Global Spine Journal
2020, Vol. 10(6) 682-691
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2192568219868218
journals.sagepub.com/home/gsj

External Multicenter Study of Reliability and Reproducibility for Lower Cervical Spine Injuries Classification Systems—Part 1: A Comparison of Morphological Schemes

Andrey Grin, MD, PhD[1,2], Vladimir Krylov, MD, PhD[1,2], Ivan Lvov, MD, PhD[1], Aleksandr Talypov, MD, PhD[1], Dmitriy Dzukaev MD[3], Anton Kordonskiy, MD, PhD[1], Vladimir Smirnov, MD, PhD[1], Vasily Karanadze, MD[1], Boburmirzo Abdukhalikov, MD[1], Ulugbek Khushnazarov, MD[1], Irina Aleynikova, MD, PhD[1], Elza Kazakova, MD[1], Olesya Bogdanova, MD[1], Alexander Peyker, MD[3], Vitaliy Semchenko, MD[3], Andrey Aksenov, MD[3], Anton Borzenkov, MD[3], Vladimir Gulyy, MD[3], Soslan Torchinov, MD[3], Sergey Bagaev, MD[3], Anton Toporskiy, MD[3], Andrey Nikitin, MD, PhD[2], Sevak Arakelyan, MD[4], Avetik Martikyan, MD[4], Stanislav Oshchepkov, MD[5], Dmitriy Hovrin, MD, PhD[6], Aslan Kojev, MD[7], and Musheg Khalatyan, MD[7]

Abstract

Study Design: Multicenter observational survey study.

Objectives: To quantify and compare the inter- and intraobserver reliability of Allen-Fergusson (A-F), Harris, Argenson, and AOSpine (AOS) classifications for cervical spine injuries, in a multicentric survey of neurosurgeons with different levels of experience.

Methods: We used data of 64 consecutive patients. Totally, 37 surgeons (from 7 centers), were included in the study. The initial assessment was returned by 36 raters. The second assessment performed after 1.5 months included 24 raters.

Results: We received 15 111 answers for 3840 evaluations. Raters reached a fair general agreement of the A-F scale, while the experienced group achieved $\kappa = 0.39$. While all groups showed moderate interrater reliability for primary assessment of Harris scale ($\kappa = 0.44$), the κ value for experts decreased from 0.58 to 0.49. The Argenson scale demonstrated moderate and substantial agreement among all raters ($\kappa = 0.47$ and $\kappa = 0.55$, respectively). The AOS scheme primary assessment general kappa value for all types of injuries and across all raters was 0.49, reaching substantial agreement among experts ($\kappa = 0.62$) with moderate agreement across beginner and intermediate groups ($\kappa = 0.48$ and $\kappa = 0.44$, respectively). The second assessment general agreement kappa value reached 0.56.

[1] Sklifosovsky Research Institute of Emergency Care, Moscow, Russia
[2] Evdokimov Moscow State University of Medicine and Dentistry, Moscow, Russia
[3] Moscow Spine Center of City Hospital No. 67, Moscow, Russia
[4] Moscow City Hospital No. 13, Moscow, Russia
[5] Pirogov National Medical and Surgical Center, Moscow, Russia
[6] Moscow City Hospital No. 7, Moscow, Russia
[7] Moscow City Hospital No. 1, Moscow, Russia

Corresponding Author:
Ivan Lvov, Sklifosovsky Research Institute of Emergency Care, B. Suharevskaya Pl. 3, Moscow 107945, Russia.
Email: dr.speleolog@gmail.com

Conclusions: We found the highest values of interobserver agreement and reproducibility among surgeons with different levels of experience with Argenson and AOSpine classifications. The AOSpine scale additionally incorporated more detailed description of compression injuries and facet-joint fractures. Agreement levels reached for Allen-Fergusson and Harris scales were fair and moderate, respectively, indicating difficulty of their application in clinical practice, especially by junior specialists.

Keywords

Allen-Fergusson classification, Harris classification, Argenson classification, AOSpine classification, cervical spine trauma, interobserver agreement, intraobserver agreement, Fleiss' kappa

Introduction

The first classification system for lower cervical spine injuries was developed by Bohler in 1929.[1] The author divided spinal fractures depending on factors including the level of trauma, type of dislocation of the injured segment, and clinical symptoms and signs of spinal cord injury (relying on both their personal experience in treating victims of World War I along with an overview of the patients' X-ray data). One of the first lower cervical spine injury classification systems used in clinical practice, was suggested by Holdsworth.[2] This system was the first to consider both mechanism and the type of injury and recommend corresponding potential surgical management strategies. The majority of classification systems for subaxial cervical spinal injuries were developed and implemented in clinical practice over the past 40 years. The most popular morphological classification systems include Holdsworth, Allen-Fergusson (A-F), Harris, Argenson, and the AOSpine (AOS) scales.[3-5] While these categorizations have been used by spinal surgeons for many years, no single system is uniformly and widely accepted. According to the survey performed by Chhabra et al. 37.5%, 40%, and 7.5% experts were found to use the A-F classification, numeric scales, and other scales, respectively, for classifying cervical spinal injuries.[6] The study also reported that some experts never used any scales.

There are just a few existing studies on the potential applications of various lower cervical spine injury classifications. The implementation of a classification system in clinical practice generally comprises 3 stages,[7] of which, stage 1 involves development of the scale and its internal validation, stage 2 consists of external validation through clinical studies (including multicenter studies) after application of the system, and stage 3 involves validation of the system through prospective studies. The most investigated classification systems of those listed herein are the AOS and the A-F scales. We did not find any information concerning the reliability of the Argenson scale. Of the few articles presenting an assessment of reliability of the subaxial classification systems, only 1 includes the results of a stage 1 study of the AOSpine subaxial cervical spine injury classification.[5] Furthermore, only 1 of the 4 other published studies that include an external validation[8-11] of this system, is multicentric.[8] We found only 1 study on the Harris scale.[8] While opinions of specialists with different levels of experience have been compared in this study,[9] the review was monocentric. A summary of the current published studies on the topic is provided in Table 1.

The main purpose of our study was to measure and compare the inter- and intraobserver reliability for the Allen-Fergusson, Harris, Argenson, and AOSpine systems of classification of subaxial cervical spinal injuries, on implementation by neurosurgeons having different levels of experience and working in different clinics.

Materials and Methods

Patient Cohort

In this retrospective, survey analysis, we used data pertaining to 64 consecutive patients who underwent surgery between January 2013 and December 2017. All patients underwent surgery in the study initiator's institute. The institutional review board approval was obtained for the study. As this was a retrospective study of anonymized patient records, the requirement of consent was waived by the appropriate ethics review board.

We used anonymized computed tomography (CT) and magnetic resonance imaging (MRI) scans of 100% and 58% of the participants, respectively, for this study. We also collated data of each patient's neurological status. Personalized information was removed from the digital imaging and communications in medicine (DICOM) archive so that each participant could independently build multiplanar and 3-dimensional reformations. Every case included in the study was assigned a unique consecutive number. Every evaluating surgeon (rater) was provided with a folder, which included data of all 64 cases arranged in a random order, which ruled out the risk of duplication of answers.

Raters

A total of 37 surgeons from 7 different clinics were chosen for the study as evaluators or raters. Five clinics were level 1 trauma centers and 2 were university clinics. The raters were divided into 3 groups depending on their level of experience. The group of beginners (n = 20) included residents, nonspinal surgeons, and junior spinal surgeons having <5 years of surgical experience. The intermediate group (n = 10) comprised neurosurgeons having 5 to 10 years of experience in spinal surgery and who were able to perform a cervical spine procedure independently, having participated in multiple surgeries.

Table 1. Current Studies of Morphological Classifications Reliability.

Study	Classifications	No. of Patients	Patients	No. of Raters	Raters	No. of Participating Centers
Stone et al (2010)[10]	Allen-Fergusson	50	Consecutive patients	5	Not identified	Single center
Vaccaro et al (2007)[8]	Allen-Fergusson, Harris	11	Selected from database	20	5 neurosurgeons 15 orthopedic surgeons	Multicenter
Urrutia et al (2016)[11]	AOSpine, Allen-Fergusson	65	Selected from database	6	3 fellowship trained spine surgeons 3 orthopedic surgery residents	Single center
Vaccaro et al (2016)[5]	AOSpine	30	Selected from database	10	Spinal surgeons	Multicenter
Silva et al (2016)[9]	AOSpine	51	Consecutive patients	5	1 second-year resident 2 final-year residents 1 neurosurgeon 1 orthopedic surgeon	Single center
Present study	Allen-Fergusson, Harris, Argenson, AOSpine	64	Consecutive patients	37	20 beginners and nonspinal neurosurgeons 10 intermediate spine surgeons 7 experienced spine surgeons	Multicenter

The experienced group (n = 7) consisted of surgeons having >10 years of experience in spinal surgery. Only 9 of the 37 raters used the AOS classification routinely, while 9 others reported using it occasionally. Only 2 surgeons employed the A-F classification in practice. The remaining 19 raters had studied the various classification scales during their residency but had never utilized one in clinical practice.

Each rater was provided a personalized package, including (1) a USB flash drive with the blinded DICOM archive data along with information regarding the corresponding patient's neurological status; (2) a reference booklet, with material consisting of the original authors' illustrations and detailed descriptions of each of the 4 systems of classification being evaluated[12]; and (3) an application form to fill in their answers.

Assessment Process

In total, 2 assessment procedures were carried out. The initial assessments were received from 36 raters. Thereafter, the order and serial numbers of cases were randomly changed, and the second-stage assessment was completed by 24 raters separately in each clinic, 1.5 months after the first assessment.

Statistical Analysis

We used the Microsoft Excel 2011 for Mac along with the Visual Basic Applications (VBA) program AgreeStat 2015.6 for assessment of reliability of all A-F, Harris, Argenson, and AOS classification systems. Interrater reliability was estimated using the Fleiss' kappa (κ) parameter. The intrarater reliability for each rater was assessed separately, using Cohen's kappa.

The kappa statistics were interpreted using the Landis and Koch system.[13] If kappa was <0.2, the degree of agreement was estimated as slight. An indicator-value ranging between 0.2-0.4, 0.4-0.6, and 0.6-0.8 designated the degree of agreement as fair, moderate, and substantial, respectively. If kappa was >0.8, then the degree of agreement was interpreted as excellent.

The nonparametric statistical tests were performed using the program PC STATISTICA (Version 8) (StatSoft@ Inc, USA). The analysis of data in three variables was performed by applying the Kruskal-Wallis one-way criterion. The derived data were interpreted using P values (the expected deviation of a null hypothesis if both groups are not different). If P was >.05, the null hypothesis was not rejected, whereas, if $P < .05$, the null hypothesis was rejected, and the differences between the 2 groups were assumed to be significant.

Results

Overall, we processed 60 completed application forms, which included 36 and 24 forms received at the first and second stage, respectively. The raters performed a total of 3840 evaluations, for which we received 15 111 answers. The raters from the beginning group were unable to diagnose the injury in 21 evaluations and raters from all experience groups failed to classify the injury using A-F, Harris, Argenson, and AOS classification systems in 51, 56, 46, and 8 evaluations, respectively. We received duplicated answers for a single patient assessed using the A-F, Harris, Argenson, and AOS categorizations in 20, 15, and 12, and 5 evaluations, respectively. Raters were unable to visualize the files due to flash drive failure in 11 evaluations.

Allen-Fergusson Scheme

The general level of agreement among raters using this system was fair for the primary assessment. The κ value was 0.3 (95% confidence interval [CI], 0.02-0.34) (Table 2). This value was almost uniform across all raters with different levels of experience, but the highest value of $\kappa = 0.39$ was observed for the

Table 2. Interrater Reliability for Allen-Fergusson Scale Compared With Current Literature Data.[a]

				Present Study							
				First Assessment				Second Assessment			
	Vaccaro et al (2007)[8]	Stone et al (2010)[10]	Urrutia et al (2016)[11]	Beginners (n = 19)	Intermediate (n = 10)	Experienced (n = 7)	All Raters (n = 36)	Beginners (n = 12)	Intermediate (n = 8)	Experienced (n = 4)	All Raters (n = 24)
CF	—	0.52	0.37	0.26	0.35	0.34	0.31	0.34	0.40	0.48	0.39
CE	—	0.34	0.29	0.15	0.09	0.34	0.17	0.26	0.21	0.27	0.26
DF	—	0.54	0.57	0.34	0.35	0.50	0.36	0.38	0.39	0.46	0.39
DE	—	0.63	0.4	0.47	0.39	0.54	0.47	0.40	0.19	0.51	0.35
VC	—	0.61	0.58	0.30	0.27	0.28	0.29	0.24	0.16	0.22	0.23
LF	—	−0.16	0.3	0.27	0.19	0.09	0.22	0.12	0.05	−0.02	0.07
Overall	0.53	0.34	—	0.28	0.28	0.39	0.30	0.32	0.30	0.38	0.33

Abbreviations: CE, compression-extension; CF, compression-flexion; DE, distraction-extension; DF, distraction-flexion; LF, lateral flexion; VC, vertical compression.

[a] Color indication:

□ slight and fair agreement (κ = 0.01-0.39); 🟨 moderate agreement (κ = 0.40-0.59); 🟩 substantial agreement (κ = 0.6-0.79).

Table 3. Interrater Reliability for Harris Classification Compared With Current Literature Data.[a]

		Present Study							
		First Assessment				Second Assessment			
	Vaccaro et al (2007)[8]	Beginners (n = 19)	Intermediate (n = 10)	Experienced (n = 7)	All Raters (n = 36)	Beginners (n = 12)	Intermediate (n = 8)	Experienced (n = 4)	All Raters (n = 24)
Flexion	—	0.22	0.39	0.43	0.27	0.27	0.26	0.41	0.29
Compression	—	0.40	0.48	0.42	0.41	0.38	0.33	0.43	0.38
Extension	—	0.36	0.37	0.58	0.36	0.29	0.22	0.39	0.29
Flexion-rotation	—	0.24	0.48	0.61	0.36	0.33	0.31	0.47	0.37
Extension-rotation	—	0.28	0.41	0.55	0.35	0.28	0.15	0.21	0.24
Lateral flexion	—	0.31	0.11	−0.01	0.15	0.04	−0.01	−0.03	0.07
Overall	0.41	0.40	0.51	0.58	0.44	0.36	0.34	0.49	0.38

[a] Color indication:

□ slight and fair agreement (κ = 0.01-0.39); 🟨 moderate agreement (κ = 0.40-0.59); 🟩 substantial agreement (κ = 0.6-0.79).

experienced group (95% CI, 0.32-0.46). The general kappa value was slightly higher for the second assessment and reached 0.33 (95% CI, 0.29-0.36). However, it never exceeded 0.39, even among experienced experts, clearly indicating a fair agreement regarding the A-F classification system across this group. Analysis of injury subtypes demonstrated the best agreement during the second assessment among experts in the compression-flexion (κ = 0.48; 95% CI, 0.35-0.62), distraction-flexion (κ = 0.46; 95% CI, 0.34-0.59), and the distraction-extension (κ = 0.51; 95% CI, 0.22-0.81) groups.

Harris Classification

The interrater reliability level for the primary assessment was moderate (0.44) for all examined groups (95% CI, 0.33-0.56) (Table 3). However, this value dropped to a fair level following the second assessment by the beginner (κ = 0.36; 95% CI, 0.26-0.45) and the intermediate (κ = 0.34; 95% CI, 0.23-0.45) groups. The agreement parameter remained at a moderate level among the experts through both first and second stages of the study, but the κ value was found to decrease even for this group, from 0.58 (95% CI, 0.47-0.70) for the first stage to 0.49

Table 4. Interrater Reliability for Argenson Classification.[a]

	Present Study							
	First Assessment				Second Assessment			
Injury Type[b]	Beginners (n = 19)	Intermediates (n = 10)	Experienced (n = 7)	All Raters (n = 36)	Beginners (n = 12)	Intermediate (n = 8)	Experienced (n = 4)	All Raters (n = 24)
A	0.39	0.42	0.55	0.42	0.51	0.48	0.40	0.49
BF, BE	0.30	0.21	0.43	0.29	0.38	0.32	0.32	0.36
C	0.38	0.30	0.55	0.39	0.52	0.39	0.50	0.47
Overall	0.43	0.46	0.61	0.47	0.60	0.52	0.53	0.55

[a] Color indication:
□ slight and fair agreement (κ = 0.01-0.39); 🟨 moderate agreement (κ = 0.40-0.59); 🟩 substantial agreement (κ = 0.6-0.79).

[b] A, compressive injury; BE, distractive extension injury; BF, distractive flexion injury; C, rotational injury.

Table 5. Interrater Reliability for AOSpine Classification Compared With Current Literature Data.[a]

				Present Study							
				First Assessment				Second Assessment			
	Vaccaro et al (2016)[5]	Urrutia et al (2016)[11]	Silva et al (2016)[9]	Beginners (n = 19)	Intermediate (n = 10)	Experienced (n = 7)	All Raters (n = 36)	Beginners (n = 12)	Intermediate (n = 8)	Experienced (n = 4)	All Raters (n = 24)
A	0.66	0.64	0.52-0.63	0.50	0.44	0.57	0.50	0.75	0.61	0.73	0.70
B	0.54	0.51		0.22	0.14	0.37	0.22	0.54	0.32	0.55	0.46
C	0.73	0.65		0.51	0.46	0.61	0.51	0.60	0.48	0.75	0.55
F	0.66	0.61	0.53-0.60	0.40	0.38	0.54	0.42	0.37	0.34	0.56	0.39
Overall	0.65	0.61	—	0.48	0.44	0.62	0.49	0.60	0.46	0.70	0.56

[a] Color indication:
□ slight and fair agreement (κ = 0.01-0.39); 🟨 moderate agreement (κ = 0.40-0.59); 🟩 substantial agreement (κ = 0.6-0.79).

(95% CI, 0.32-0.67) for the second-stage assessment. A group analysis demonstrated that the best results for an application of this categorization were derived for the primary assessment. Moderate and substantial levels of agreement were observed among experts for all types of injury, except for lateral flexion (κ = −0.01; 95% CI, −0.02 to 0.00).

Argenson Scheme

The primary assessment of interrater reliability for this classification demonstrated moderate agreement level across all raters (κ = 0.47; 95% CI, 0.40-0.54) (Table 4). We observed a substantial level of agreement with regard to all types of injuries among experienced surgeons (κ = 0.61; 95% CI, 0.52-0.70). The second assessment revealed an improvement in agreement across all raters to a substantial level in some cases (κ = 0.55; 95% CI, 0.48-0.63), which happened due to an improvement in agreement values for the beginner (κ = 0.60; 95% CI, 0.52-0.68) and intermediate groups (κ = 0.52; 95% CI, 0.43-0.61) associated with a decrease in the value of agreement among experienced surgeons (κ = 0.53; 95% CI, 0.40-0.65). An analysis of the interrater reliability for agreement with regard to injury subtypes demonstrated the best results for compression (type A) and rotational (type C) injuries, for which the Fleiss' kappa value was estimated to be 0.49 (95% CI, 0.4-0.58) and 0.47 (95%, 0.37-0.56), respectively. The level of agreement for distraction injuries in most cases was fair and reached 0.36 (95% CI, 0.30-0.41).

AOSpine Classification

The general kappa value for all injury types and across all raters for the primary assessment using this classification system, was 0.49 (95% CI, 0.41-0.57) (Table 5). While the agreement level

Table 6. Intrarater Reliability for Morphological Classifications Comparied With Current Literature Data.[a]

	Vaccaro et al (2016)[5]	Vaccaro et al (2007)[8]	Silva et al (2016)[9]	Stone et al (2010)[10]	Urrutia et al (2016)[11]	Present Study			
						Beginners, n = 11 (Range)	Intermediate, n = 7 (Range)	Experienced, n = 4 (Range)	All Raters, n = 22 (Range)
Allen-Fergusson	—	0.63	—	0.91	0.66	0.27 (0.04-0.49)	0.29 (0.02-0.47)	0.44 (0.34-0.51)	0.30 (0.02-0.51)
Harris	—	0.53	—	—	—	0.37 (0.09-0.67)	0.44 (0.13-0.70)	0.49 (0.43-0.54)	0.41 (0.09-0.70)
Argenson	—	—	—	—	—	0.41 (0.03-0.72)	0.46 (0.17-0.71)	0.46 (0.32-0.63)	0.44 (0.03-0.72)
AOSpine	0.74	—	0.66-0,95	—	0.68	0.45 (0.12-0.72)	0.44 (0.20-0.71)	0.56 (0.51-0.61)	0.46 (0.12-0.72)

[a] Color indication:
⬜ slight and fair agreement (κ = 0.01-0.39); 🟧 moderate agreement (κ = 0.40-0.59); 🟩 substantial agreement (κ = 0.6-0.79).

among experienced surgeons was substantial (κ = 0.62; 95% CI, 0.51-0.72), it decreased to moderate for the beginner (κ = 0.48; 95% CI, 0.39-0.56) and the intermediate (κ = 0.44; 95% CI, 0.36-0.52) groups. The general agreement kappa value for the second assessment reached 0.56 (95% CI, 0.48-0.64), with a substantial level of agreement among beginners (κ = 0.60; 95% CI, 0.54-0.66) and experienced surgeons (κ = 0.70; 95% CI, 0.60-0.81). The κ value reached moderate level for the intermediate group (κ = 0.46; 95% CI, 0.35-0.56). An analysis of the classification system according to injury subtypes demonstrated the best agreement among all raters for compression injuries during the second assessment (κ = 0.70; 95% CI, 0.63-0.76). The agreement level ranged from moderate to substantial for different cases (κ = 0.55; 95% CI, 0.45-0.65) of translational injuries (C type), while it reached only up to a moderate level for distraction injuries (κ = 0.46; 95% CI, 0.39-0.53).

Reproducibility of All Examined Scales

The reproducibility of the A-F scale was found to be the lowest (κ = 0.30; range: 0.04-0.51) with even experienced surgeons reaching only a moderate level of agreement (κ = 0.44; range: 0.34-0.51), on applying the system (Table 6). An assessment of the Harris classification revealed moderate agreement among all raters during both stages of the study (κ = 0.41; range: 0.09-0.70) and showed no significant variations between surgeons with different levels of experience. The intraobserver agreement level for the Argenson classification was also moderate (κ = 0.44; range: 0.03-0.72), but it reached substantial level only for some surgeons from all three groups. The AOS showed the highest intraobserver agreement among morphological classifications with regard to reproducibility, with all examining surgeons reaching a moderate level of agreement for all types of injuries (κ = 0.46; range: 0.12-0.72) and showing a κ value of 0.56 for the group of experienced surgeons (range: 0.51-0.61).

The reproducibility of both Argenson and AOS scales was found to be higher among surgeons using the AOS classification routinely or occasionally in their practice (Kruskal-Wallis [K-W] test, *P* = .01 and .03, respectively) (Figure 1). A regular application of the AOS system also resulted in an increase in reproducibility of the Harris and A-F scales (K-W test, *P* = .048 and .02, respectively) (Figure 1).

The group of experienced surgeons demonstrated the highest level of reproducibility. However, this difference was statistically insignificant (Figure 2). The *P* values for reproducibility of A-F, Harris, Argenson, and AOS scales, as per the K-W test were .08, .46, .89, and .49, respectively.

Discussion

The controversy surrounding the development of an ideal spine injury classification system has remained unsolved for a long time. A widely applicable and accepted system of classification was required to meet several conditions[14]: (1) to contain clear terms for the stratification, thus ruling out any ambiguous or free interpretation; (2) to have categories that are comprehensive and mutually exclusive; (3) to contain clear graphic illustrations; (4) to be simple and suitable for routine clinical application; (5) to contain a limited number of categories; (6) to be structured in such a manner as to indicate a gradual increase in injury severity; (7) to indicate each defined group and subgroup with a unique alpha-arithmetic name; and (8) to represent clear and distinctive signs of injury detected on diagnostic imaging studies.

Unfortunately, there is no ideal classification that fulfills all these criteria. Some descriptive classification systems that include all types of injury patterns are too intricate and cumbersome for recall, even with a gentle learning curve. Simplifying the classification can result in insufficient or incomplete representation of the fracture structure. Finding a balance between scale simplicity and reproducibility on one hand and complicated content of the classification system on the other, is a challenging and controversial issue.

We believe that all injury patterns are covered in the most comprehensive manner by the A-F classification system. Each type of injury corresponds to a specific combined traumatic mechanism. All included subtypes are graded to indicate a gradual increase in the severity of injury and the degree of vertebral dislocation. The scale illustrates different types of injuries in the most obvious manner. Therefore, a classification

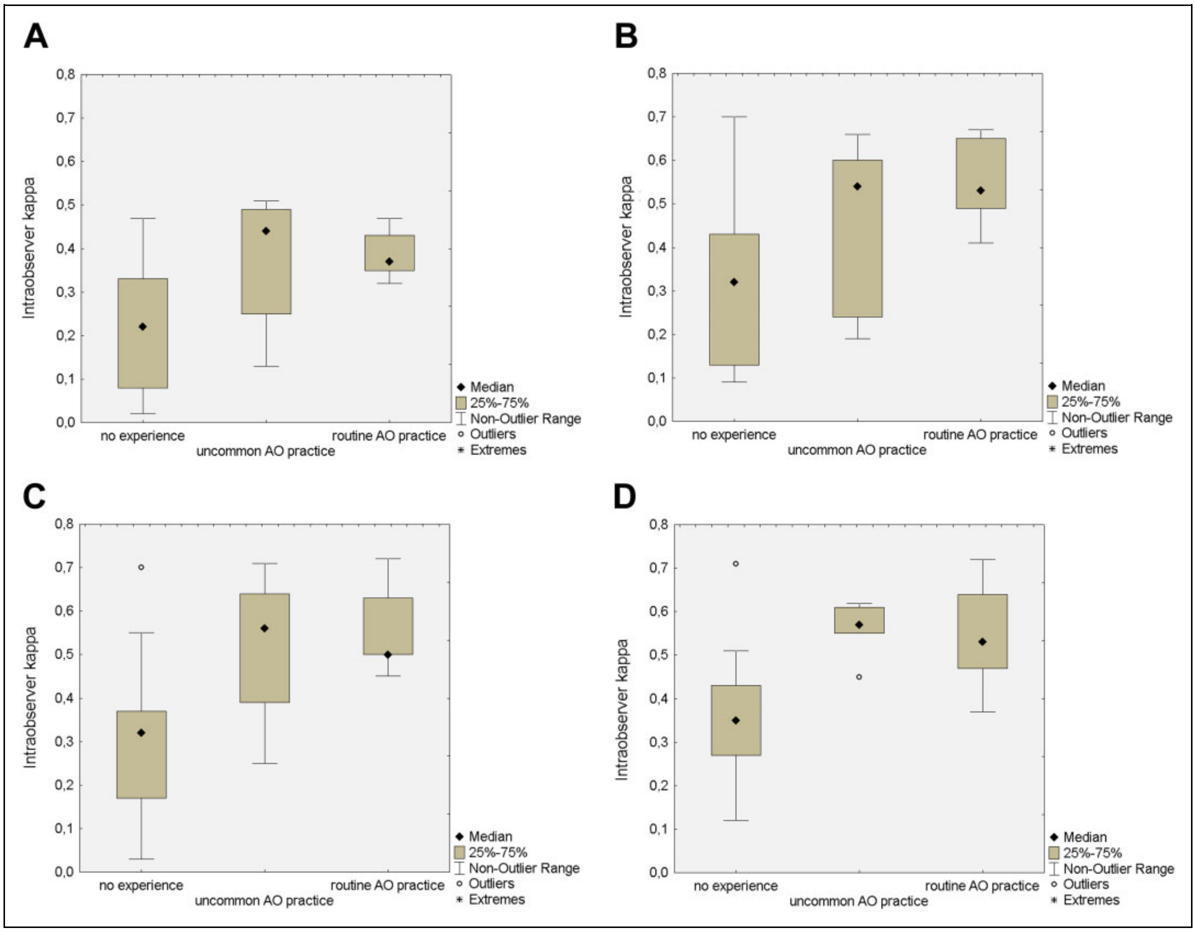

Figure 1. Reproducibility data depending on prior experience (of applying AO classification routinely) on implementation of (a) Allen-Fergusson, (b) Harris, (c) Argenson, and (d) AOSpine scales.

of the fracture type and its stage according to the A-F system enables a clear understanding of factors including the direction and degree of dislocation of the fractured fragment, the consistency of posterior spinal structures, and other important aspects of the fracture. However, as seen in this study, an application of the A-F system with its large number of subtypes (26) resulted in a high variability in the answers provided by the raters. While best results were elicited from experienced surgeons, only a moderate level of agreement was reached even by this group for most types of injuries. Previous studies[8,10,11] reporting an assessment of A-F scale only included an application of the scale by highly experienced surgeons. Results of our assessment of interobserver agreement are comparable to those reported in the existing published data. The interobserver kappa value was much lower for the group of beginners due to the high complexity of this scale (as compared with others) and the need for combining primary mechanisms of injury (flexion and extension) with additional force vectors (compression and distraction). As opined by the most examined raters, major disadvantages of the A-F scale include (1) the complexity of differential diagnosis of compression-flexion and compression injuries; (2) inability to classify traumatic disk herniations; (3) a challenging classification of rotational injuries combined with articular joint fractures; (4) a challenging classification of distraction injuries combined with disk ruptures and posterior ligament injuries (ie, B2 type in AO classification and 1a type in Harris classification); (5) complicated gradation of some injury types, such as a floating lateral mass; and (6) the absence of differentiation between stable and unstable fractures of the facet joint. In our study, the reproducibility of this scale was found to be much lower than that reported in current literature, even among experienced surgeons.[8,10,11] This discrepancy may have occurred due to the commonly tested experienced surgeons having previously faced the scale only in theory, without experiencing any practical application.

The Harris classification system appears to be much simpler compared with the A-F scale. This system considers only the major force vectors (flexion, compression, and extension) along with 2 distinct types of rotational injuries, thus accommodating 12 subtypes of lower cervical spine injuries. This simplification, however, has an impact on the interobserver agreement. In a previous study by Vaccaro et al,[8] experienced surgeons reached a moderate level of agreement for this classification. Our study demonstrated a higher agreement among the group of experienced surgeons, reaching the substantial

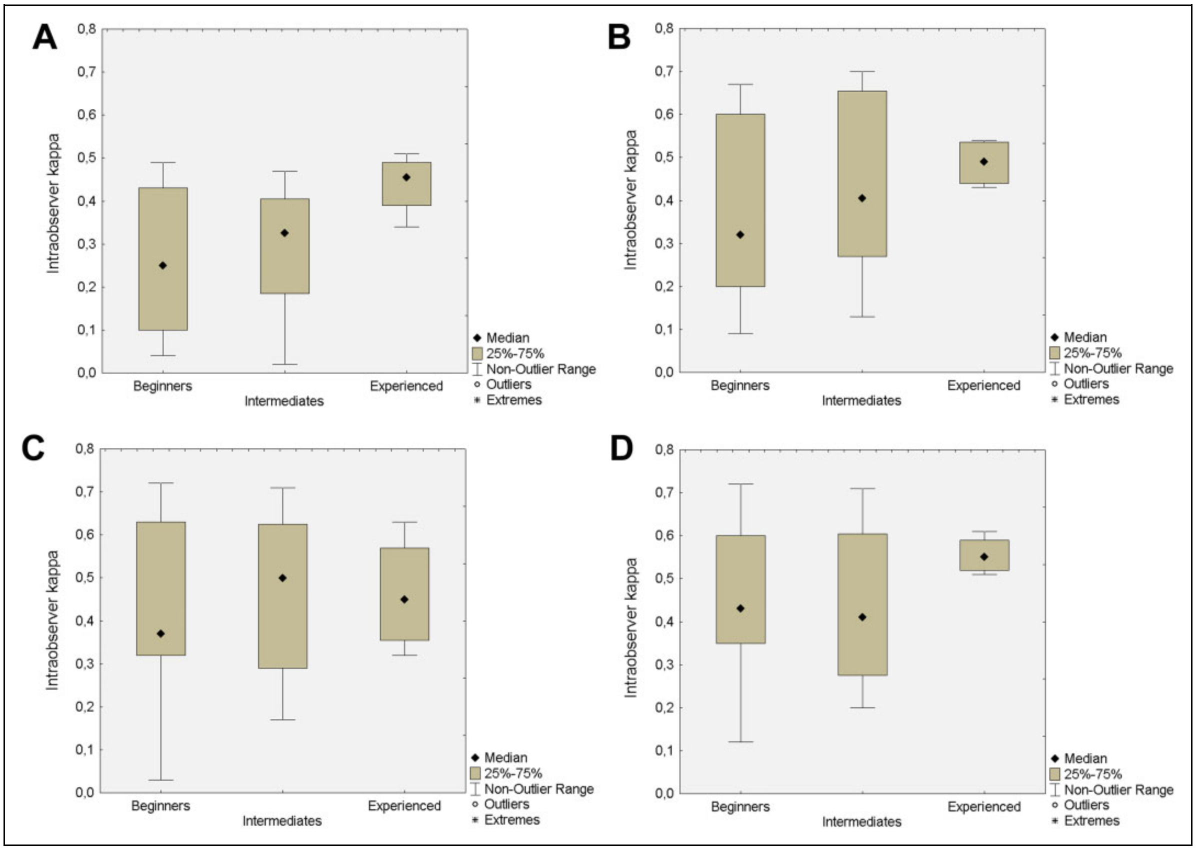

Figure 2. Reproducibility data depending on surgeons' levels of experience for (a) Allen-Fergusson (b) Harris, (c) Argenson, and (d) AOSpine scales.

level. Interobserver kappa for the group of beginners was the same as that for the A-F scale. The raters revealed some disadvantages of this system: (1) absence of gradation for compression fractures; (2) a challenging differentiation for certain injuries, for example, a floating lateral mass; and (3) absence of gradation for traumatic disk herniations. The reproducibility of this scale among experienced surgeons corresponded to that reported in current literature,[8] with achievement of a moderate level across all surgeons.

The Argenson classification describes only 3 types of injuries, resulting in significant increase in interobserver agreement. This effect was observed across all groups of surgeons. During the second stage of the study, substantial agreement was observed among beginners, indicating the potential for a faster learning and adoption of the scale, if routinely applied. The general agreement was close to substantial level for almost all raters, indicating that this scale was one of the most convenient for practical application. Some disadvantages of the system enumerated by the raters included (1) an absence of gradation of compression fractures, (2) an absence of gradation of traumatic disk herniations, (3) an unclear description of flexion-distraction and extension-distraction injuries, and (4) a challenging gradation of facet-joint fractures without dislocation. The general reproducibility of the scale was found to be at a moderate level, which was slightly higher than that for the Harris scale. In the assessment performed across all groups, its value reached substantial level for some raters.

The AOSpine scale comprises groups, of which 2 are based on the primary mechanism of trauma that is, compression and distraction injuries. Another group includes vertebral dislocation in any direction. We believe that this simplified stratification resulted in the highest values of interobserver kappa achieved across all rater groups. During the second assessment, the expert group demonstrated significantly higher values than those using the A-F, Harris, and Argenson classification systems, a finding in accordance with data reported in current literature.[5,9,11] The main disadvantages of the AOS scale as reported by the raters included (1) a challenging identification of F1 and F2 fractures, (2) common confusion between A2 and A4 types in vertically split vertebral injuries, and (3) a frequent absence of facet-joint trauma classification in addition to the major injury. The intraobserver kappa was found to be higher for the AOS scale than for the Argenson system in our study and reached a moderate level.

Therefore, the highest level of agreement across all raters with different degrees of experience was obtained for the AOS and the Argenson scales. The highest value of Fleiss kappa was determined for the AOS scale. This may have also occurred because most raters had prior experience in applying this scale in practice. However, considering certain features of the Fleiss kappa calculation, no clear statistical correlation can be determined.

The A-F and Harris scales meet only half of the criteria defining an ideal classification scale.[14] A major disadvantage includes an absence of clear graphic illustrations supplemented with a detailed description. We found a clear graphic implementation of these scales described only in one book,[15] but the associated description of injury subtypes was incomplete and superficial. This could result in some degree of inconsistency in implementation and consequently in the answers provided by the raters. Therefore, for this study, we developed a special reference book including illustrations used by the original authors of both these systems of classification, along with a comprehensive description of each stage of injury, based on the original articles.[12,16,17] In contrast to these scales, the Argenson and AOS classifications are supplemented with good illustrations and descriptions. Nevertheless, only the AOS scale met almost all criteria of an ideal classification. The Argenson scale missed certain injury subtypes, and according to some raters, the original reference for the system contained an incomplete description of distraction injuries. The only major disadvantage of AOS found by raters, was the potentially incorrect interpretation of the description and illustrations of stable and unstable fractures involving articular joints.

While estimating the reproducibility of the scales, we derived widely variable kappa values. At the beginning of the study more than half of the raters had no experience of using scales in clinical practice. We did not conduct any orientation seminar prior to the first stage of the study. Therefore, raters had to learn these scales individually, using the provided reference materials. By the second stage of the study, the raters acquired practical experience of applying the system. Consequently, some raters changed their opinion regarding certain cases, resulting in a change in their reported answers in the second stage of the study. This, in turn, resulted in an increase in interrater reliability for A-F, Argenson, and AOS scales associated with a decrease in intrarater reliability. The remaining raters, who used the AOS scale routinely or occasionally, demonstrated a clear understanding of injury mechanisms while applying the system. The intrarater reliability was significantly higher and more stable for these raters not only for the AOS system but also for other morphological scales.

Furthermore, in almost all previous studies,[5,8,11] patients or cases with the most characteristic and distinctive injuries were chosen for examination. The controversial and most complicated cases could be excluded, resulting in higher values of the level of agreement and more stable values for reproducibility. In our study as well as in 2 other published studies,[9,10] data of all patients within a certain time interval were included, which made the case-sampling divergent, thus increasing resemblance to actual clinical scenarios.

We believe that the major advantage of this study is not only its multicentric nature (involving surgeons from different clinics and different surgical schools) but also the evaluation of implementation of the classification systems by specialists having varying levels of experience. The residents and junior surgeons are usually the first to see the patients in the clinic. Interpreting and presenting an accurate diagnosis is one of the most important aspects of managing a patient with cervical spinal trauma at admission and going forward. The highest values of interobserver agreement among specialists with different levels of experience were found on implementation of Argenson and AOS scales. We suggest that a thorough knowledge of these scales acquired during residency followed by a routine practical application, may increase their reproducibility to an excellent level.

Conclusion

Our results showed that the highest values for both interobserver agreement and reproducibility among surgeons with varying levels of experience, were found with an implementation of Argenson and AOSpine classification systems. The AOSpine scale additionally incorporated more detailed description of compression injuries and facet-joint fractures. The levels of agreement for A-F and Harris scales were fair and moderate, respectively, indicating that their application in clinical practice, especially by junior specialists, would be challenging. Future studies on the reliability of current classification systems should not only involve spinal surgeons, but also various specialists from related branches.

Acknowledgments

We would like to thank Editage (www.editage.com) for English language editing.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Ivan Lvov, MD, PhD https://orcid.org/0000-0003-1718-0792

References

1. Böhler L. *The Treatment of Fractures. A Translation of "Technik der Knochenbruchbehandlung im Frieden und im Kriege" by M. E. Steinberg*. Vienna, Austria: Wilhelm Maudrich; 1929.
2. Holdsworth F. Fractures, dislocations, and fracture-dislocations of the spine. *J Bone Joint Surg Am*. 1970;52:1534-1551.
3. Walters BC, Hadley MN, Hurlbert RJ, et al. Guidelines for the management of acute cervical spine and spinal cord injuries: 2013 update. *Neurosurgery*. 2013;60(CN suppl 1):82-91.
4. Argenson C, de Peretti F, Ghabris A, Eude P, Lovet J, Hovorka I. Classification of lower cervical spine injuries. *Eur J Orthop Surg Traumatol*. 1997;7;215-229.
5. Vaccaro AR, Koerner JD, Radcliff KE, et al. AOSpine subaxial cervical spine injury classification system. *Eur Spine J*. 2016;25: 2173-2184.

6. Chhabra HS, Kaul R, Kanagaraju V. Do we have an ideal classification system for thoracolumbar and subaxial cervical spine injuries: what is the expert's perspective? *Spinal Cord*. 2015;53: 42-48.
7. Audigé L, Bhandari M, Hanson B, Kellam J. A concept for the validation of fracture classifications. *J Orthop Trauma*. 2005;19: 401-406.
8. Vaccaro AR, Hulbert RJ, Patel AA, et al; Spine Trauma Study Group. 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 (Phila Pa 1976)*. 2007;32:2365-2374.
9. Silva OT, Sabba MF, Lira HI, et al. Evaluation of the reliability and validity of the newer AOSpine subaxial cervical injury classification (C-3 to C-7). *J Neurosurg Spine*. 2016;25:303-308.
10. Stone AT, Bransford RJ, Lee MJ, et al. Reliability of classification systems for subaxial cervical injuries. *Evid Based Spine Care J*. 2010;1:19-26.
11. Urrutia J, Zamora T, Campos M, et al. A comparative agreement evaluation of two subaxial cervical spine injury classification systems: the AOSpine and the Allen and Ferguson schemes. *Eur Spine J*. 2016;25:2185-2192.
12. Grin AA, Lvov IS, Arakelyan SL, et al. Currently available classification systems for lower cervical spine injuries. Part 1. Overview of the most popular scales and classifications [in Russian]. *Russian J Neurosurg*. 2019;21:90-102. doi:10.17650/1683-3295-2019-21-1-90-102
13. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159-174.
14. van Middendorp JJ, Audigé L, Hanson B, Chapman JR, Hosman AJ. What should an ideal spinal injury classification system consist of? A methodological review and conceptual proposal for future classifications. *Eur Spine J*. 2010;19: 1238-1249.
15. Chapman JR, Dettori JR, Norvel DC, eds. *Spine Classifications and Severity Measures*. 2nd ed. Stuttgart, Germany: Thieme; 2009.
16. Allen BL Jr, Ferguson RL, Lehmann TR, O'Brien RP. A mechanistic classification of closed, indirect fractures and dislocations of the lower cervical spine. *Spine (Phila Pa 1976)*. 1982;7:1-27.
17. Harris JH Jr, Edeiken-Monroe B, Kopaniky DR. A practical classification of acute cervical spine injuries. *Orthop Clin North Am*. 1986;17:15-30.