. However, only 3 examinations were performed on patients with an incomplete injury. Total scores for LT, PP, and motor were obtained by adding the individual scores. Statistical analyses were conducted using R, a language and environment for statistical computing (http://www.R-project.org).

Inter-rater Reliability Analyses

Inter-rater reliability for LT, PP, and motor total scores was calculated using ICC (1-way, random effects model) on day-1 data. Because recent research indicates that the motor score should be divided into an upper extremity motor score (UEMS) and a lower extremity motor score (LEMS) (12), reliability of motor scores was repeated for these subscales. Reliability values of motor subscale scores were calculated using only data from patients with tetraplegia for UEMS and only patients with incomplete

Table 1. Descriptive Statistics of Day 1 Scores (n = 16)

Scale (Maximum Score)	Mean	SD	Median	Minimum	Maximum
Light touch total score (112)	50.7	25.1	48.8	17.7	93.3
Pin prick total score (112)	45.7	23.0	46.0	16.0	82.0
Total motor score (100)	45.7	19.4	50.0	20.0	87.0
Upper extremity motor score (50)*	36.5	12.7	38.5	20.0	50.0
Lower extremity motor score (50)*	9.2	15.5	0.0	0.0	45.3

*Values are based on average scores from day 1 examinations on each subject.

injuries for LEMS. This in effect eliminated patients with UEMS of 50 from UEMS calculations and patients with LEMS of zero from LEMS calculations. Repeatability of the LT, PP, and motor scores was calculated using the method of Bland and Altman (10). Repeatability equals $\sqrt{2} \times 1.96 \times \mathrm{SDw}$, where SDw is the square root of the residual mean square in a 1-way analysis of variance.

Intra-rater Reliability Analyses

Intra-rater reliability and repeatability were calculated for the 3 scales using the ICC (2-way, random effects model) and the Bland-Altman repeatability statistic, respectively. Analyses were limited to patients with complete injuries because there were insufficient data on incomplete patients (only 3 pairs of examinations).

RESULTS

Descriptive statistics for motor, LT, and PP scores for the 16 patients obtained on day 1 are found in Table 1. Scores covered most of the range of the 3 total scores. The median LEMS was zero due to the high number of patients with complete injuries tested, all of whom had no motor function in the lower extremities.

Inter-rater Reliability

Inter-rater reliability values for sensory and motor scores were very good to excellent in most cases, except for PP in incomplete patients (Table 2). For all patients, the ICC for total motor score was 0.98; for LT 0.96, and for PP 0.89. The reliability values for incomplete patients were lower, with wide confidence intervals, due to the small number of patients. The lower limits of the confidence interval for sensory scores in the patients with incomplete injuries were below 0.75, a value that some propose as a minimally acceptable level of reliability (13,14). Reliability values for physicians and therapists were similar.

Repeatability values (smallest real difference) for scores are found in Table 3. When retesting stable patients, any differences from baseline scores are expected to be no greater than these values in 95% of pairs of observations. As can be seen, repeatability values are small for complete injuries but large for incomplete injuries. As a percentage of total scale score, LT repeatability (21%) is almost twice as large as total motor score (12%), and PP is nearly 3 times as large (31%). For UEMS and LEMS the repeatability values are equal to 10% and 14% of total scale scores, respectively.

Intra-rater Reliability

Twelve examiners evaluated 4 patients with complete injuries on both days, with each patient having 3 pairs of examinations. Two patients had tetraplegia, and 2 had paraplegia. This allowed us to evaluate intra-rater reliability for sensory scores and UEMS (Table 4). Some improvement in repeatability can be seen for sensory scores when the same evaluator performs all examinations. It should be noted that one outlier value was dropped from the UEMS calculation. One evaluation sheet was felt to have a recording error, where “2” was recorded instead of “5” for normal strength, resulting in an 18-point difference in motor scores between examinations. All other evaluations by the examiner were consistent with those of other examiners.

Table 2. Inter-rater Reliability Coefficients

	All Patients		Complete Injury		Incomplete Injury	
Group	ICC	95% CI	ICC	95% CI	ICC	95% CI
Light touch	0.96	0.90–0.98	0.99	0.98–1.00	0.86	0.57–0.98
Pin prick	0.89	0.77–0.96	0.99	0.97–1.00	0.69	0.25–0.94
Total motor score	0.98	0.96–0.99	1.00	0.99–1.00	0.95	0.83–0.99
Upper extremity motor score (tetra)	0.96	0.88–0.99				
Lower extremity motor score					0.98	0.92–1.00

ICC, intraclass correlation coefficient.

Table 3. Repeatability Values (points)

	All	Complete	Incomplete
Light touch	15.0	5.7	23.3
Pin prick	22.1	6.4	35.2
Total motor score	7.4	1.9	11.8
Upper extremity motor score (tetra)	5.1		
Lower extremity motor score			6.9

DISCUSSION

This study demonstrates that the motor and sensory examination of the Standards can be reliable when conducted by trained examiners. PP scores had the lowest reliability, possibly because determining PP values is more difficult than LT because it requires distinguishing sharp from dull and the degree of sharpness compared with normal. Our results are comparable to those of Cohen et al (2), considering differences in sample size. In that study, inter-rater reliability values for LT, PP, and motor scores were 0.96, 0.96, and 0.98, respectively. The corresponding values for intra-rater reliability were 0.99, 0.98, and 0.99. In Cohen's study, reliability was tested using 29 examiners and 32 men with SCI; our study consisted of 16 examiners and patients. More recently, Savic et al had 2 experienced examiners test 45 patients with SCI and found ICC values above 0.98 for motor and sensory total scores (15). Mulcahey et al evaluated the intra-rater reliability of the motor and sensory data in children and youths (16). They found that the examination was not reliable or could not be done in children younger than 4 years, and children younger than 10 years were distressed by the PP examination, limiting the usefulness in these age groups. In children 4 years and older, ICC values were generally high, although results were inconclusive for total motor scores in children younger than 15 years due to wide confidence intervals.

We have estimated the magnitude of difference in scores required to represent a true change rather than measurement error. Our results indicate that the scales are very sensitive to change in complete patients but only moderately sensitive in incomplete patients. Jonsson et al (3) conducted an inter-rater reliability study of the 1992 ASIA standards, incorporating changes made in 1996. However, the study looked at agreement for individual sensory points and muscle grades only, not total scores, and found limited agreement among raters. Individual item scores would be expected to have lower reliability and agreement than scale scores, because random error tends to cancel out in the total scores (13). We agree that one should not place much importance in changes of an individual sensory score or in small changes in a muscle grade.

Savic et al found that 95% of repeat LT and motor scores would differ by less than 4 points, while PP scores would differ by less than 8 points (15). These values are better than what we obtained, possibly due to differences in methods and sample composition. Savic's study used only 2 examiners, whereas we used 16 different examiners. The examiners in Savic's study also spent more time establishing uniform techniques than in our study. Finally, the sensory reliability sample in Savic's study had a preponderance of complete subjects (22/30), which may have contributed to the better repeatability values.

This study has some limitations. Patients were drawn from volunteers, and the majority had complete injuries. Estimates of reliability in patients with incomplete injuries are therefore less precise than for patients with complete injuries, resulting in wide confidence intervals. The intra-rater reliability testing was performed over 2 days. Although it is generally desirable to have a longer period between testing to reduce the influence of memory on scoring, this was not possible in the setting of investigator training. However, given the large number of muscles and sensory points tested on day 1, we believe examiners were unlikely to remember exact scores the next day. The study only looked at reliability shortly after a training session. Whether reliability deteriorates over time is unknown. Finally, because of the small number of subjects, results could have been influenced by a patient with erratic responses or an examiner with poor reliability. The patient-examiner pairings did not permit separation of patient and examiner effects in the inter-rater reliability testing.

The study and analysis revealed some practical considerations for neurological testing, especially as it relates to documenting for research purposes. The current ASIA worksheet has very small boxes for

Table 4. Intra-rater Reliability and Responsiveness for Patients With Complete Injuries

	No. of Examiners	ICC	95% CI	Smallest Real Difference
Light touch	12	0.99	0.97–1.00	4.1
Pin prick	12	0.99	0.94–1.00	5.9
Upper extremity motor score* (tetra)	5	0.98	0.79–1.00	2.0

*One outlier value dropped. See text for details.
ICC, intraclass correlation coefficient.

recording scores, making it difficult at times to read the number. Error can also be introduced if the examiner also is responsible for filling out the worksheet. Individual scores may be confused unless each score is recorded immediately. It is possible that the examiner with the large motor score difference during reliability testing confused the motor and sensory scoring. Having just competed sensory testing, in which a score of 2 is normal, this examiner may have continued to write a "2" in the motor boxes instead of "5" to indicate normal. Therefore, it may be prudent to have an assistant available to record scores and be sure the correct score is legibly recorded in the correct location.

Error due to the examiner could not be separated from error related to the patient being examined in the current study. Training focused on the examiners. However, some patients have difficulty with sensory testing and require several trials in the same dermatome before deciding whether the stimulus is the same as or different from the face. Some patients are too stringent in their criteria, indicating differences in areas that are well above the injury level. It would be interesting to standardize a "training" session with patients—explaining in detail the process of testing and testing a few dermatomes above, at, and below the injury level—to determine whether this improved reliability.

CONCLUSION

Inter-rater reliability of the summed scores for LT can be high in trained examiners. In this study, the scores generally exceeded recommended reliability values; more training may be required to achieve acceptable reliability of PP scores. Repeatability values are reasonably small for patients with complete injuries. Additional studies involving patients with incomplete injuries are needed to determine more precise estimates of repeatability. It is important to keep in mind that experimental power (ie, the ability to show effectiveness in clinical trials) is greatly dependent on reliable measurement. Consequently, we suggest that all examiners in clinical trials be preassessed for reliability and corrective action be taken when minimal standards are not achieved.

REFERENCES

1. American Spinal Injury Association. *International Standards for Neurological Classification of Spinal Cord Injury.* Revised 2000, reprinted 2002. Chicago, IL: American Spinal Injury Association; 2002.
2. Cohen ME, Bartko JJ. Reliability of ISCSCI-92 for neurological classification of spinal cord injury. In: Ditunno JF, Donovan WH, Maynard FM, eds. *Reference Manual for the International Standards for Neurological and Functional Classification of Spinal Cord Injury.* Atlanta, GA: American Spinal Injury Association; 1994:59–66.
3. 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.
4. American Spinal Injury Association. *International Standards for Neurological and Functional Classification of Spinal Cord Injury.* Revised 1992. Chicago, IL: American Spinal Injury Association; 1992.
5. American Spinal Injury Association. *International Standards for Neurological Classification of Spinal Cord Injury.* Revised 1996. Chicago, IL: American Spinal Injury Association; 1996.
6. American Spinal Injury Association. *International Standards for Neurological Classification of Spinal Cord Injury.* Revised 2000. Chicago, IL: American Spinal Injury Association; 2000.
7. American Spinal Injury Association. *Reference Manual for the International Standards for Neurological Classification of Spinal Cord Injury*. Chicago, IL: American Spinal Injury Association; 2003.
8. Kirshner B, Guyatt G. A methodological framework for assessing health indices. *J Chron Dis.* 1985;38:27–36.
9. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. *Controlled Clin Trials.* 1991;12:142S–158S.
10. Bland JM, Altman DG. Measurement error. *BMJ.* 1996;312: 1654.
11. Beckerman H, Roebroeck ME, Lankhorst GJ, Becher JG, Bezemer PD, Verbeek ALM. Smallest real difference, a link between reproducibility and responsiveness. *Qual Life Res.* 2001;10:571–578.
12. Marino RJ, Graves DE. Metric properties of the ASIA motor score: subscales improve correlation with functional activities. *Arch Phys Med Rehabil.* 2004;85:1804–1810.
13. Streiner DL, Norman GR. *Health Measurement Scales: A Practical Guide to Their Development and Use.* 2nd ed. Oxford: Oxford University Press; 1995.
14. Hinderer SR, Hinderer KA. Principles and applications of measurement methods. In: DeLisa JA, ed. *Physical Medicine and Rehabilitation: Principles and Practice.* 4th ed. Philadelphia, PA: Lippincott Williams and Wilkins; 2005:1139–1162.
15. Savic G, Bergstrom EMK, Frankel HL, Jamous MA, Jones PW. Inter-rater reliability of motor and sensory examinations performed according to American Spinal Injury Association standards. *Spinal Cord.* 2007;45:444–451.
16. Mulcahey MJ, Gaughan J, Betz RR, Johansen KJ. The International Standards for Neurological Classification of Spinal Cord Injury: reliability of data when applied to children and youths. *Spinal Cord.* 2007;45:452–459.