19–26

Reliability of classification systems for subaxial cervical injuries

Authors Addison T Stone[1], Richard J Bransford[2], Michael J Lee[2], Marcelo D Vilela[3], Carlo Bellabarba[2], Paul A Anderson[4], Julie Agel[2]

Institution [1]Orthopedics International, Kirkland, WA, USA

[2] Department of Orthopaedics and Sports Medicine, University of Washington, Seattle, WA, USA

[3] Mater Dei Hospital, Belo Horizonte MG, Brazil

[4] Department of Orthopaedic Surgery and Rehabilitation, University of Wisconsin, Madison, WI, USA

Methodological principle	
Broad spectrum of patients with expected condition	
Adequate description of methods for replication	•
Blinded/independent comparison of tests/ interpretations	•
Evidence level	**II**

The definiton of the different level classes of evidence for netability studies is available on page 55.

ABSTRACT

Study design: Interobserver and intraobserver reliability

Objective: To measure and compare the interobserver and intraobserver reliability of the cervical spine injury severity score (CSISS), the subaxial injury classification (SLIC) and severity scale, and the Allen-Ferguson system in patients with subaxial cervical spine injuries presenting to the emergency department.

Methods: Five examiners independently reviewed c-spine x-rays (CT/MRI) of 50 consecutive patients with subaxial cervical-spine injuries. They classified each case using CSISS, SLIC, and the Allen-Ferguson system. Examiners also documented if they believed the case required surgical management. At least 6 weeks later, the above steps were repeated for ten randomly chosen cases.

Results: The interobserver and intraobserver reliability for the total CSISS and total SLIC score are excellent. There is poor interobserver reliability and excellent intraobserver reliability when a total kappa score is calculated using all 21 groups for the Allen-Ferguson system. With respect to surgical management decisions, the interobserver agreement is moderate and the intraobserver agreement is excellent.

Conclusions: There is no universally accepted classification scheme for subaxial cervical-spine injuries. A useful classification system must have excellent reliability to consistently and accurately describe injury patterns between different observers and allow for comparison across systems or cohorts. Both the CSISS and the SLIC and severity scale are promising classification systems with excellent interobserver and intraobserver reliability. Future studies will need to determine if their quantitative scores correlate with management and clinical outcomes.

No financial support was received for this study.
IRB approval received.

STUDY RATIONALE AND CONTEXT

The identification and appropriate treatment of subaxial cervical-spine injuries is essential to optimize outcomes. Injuries to the cervical spine are present in only 1%–3% of people who sustain blunt trauma; however, the morbidity and mortality associated with these injuries can be devastating [1, 2]. Numerous classification systems have been proposed to describe these injuries, predict stability, and dictate treatment; still, none of them are universally accepted [3–17]. The "ideal" classification system must have excellent interobserver and intraobserver reliability, quantify stability, predict prognosis, and dictate treatment. We rely on a universal classification system as a prerequisite for comparison of clinical outcomes across different techniques and researchers. Newer systems have started to attempt to quantify injuries on a continuum in the form of objectively obtainable injury severity scales instead of differentiating injuries into various subtypes. To date, no studies have simultaneously evaluated the CSISS and the SLIC as two examples of a severity scale for cervical-spine injuries, and the Allen-Ferguson system as the most representative example of a typical classification system with a phylogeny of injury categories.

OBJECTIVE

To measure and compare the interobserver and intraobserver reliability of CSISS, SLIC and the Allen-Ferguson system in patients with subaxial cervical-spine injuries.

METHODS

Study design: Interobserver and intraobserver reliability.

Inclusion criteria: Patients seen in the emergency department with significant subaxial cervical injury with adequate imaging showing the morphology on computed tomography (CT) and/or magnetic resonance imaging (MRI).

Exclusion criteria: Patients with spine fractures outside the subaxial cervical region were excluded.

Patient population: Fifty consecutive patients seen in the emergency department at Harborview Medical Center (Seattle, WA) from April 2007 to August 2007 meeting the inclusion criteria. CT was available for 100% of patients and MRI was available for 70% of patients.

Classification systems evaluated: (please see web appendix at www.aospine.org/ebsj for additional details)

- CSISS is an ordinal score which divides the subaxial cervical spine into four columns: anterior, posterior, right pillar (right lateral column), and left pillar (left lateral column) and takes into account fractures as well as ligamentous injuries. Each column is given a score from zero (no injury) to five (most significant injury possible to that column) based on the severity of injury (**Fig 1**). The total quantitative scored is determined by adding the scores for each column at a given level of injury for a maximum score of 20. If there are multiple levels of injury, the highest quantitative score is used after determining the score for each individual level of injury (**Figs 2a–b**) [4, 14].
- SLIC and severity scale is an ordinal score comprised of three components: (1) injury morphology as determined by the pattern of spinal column disruption on available imaging studies; (2) integrity of the discoligamentous complex (DLC) represented by both anterior and posterior ligamentous structures as well as the intervertebral disc, and (3) neurological status of the patient [16]. Higher scores represent more severe injuries (**Table 1, Figs 2a–b**). Although both CSISS and SLIC are based on injury morphology and the integrity of the DLC, only SLIC takes into account neurological status.
- The Allen-Ferguson system is based on mechanism of injury. Six different phylogenies (compressive flexion, vertical compression, distractive flexion, compressive extension, distractive extension, and lateral flexion) are evaluated. There are different stages, based on severity, within each phylogeny for a total of 21 different possible classification types. It is a nonordinal system that does not quantify severity or dictate treatment.

Assessment process: Patient studies were de-identified and a new identity number randomly assigned to facilitate reviewer blinding. The images were copied to DVDs and distributed to the reviewers. Original papers and quick reference guides describing each classification system were provided to five spine surgeons who independently reviewed cervical spine radiographs (CT and MRI). To determine neurological status for the SLIC, reviewers were provided documented physical examinations from

each patient's chart. The reviewers independently classified each patient's injuries for all three classification systems and recorded whether surgery was indicated. At least 6 weeks after the interobserver data were collected, 10 of the original 50 cases were randomly chosen for the intraobserver results. The above steps were then repeated. Reviewers were blinded to the results of the previous assessment.

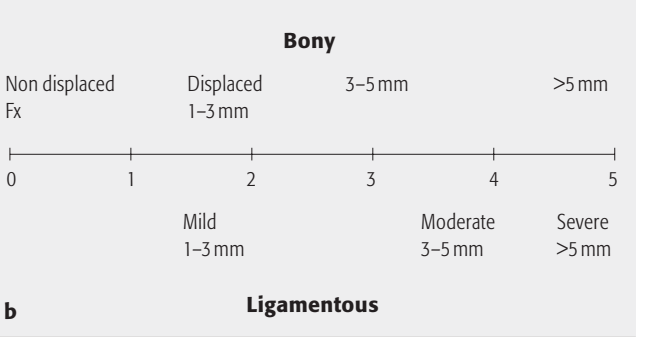
Analysis: Interobserver reliability for both CSISS and SLIC and severity scale was determined with intraclass correlation coefficient (ICC) using two-way random effects. ICC was used since we had more than two raters and because these systems are ordinal with higher scores representing more severe injuries. Interobserver reliability was calculated for the Allen-Ferguson system as well as for management of these injuries using kappa (INTER_RATER.MAC in SAS version 9.1.3 for Windows). We used kappa for the Allen-Ferguson classification since this is a nominal system with no natural ordering to the different phylogenies. Cohen's kappa was not used since it determines agreement between two raters only [18–20]. We considered ICC and kappa scores > 0.75 as excellent, 0.4–0.75 as moderate, and scores < 0.4 as poor [21].

Additional methodological and technical details are provided in the web appendix at www.aospine.org/ebsj.

RESULTS

- ***Interobserver variability:*** ICC values for CSISS and SLIC suggest excellent reliability; however, the kappa scores for the Allen-Ferguson system and management decisions were within the range of moderate to poor reliability.
  - For CSISS, the ICCs were 0.92, 0.94, 0.92, and 0.93 for the anterior column, posterior column, right pillar, and left pillar, respectively; with ICC for the total CSISS of 0.96.
  - For SLIC, the ICC values for injury morphology, DLC, and neurological status were 0.86, 0.90, and 0.98, respectively; and 0.79 for the total score.
  - For the Allen-Ferguson system, the overall kappa values for each of the six phylogenies are listed in **Table 2**. When a total kappa score was calculated using all 21 groups the value was 0.34. When it was determined from only the six main phylogenies it increased to 0.50.

Fig 1 The linear graph represent a 0-5 points severity of injury scale for bony cervical-spine injuries (top half) and ligamentous injuries (bottom half). There are general descriptors added to aid the clinician in attributing the most fitting point scale to the injury present.



Reprinted with permission from **Anderson PA, Moore TA, Davis KW, et al** (2007) Cervical spine injury severity score: assessment of reliability. *J Bone Joint Surg Am;* 89(5):1057–1065.

Fig 2a An axial image of a C7 burst fracture.

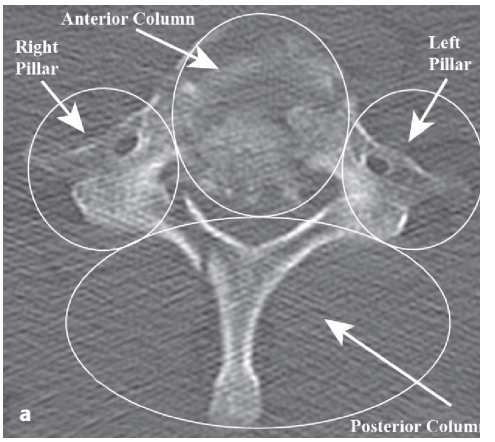
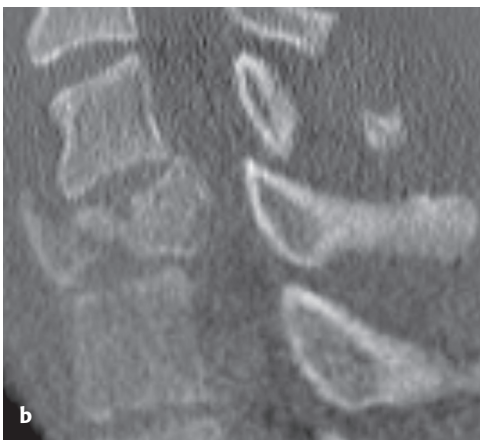


Fig 2b A sagittal view of C7 from the same patient.



  - The interobserver kappa agreement for management of the 50 cases was 0.55.
- ***Intraobserver variability:*** ICC and kappa values all suggested excellent intrarater reliability.
  - For CSISS, the ICCs were 0.94, 0.98, 0.92, and 0.98 for the anterior column, posterior column, right pillar, and left pillar, respectively; and 0.98 for the total CSISS.
  - For SLIC, the ICC values for injury morphology, DLC, and neurological status were 0.94, 0.94, and 0.99, respectively; and 0.98 for the total SLIC score.
- Using all 21 groups for the Allen-Ferguson system, the kappa value was 0.91. The intraobserver kappa agreement for management of the 50 cases was 0.91 (**Table 3**).

AN ILLUSTRATIVE CASE

The images in Figs 2a and b demonstrate how the CSISS and SLIC scores are calculated. The anterior column of this patient is displaced more than 5 mm, so according to the analog scale of the CSISS it would receive a score of 5. Both the right and lateral pillars have no injuries so they are assigned a score of 0. The posterior column has mild displacement (~2 mm), scoring a value of 2. Therefore, the total CSISS score for all four columns is 7. If there were multiple levels of injury within the subaxial cervical spine, the level with the highest total CSISS score would be used. With respect to the SLIC scale theit would receive a score of 2 for morphology, since it is a burst-type fracture. The discoligamentous complex is intact (this was confirmed also on MRI), scoring 0 for that component. The final component of the SLIC score is neurological status. The examinations from the patient's chart show that the patient had 0/5 strength in the lower extremities and no sensation below T2 including absent perianal sensation. Based on these clinical findings the score for neurological status would be a 2 as the patient has a complete cord injury. Therefore, the total SLIC score is 4.

Table 1 SLIC scale

	Points
Morphology	
No abnormality	0
Compression	1
Burst	+1=2
Distraction (eg, facet perch, hyperextension)	3
Rotation/translation (eg, facet dislocation, unstable teardrop or advanced staged flexion compression injury)	4
Discoligamentous complex (DLC)	
Intact	0
Indeterminate (eg, isolated interspinous widening, MRI signal change only)	1
Disrupted (eg, widening of the disc space, facet perch, or dislocation)	2
Neurological status	
Intact	0
Root injury	1
Complete cord injury	2
Incomplete cord injury	3
Continuous cord compression in setting of neurodeficit ('neuromodifier')	+1

SLIC=subaxial injury classification
MRI = magnetic resonance imaging

Table 2 Interobserver reliability

Measure	Intraclass correlation	Kappa
Cervical spine injury severity score (CSISS)		
Anterior column	0.93	NA
Posterior column	0.94	NA
Right pillar	0.92	NA
Left pillar	0.93	NA
Total CSISS	0.96	NA
Subaxial injury classification (SLIC) and severity scale		
Injury morphology	0.86	NA
DLC	0.90	NA
Neurological status	0.98	NA
Total SLIC	0.79	NA
Allen-Ferguson system		
Compressive flexion	NA	0.52
Vertical compression	NA	0.61
Distractive flexion	NA	0.54
Compressive extension	NA	0.34
Distractive extension	NA	0.63
Lateral flexion	NA	-0.16
Total (six phylogenies, listed above)	NA	0.50
Total (all 21 stages)	NA	0.34
Management (operative versus nonoperative)	NA	0.55

NA = not available
DLC = discoligamentous complex

Table 3 Intraobserver reliability

Measure	Intraclass correlation (range)	Kappa
Cervical spine injury severity score (CSISS)		
Anterior column	0.94 (0.85–1.00)	NA
Posterior column	0.98 (0.78–1.00)	NA
Right pillar	0.92 (0.30–0.97)	NA
Left pillar	0.98 (0.72–1.00)	NA
Total CSISS	0.98 (0.78–0.99)	NA
Subaxial injury classification (SLIC) and severity scale		
Injury morphology	0.94 (0.26–1.00)	NA
DLC	0.94 (0.47–1.00)	NA
Neurological status	0.99 (0.94–1.00)	NA
Total SLIC	0.98 (0.61–1.00)	NA
Allen-Ferguson system		
Total (all 21 stages)	NA	0.91 (0.30–0.99)
Management (operative versus nonoperative)	NA	0.91 (0.36–1.00)

NA = indicates not available
DLC = discoligamentous complex

DISCUSSION

- Reliability has been independently reported for both CSISS and SLIC and severity scale; however, these two classification systems have not been compared in the same study [4, 14, 16].
- Studies of CSISS: Moore et al. [14] determined the interobserver and intraobserver reliability, reporting a mean ICC for interobserver reliability of 0.88 for all four columns and the total CSISS. The ICC for intraobserver reliability was equal to or greater than 0.97 for all cases. This is consistent with our results. Anderson et al. [4] calculated an ICC of 0.82 for the anterior column; 0.76, for the posterior column; 0.79, for the right pillar; 0.74 for the left pillar; and 0.88 for the total CSISS. Our results showed superior interobserver reliability with a calculated ICC of 0.92 for the anterior column; 0.94, for the posterior column; 0.92, for the right pillar; 0.93, for the left pillar; and 0.96 for the total CSISS. They reported a mean intraobserver ICC of 0.98 for the total CSISS, which is identical to the score we obtained.
- Studies of SLIC: Vaccaro et al. evaluated the interobserver and intraobserver reliability of the SLIC severity scale and compared them with both the Harris and Allen-Ferguson system using both CT and MRI images [16]. They calculated an interobserver ICC of 0.57 for injury morphology; 0.49 for the DLC; 0.87 for neurological status; and 0.71 for total SLIC. Our results showed an improved ICC for all categories. They calculated an intraobserver ICC of 0.75 for injury morphology; 0.66, DLC; 0.90, neurological status; and 0.83, total SLIC. Our study had an intraobserver ICC of more than 0.93 for all three categories and the total SLIC score.
- Allen-Ferguson system evaluation: Vaccaro et al. determined the interobserver and intraobserver reliability and reported a total interobserver Cohen kappa of 0.53 (moderate interobserver agreement) [16]. Our results echo these results to a considerable degree. Our total kappa score for the six phylogenies was 0.50 (moderate agreement), when all 21 stages were considered the interobserver rating dropped to 0.34 (poor agreement).
- Strengths: Five surgeons reviewed more cases than reported in previous studies, which may have increased our statistical power. Use of two different methods to determine the interobserver and intraobserver reliability of the Allen-Ferguson system allowed us to evaluate the interobserver reliability of both the 21 stages and six main phylogenies, the results of which suggest better agreement when the later approach is taken. Unlike the Moore et al and An-

derson et al studies, where MRIs were not given to the reviewers, our reviewers were able to review MRIs for 70% of the cases which may have improved interobserver and intraobserver reliability. Finally, we calculated kappa from INTER_RATER.MAC in SAS because this algorithm has been shown to be superior when there are more than two raters [18, 19, 21].

- Limitations: MRI was only available for 70% of patients. Patients with neurological deficits were more likely to have obtained an MRI. The extent to which this would influence classification is not clear and further study may be warranted. Another limitation of this study is that the reviewers were from institutions that treat a high volume of spine trauma. The reliability obtained among observers with less experience in evaluating and treating spine trauma may not be consistent with our findings.
- Clinical relevance and impact: CSISS and SLIC and severity scale have excellent interobserver and intraobserver reliability. Accurate classification of injuries is important to determining optimal treatment. The "ideal" classification system must have excellent interobserver and intraobserver reliability, quantify stability, predict prognosis, and dictate treatment.
- Future studies: The next step is to determine if the quantitative scores of CSISS and SLIC correlate with management of subaxial cervical-spine injuries and clinical outcomes.

SUMMARY AND CONCLUSIONS

- Currently, there is no universally accepted and reproducible classification system for subaxial cervical-spine injuries.
- We report excellent interobserver and intraobserver reliability with use of CSISS and the SLIC and severity scale.
- Based on all 21 stages of the Allen-Ferguson system, the interobserver reliability is poor and the intraobserver reliability is excellent.
- There is moderate interobserver reliability and excellent intraobserver reliability with respect to management of subaxial cervical-spine injuries.

REFERENCES

1. **Kwon BK, Vaccaro AR, Grauer JN, et al** (2006) Subaxial cervical spine trauma. *J Am Acad Orthop Surg;* 14(2):78–89.
2. **Lowery DW, Wald MM, Browne BJ, et al** (2001) Epidemiology of cervical spine injury victims. *Ann Emerg Med;* 38(1):12–16.
3. **Allen BL Jr, Ferguson RL, Lehmann TR, et al** (1982) A mechanistic classification of closed, indirect fractures and dislocations of the lower cervical spine. *Spine;* 7(1):1–27.
4. **Anderson PA, Moore TA, Davis KW, et al** (2007) Cervical spine injury severity score: assessment of reliability. *J Bone Joint Surg Am;* 89(5):1057–1065.
5. **Apley AG** (1970) Fractures of the spine. *Ann R Coll Surg Engl;* 46(4):210–223.
6. **Babcock JL** (1976) Cervical spine injuries: diagnosis and classification. *Arch Surg;* 111(6):646–651.
7. **Beatson TR** (1963) Fractures and dislocations of the cervical spine. *J Bone Joint Surg Br;* 45:21–35.
8. **Bohlman HH** (1979) Acute fractures and dislocations of the cervical spine: an analysis of three hundred hospitalized patients and review of the literature. *J Bone Joint Surg Am;* 61(8):1119–1142.
9. **Harris JH Jr, Edeiken-Monroe B, Kopaniky DR** (1986) A practical classification of acute cervical spine injuries. *Orthop Clin North Am;* 17(1):15–30.
10. **Holdsworth FW** (1963) Fractures, dislocations, and fracture-dislocations of the spine. *J Bone Joint Surg Br;* 45B:6–20.
11. **Holdsworth FW** (1970) Fractures, dislocations, and fracture-dislocations of the spine. *J Bone Joint Surg Am;* 52(8):1534–1551.
12. **Jones RW** (1934) The treatment of fractures and fracture dislocations of the spine. *J Bone Joint Surg Am;* 16:30–45.
13. **King DM** (1967) Fractures and dislocations of the cervical part of the spine. *Aust N Z J Surg;* 37(1):57–64.
14. **Moore TA, Vaccaro AR, Anderson PA** (2006) Classification of lower cervical spine injuries. *Spine;* 31(11 Suppl):S37–43; discussion S61.
15. **Nicoll EA** (1949) Fractures of the dorso-lumbar spine. *J Bone Joint Surg Br;* 31B:376–394.

16. **Vaccaro AR, Hulbert RJ, Patel AA, et al** (2007) The subaxial cervical spine injury classification system: a novel approach to recognize the importance of morphology, neurology, and integrity of the disco-ligamentous complex. *Spine;* 32(21):2365–2374.
17. **Whitley JE, Forsyth HF** (1960) The classification of cervical spine injuries. *Am J Roentgenol Radium Ther Nucl Med;* 83:633–644.
18. **Gwet K** (2001) Handbook of Inter-rater Reliability: How to Estimate the Level of Agreement Between Two or MultipleRraters. Gaithersburg, Maryland: Stataxis.
19. **Gwet K** (2002) Computing inter-rater reliability with the SAS system. *Statistical methods for inter-rater reliability assessment, No. 3.*
20. **McGinn T, Wyer PC, Newman TB, et al** (2004) Tips for learners of evidence-based medicine: 3. Measures of observer variability (kappa statistic). *CMAJ;* 171(11):1369–1373.
21. **Shrout PE, Fleiss JL** (1979) Intraclass correlations: uses in assessing rater reliability. *Psychol Bull;* 86(2):420–428.

EDITORIAL STAFF PERSPECTIVE

As the authors point out, in order for any type of classification scheme to be useful, in addition to measuring what it is intended to measure (validity), the measurements need to be reproducible (reliability). This well done study does a thorough and thoughtful job of evaluating interobserver and intraobserver variability of the measures in a consecutive group of patients presenting to a regional tertiary care trauma center. The primary methodological strengths of this study include attention to details of study blinding, random selection of cases for second review, interpretation of the second review for intrarater reliability without prior knowledge of the previous reading, and ensuring that sufficient time had elapsed between reviews to avoid influence of the first interpretation on the results of the second. The authors also, for the first time to our knowledge, compare different 'severity scales,' which are starting to replace the more traditional 'classification systems.' These severity scales emphasize the continuum of any given disease by using a point system rather than attempting to separate disorders into artificially created categories ('phylogenies'). The reviewers applaud the authors in taking an important step beyond comparing traditional classification systems and looking into the potential applications of severity scales for everyday use.

Methods: *An important question that needs to be addressed in reliability studies is: Will these measures be reproducible across a range of severity conditions and among reviewers of different experience levels or schooled in different assessment strategies? For reliability studies, the study population generally should comprise those with a broad spectrum of the suspected condition who are likely to have the measure applied now or in the future. For instance, differences of body habitus and condition characteristics may influence measurements and the ability to reproduce the results. If conducted in a population primarily composed of those with known or severe disease, a classification scheme may give different results compared with studies on a group of more healthy individuals/less severe disease and may not give an accurate picture of overall reproducibility across condition severity. If patients with less severe disease image differently than those with more severe disease, this could affect the interpretation of x-rays and classification. The fact that the authors used selected consecutive patients somewhat increases the possibility that those with less severe as well as more severe injury are included. However, the authors also point out that the 70% of patients who had MRI were more likely to have had neurological deficits. The range of severity conditions is not described in this study, so the extent to which these scales are reproducible across ranges of severity is not clear.*

Some indication of the breadth of condition severity in study populations provides important information regarding generalizing the results to other settings as well. One could ask: In my setting, with the range of patients I see (outside of a regional trauma center) will I have the same reproducibility in applying these measures?

Answering these questions may be a helpful step in further establishing these as the appropriate measures for assessing patients with subaxial cervical-spineinjuries in addition to the next steps the authors suggest, namely evaluating the correlation of quantitative CSISS and SLIC scores with management decision and clinical outcomes. This study takes a big step forward in supporting the use of severity scales over the more traditional classification systems, such as the Allen-Ferguson system for the challenging topic of subaxial cervical-spine trauma.