AOSPINE

Global Spine Journal
2019, Vol. 9(2) 231-242
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2192568218806847
journals.sagepub.com/home/gsj
SAGE

Reliability and Validity of the AOSpine Thoracolumbar Injury Classification System: A Systematic Review

Aidin Abedi, MD[1], Lidwine B. Mokkink, PhD[2,3], Shayan Abdollah Zadegan, MD[4], Permsak Paholpak, MD[5], Koji Tamai, MD[6], Jeffrey C. Wang, MD[1], and Zorica Buser, PhD[1]

Abstract

Study Design: Systematic review.

Objectives: The AOSpine thoracolumbar injury classification system (ATLICS) is a relatively simple yet comprehensive classification of spine injuries introduced in 2013. This systematic review summarizes the evidence on measurement properties of this new classification, particularly the reliability and validity of the main morphologic injury types with and without inclusion of the subtypes.

Methods: A literature search was performed using PubMed and Embase in September 2016. A revised version of the COSMIN checklist was used for evaluation of the quality of studies. Two independent reviewers performed all steps of the review.

Results: Nine articles were included in the final review, all of which evaluated the reliability of the ATLICS and had a fair methodological quality. The reliability of the modifiers was unknown. Overall, the quality of evidence for reliability of the morphologic and neurologic classification sections was low. However, there was moderate evidence for poor interobserver reliability of the morphologic classification when all subtypes were included, and moderate evidence for good intraobserver reliability with exclusion of subtypes. The reliability of the morphologic classification was independent of the observer's experience and cultural background.

Conclusions: ATLICS represents the most current system for evaluation of thoracolumbar injuries. Based on this review, further studies with robust methodological quality are needed to evaluate the measurement properties of ATLICS. Shortcomings of the reliability studies are discussed.

Keywords

reliability, reproducibility of results, validity, spine, thoracolumbar, trauma, spinal injuries, spinal fractures, injury severity score, classification

Introduction

Classification of spinal injuries is an ongoing challenge. Since the first classification of thoracolumbar injuries proposed by Watson-Jones[1] in 1938, substantial efforts have been made to design an ideal scheme which is valid and reliable. Although various classification systems have been developed, none are universally accepted.[2]

The AOSpine thoracolumbar injury classification system (ATLICS) was introduced by Vaccaro et al[3] in 2013, as a relatively simple yet comprehensive classification system. ATLICS employs the features of 2 previous classification systems: the Magerl system[4] and the Thoracolumbar Injury Classification and Severity Score (TLICS).[5] Magerl et al[4] established their classification in 1994, based on 3 main mechanisms of fracture: vertebral body compression (type A), anterior and posterior element injuries with distraction (type B), and anterior and posterior element injuries with

[1] University of Southern California, Los Angeles, CA, USA
[2] VU University Medical Center, Amsterdam, the Netherlands
[3] Amsterdam Public Health Research Institute, Amsterdam, the Netherlands
[4] Tehran University of Medical Sciences, Tehran, Iran
[5] Khon Kaen University, Khon Kaen, Thailand
[6] Osaka City University, Osaka, Japan

Corresponding Author:
Zorica Buser, Department of Orthopaedic Surgery, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA.
Email: zbuser@usc.edu

rotation (type C). The TLICS was established in 2005 by Vaccaro et al.[5] They used the following 3 primary morphologies: compression, translation/rotation, and distraction. In addition, they included the integrity of the posterior ligamentous complex and the neurologic status of the patient and also weighted the injury severity with a point system.

Despite the relatively short period after its development, ATLICS has gained attention as a substitute for its predecessor, the TLICS. There have been ongoing efforts to evaluate the properties of ATLICS, to develop an injury severity score and a surgical algorithm based on this classification.[6-9] Selection of a classification system is often a conundrum that needs a thorough evaluation and critical appraisal of the literature on reliability and validity of existing measures. Therefore, this systematic review summarizes the evidence on measurement properties of ATLICS, mainly focusing on the reliability and validity of the morphologic classification with and without inclusion of subtypes. Our secondary objective was to assess the effect of observers' experience and professional background on measurement properties of ATLICS.

Materials and Methods

Protocol

This systematic review was designed in line with the recommendations of the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) initiative[10,11] and reported in compliance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Protocols (PRISMA-P) statement.[12] All steps of the review process were performed independently by 2 reviewers and controversies were resolved through consensus.

Literature Search

A literature search was performed using PubMed and Embase electronic databases in September 2016, with the keywords "AOSpine," "injury," "fracture," and "classification" (Appendix A). The search was supplemented using related Medical Subject Headings (MeSH) and EMTREE terms. No restriction was applied regarding the publication type, language, date, or other search limits. Additionally, the references of the included studies were screened for potentially relevant articles that were not identified in the electronic search.

Eligibility Criteria

Inclusion criteria:

1- The aim of the study was to evaluate the reliability and/or validity of the ATLICS.

OR

2- The focus of the study was on development or modification of the ATLICS, the surgical algorithm or the injury severity score derived from this classification.

Exclusion criteria:

1- Studies with no data on reliability and/or validity of the ATLICS.

2- Studies evaluating the preliminary version(s) of the ATLICS.

3- Studies with duplicate data.

Study Selection

Two authors (AA and ZB) independently screened the titles and abstracts of all identified records and appraised them using the eligibility criteria above. Other relevant records such as reviews and technical notes were initially considered for inclusion to be used for reference tracking. Same criteria were applied for screening the references of included studies, and for selection of the final set of full-texts for data extraction and best evidence synthesis. All discrepancies during the selection process were resolved through consensus between the 2 reviewers.

Measurement Properties

Measurement properties of interest were defined according to the COSMIN taxonomy.[13] Only the following measurement properties that are relevant to ATLICS were included: face and content validity, construct validity, reliability (both inter- and intraobserver reliability) and measurement error. Internal consistency and structural validity were irrelevant, as these measurement properties are only relevant for multi-item instruments based on a reflective model. Criterion validity was irrelevant as no gold standard exists to classify spine injuries.

Evaluation of the Quality of Studies

Since there are no widely accepted standards for assessing the methodological quality of studies on the measurement properties of classification systems, a content comparison of 3 existing instruments was performed: the COSMIN checklist,[14] the Quality Appraisal of Reliability Studies (QAREL) checklist,[15] and the quality criteria proposed by Audigé et al.[16] One of the main disadvantages of QAREL is that the items cover the generalizability of results rather than the methodological quality of the studies. Furthermore, lack of details in QAREL items concerning the statistical methods hinders a comprehensive assessment of the quality of studies. Besides, QAREL provides only 2 options for each item, which negatively influences the flexibility and precision of the quality assessment process. Therefore, COSMIN checklist[14] was selected as the most appropriate tool, as it includes the most comprehensive, systematic and transparent methodology for assessment of the methodological quality of studies on reliability, validity, and responsiveness.[14] However, since COSMIN was developed for patient-reported outcome measures, it was modified to meet the purpose of this study (Appendix B). The box Reliability and the box Measurement error of the COSMIN were modified, by removing the questions/standards addressing "missing items," since they are irrelevant considering the output of the ATLICS. Furthermore, the item that concerns the sample size was removed, as there is no consensus on the adequate sample size for reliability studies with multiple raters. Moreover, sample sizes are now included in the data syntheses phase. Additionally, as suggested in other published standards, lack of blinding to patients' clinical findings was included as a minor methodological flaw in the modified checklist when the morphologic classification section of ATLICS was the only focus of a study.[15,16] Following these modifications, 3 authors (AA, ZB, and LBM) pilot-tested the checklist in three papers to increase the agreement among reviewers. Similarly as done in systematic reviews of patient-reported outcome measures that use the COSMIN checklist, the overall quality of each study was determined based on the COSMIN item with the lowest score (ie, worst-score-counts method).[17]

Data Extraction

The following information was extracted from all included studies: the sampling method, patient characteristics (ie, the number of injury levels and distribution of injury based on ATLICS), observer characteristics (ie, their specialties and level of experience), imaging modalities and findings. For reliability studies, kappa values and percentage agreement were extracted as indicators of reliability and measurement error, respectively. For content validity studies, the findings on relevance of ATLICS (ie, to aspects such as the construct of thoracolumbar injury, target population and its discriminative application), its comprehensiveness and comprehensibility were considered for extraction. For construct validity studies, the hypotheses (when formulated in advance) and findings regarding the correlation of ATLICS with other measures and its ability to discriminate between important patient subgroups were considered for extraction.

Best Evidence Synthesis

The results of each study were evaluated using the criteria proposed by Terwee et al[18] (Appendix C). To generate an overall rating, qualitative summaries were produced and rated as follows: "positive (+)" overall findings when at least 75% of the summarized results met the criteria for good measurement properties (Appendix C); "negative (−)" when less than 25% of the summarized results met the criteria for good measurement properties; and "conflicting" when between 25% and 75% of the summarized results met the criteria for good measurement properties. Consequently, the quality of evidence (QoE) was determined based on the quality of studies and results, according to the criteria proposed by Prinsen et al[19] (Appendix D).

Results

Overview of the Studies

The literature search identified 63 unique records, of which 17 were selected for the full-text review and 9 articles[3,20-27] were included in the final review (Figure 1). There was a substantial variation in the characteristics of included studies in terms of the number of observers and cases and imaging modalities (Table 1). The sampling method was random in 2 studies, consecutive in 2, not reported in 1, and purposive in the remaining 4 studies. The observers in all studies were surgeons with different levels of training and experience. All studies used the same static images for repeated assessments. All included studies evaluated the reliability or measurement error. However, there was no study on content, face, or construct validity of the ATLICS. While all studies focused on the morphologic classification section of the ATLICS, the reliability of the neurological classification was evaluated in only 1 study, and the modifiers were not addressed in any of the studies. The overall quality of all included studies was fair. Lack of blinding to the clinical findings and use of unweighted kappa statistical method were the most common methodological flaws of the studies. Details of the results of included studies are presented in Table 2. Summary of the evidence on intra- and interobserver reliability of the ATLICS is presented in Tables 3 and 4, respectively.

Neurologic Classification

Reliability of the neurologic classification was assessed only in 1 study, using the medical data provided to the observers. This study showed positive results for inter- and intraobserver reliability, with kappa values of 0.85 and 0.91, respectively (QoE: low) (Table 2).

Morphologic Classification

Reliability of the morphologic classification was assessed at 4 different levels: the overall morphologic classification (ie, 3 main types of injury and their subtypes), overall morphologic classification with exclusion of subtypes (ie, 3 main types of injury) and separately for each main type and subtype.

Intraobserver reliability of the overall morphologic classification was evaluated in 5 studies, of which 2 studies showed positive results (QoE: low) (Table 3). With exclusion of subtypes, the intraobserver reliability improved in all studies, demonstrating positive results in 4 out of 5 studies (QoE: moderate). The intraobserver reliability of the type A and type B injuries was reported in 4 studies, of which 3 studies showed positive results for type A and 2 studies for type B (QoE: low). The intraobserver reliability of type C injury was positive in 2 studies (QoE: moderate).

Interobserver reliability of the overall morphologic classification was evaluated in 4 studies, all of which showed negative results (QoE: moderate) (Table 4). Although in all studies the kappa values increased after exclusion of subtypes, the minimum requirement for good reliability was fulfilled only in 2 studies (QoE: low). Regarding the main injury types, the types A and C had positive interobserver reliability in 4 out of 6 studies. However, the proportion of studies with positive findings did not reach the 75% cutoff for having an overall acceptable interobserver reliability (QoE: low). For type B injuries, only 1 study showed good reliability, while other 5 studies showed negative results with kappa values as low as 0.22 (QoE:

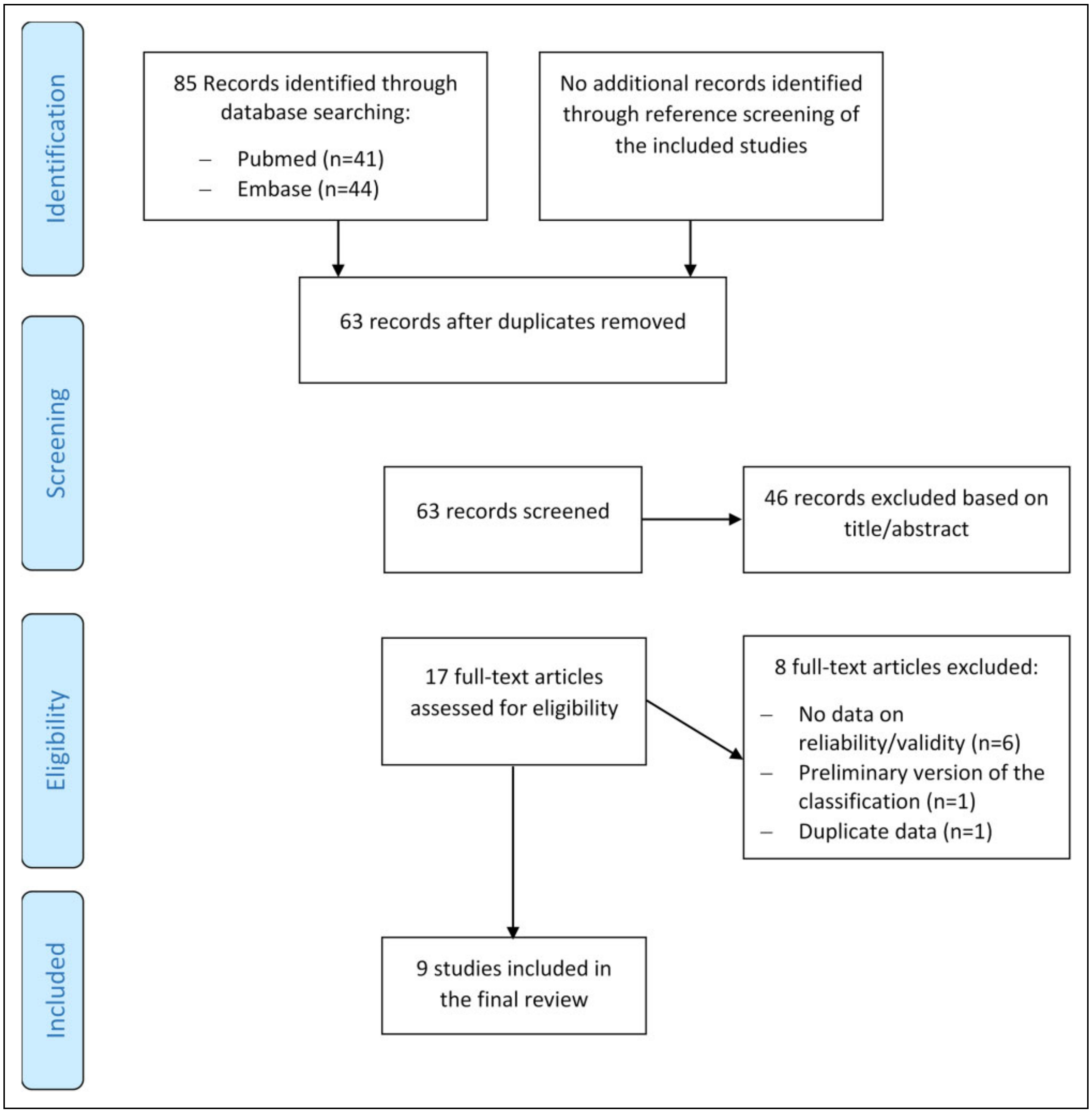
Identification
Screening
Eligibility
Included
85 Records identified through database searching:
– Pubmed (n=41)
– Embase (n=44)
No additional records identified through reference screening of the included studies
63 records after duplicates removed
63 records screened
46 records excluded based on title/abstract
17 full-text articles assessed for eligibility
8 full-text articles excluded:
– No data on reliability/validity (n=6)
– Preliminary version of the classification (n=1)
– Duplicate data (n=1)
9 studies included in the final review

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram[33] of the screening and selection process.

moderate). Interobserver reliability of the subtypes was reported in 4 studies. Subtypes A0 and B3 had positive interobserver reliability in multiple studies (QoE: low), whereas the interobserver reliability of other subtypes was negative in all studies (QoE: moderate).

The interobserver agreement was reported in 7 studies. In some studies, the agreement was assessed between observers and a predefined gold standard, whereas others evaluated the agreement among participating observers. The interobserver agreement of the overall morphologic classification ranged between 0% and 55.5% (Table 2). In all the studies, the agreement improved with exclusion of subtypes and ranged between 12% and 64.8%. The interobserver agreement of the subtypes was reported only in 1 study, ranging from 0% (A2) to 98% (B3).

The intraobserver agreement of the overall morphologic classification was reported in 2 studies, ranging from 75% to 85% (Table 2). With exclusion of subtypes, the agreement increased beyond 85% in both studies. Furthermore, the intraobserver agreement of the main injury types was reported in one study, with 87% agreement for type A, 89% for type B, and 94% for type C.

Among all studies, the article by Kepler et al[23] was unique due to a large number of observers. Furthermore, the findings of this study were analyzed in 2 subsequent studies.[24,25] Kepler et al[23] evaluated the reliability of the ATLICS, using high-quality computed tomography scans of 25 patients and a large group of observers consisting of 100 AOSpine-affiliated surgeons from 5 continents. Although they showed good intraobserver reliability of the overall morphologic classification with exclusion of subtypes (mean kappa=0.81), their findings were low for the overall morphologic classification ($\kappa = 0.68$), type A ($\kappa = 0.57$), and type B ($\kappa = 0.43$) injuries (Table 2).

Table 1. Characteristics of the Included Studies.

		Cases			Observers		
First Author (Year)	Sampling Method	n	Injury Distribution[a]	No. of Injury Levels	n	Characteristics	Imaging Modality
Vaccaro (2013)[3]	Random	n = 40	Type A: 54% Type B: 24% Type C: 22%	NR	n = 9	Spine surgeons	NR
Kepler (2016)[23]	Purposive	n = 25	Type A: 53% Type B:=32.4% Type C: 13.4%	NR	n = 100	Surgeons from Africa, Asia, Europe, North America, and South America	High-quality CT
Azimi (2015)[20]	Random	n = 56	Type A: 41.9% Type B: 28.4% Type C: 29.7%	74	n = 2	Spine surgeons	Plain X-ray, CT, and MRI
Barcelos (2016)[21]	Consecutive	n = 43	Type A: 32.5% Type B: 16.3% Type C: 51.2%	NR	n = 3	Spine surgeons	CT (axial, sagittal, and coronal)
Kaul (2016)[22]	Consecutive	n = 50	Type A: 39.45% Type B: 24% Type C: 36.55%	NR	n = 11	Surgeons from 4 countries. Orthopedic surgeons: n = 10 Neurosurgeon: n = 1	Plain X-ray, CT, and MRI
Sadiqi (2015)[24]	Purposive	n = 25	Subtype A0: 4% Other subtypes: 12% each	NR	n = 100	International group of spine surgeons naïve to the classification. Subgroups (years of experience): Subgroup 1 (≤10 years): n = 30 Subgroup 2 (11-20 years): n = 44 Subgroup 3 (>20 years): n = 26	High-quality CT
Schroeder (2015)[25]	Purposive	n = 6	Subtype A3: n = 3 Subtype A4: n = 3	NR	n = 100	International group of spine surgeons naïve to the classification	High-quality CT
Urrutia (2015)[26]	Purposive	n = 70	Type A: 49.8% Type B: 30% Type C: 20.2%	NR	n = 6	Spine surgeons (fellowship-trained): n = 3 Orthopedic surgery residents: n = 3	Plain X-ray (anteroposterior and lateral), multislice 64 channel CT (axial view and sagittal reconstruction)
Yacoub (2015)[27]	NR	n = 54	NR	NR	n = 2	Spine surgeon: n = 1 Neurosurgery resident: n = 1	Multislice 64-channel CT with reconstruction

Abbreviations: NR, not reported; CT, computed tomography; MRI, magnetic resonance imaging.

[a] Based on AOSpine Thoracolumbar Injury Classification System.[3]

Furthermore, they showed good interobserver reliability for type A ($\kappa = 0.8$), type C ($\kappa = 0.72$), and the overall morphologic classification with exclusion of subtypes ($\kappa = 0.74$). However, the reliability of the overall morphologic classification and type B injuries was insufficient, with kappa values of 0.56 and 0.68, respectively. Meanwhile, except for subtype A0 ($\kappa = 0.96$), all subtypes showed low reliability, with kappa values ranging between 0.19 and 0.68. The overall agreement between all observers for the morphologic classification with and without inclusion of subtypes was 0% and 12%, respectively.

Effect of Observers' Experience, Training and Cultural Background

Sadiqi et al[24] further analyzed the results of the study done by Kepler et al,[23] exploring the differences in reliability of ATLICS among surgeons with low, moderate, and high experience. This study demonstrated negative results for intraobserver reliability of the overall morphologic classification for all 3 groups, with mean kappa values ranging between 0.67 and 0.69 (Table 2). With exclusion of subtypes, the kappa values increased in all groups (mean kappa

Table 2. Summary of the Results of Studies on Reliability and Measurement Error of the AOSpine Thoracolumbar Injury Classification System.

	Findings				
	Intraobserver			Interobserver	
First Author (Year)	Time Interval	Reliability (Kappa Values)	Measurement Error (% Agreement)	Reliability (Kappa Values)	Measurement Error (% Agreement)
Vaccaro (2013)[3]	1 month	Overall: Mean = 0.77 (range: 0.6-0.97) Overall without subtypes: Mean = 0.85 (range: 0.75-0.96) Type A: 0.72 Type B: 0.43		Overall: 0.64 Overall without subtypes: 0.72 Type A: 0.72 Type B: 0.58 Type C: 0.7 Subtypes: 0.34 (B2) to 1 (A0)	Overall: 35% Overall without subtypes: 60%
Kepler (2016)[23]	1 month	Overall: mean = 0.68 (range: 0.22-1) Overall without subtypes: mean = 0.81 (range: 0.32-1.0) Type A: 0.57 Type B: 0.43		Overall: 0.56 Overall without subtypes: 0.74 Type A: 0.80 Type B: 0.68 Type C: 0.72 Subtypes: 0.19 (A4) to 0.96 (A0)	Overall: 0% Overall without subtypes: 12%
Azimi (2015)[20]	5 weeks	Type A: 0.84 (95% CI: 0.82-0.91) Type B: 0.83 (95% CI: 0.81-0.88) Type C: 0.86) 95% CI: 0.83-0.92)		Type A: 0.88 (95% CI: 0.80-0.94) Type B: 0.86 (95% CI: 0.83-0.93) Type C: 0.89 (95% CI: 0.84-0.94)	
Barcelos (2016)[21]					
First assessment				Overall without subtypes: 0.526 Type A: 0.535 Type B: 0.215 Type C: 0.654	
Second assessment				Overall without subtypes: 0.645 Type A: 0.763 Type B: 0.230 Type C: 0.688	
Kaul (2016)[22]	6 weeks	Overall: 0.61 (SE = 0.13) Overall without subtypes: 0.68 (SE = 0.13) Neurological injury: 0.91 (SE = 0.08)		Overall: 0.45 (SE = 0.01) Overall without subtypes: 0.59 (SE = 0.01) Type A: 0.64 Type B: 0.40 Type C: 0.71 Neurological injury: 0.85 (SE = 0.01)	Overall: 32%
Sadiqi (2015)[24]	1 month	Overall: Subgroup 1: Mean = 0.69 (range: 0.44-0.91) Subgroup 2: Mean = 0.69 (range: 0.22-1.00) Subgroup 3: Mean = 0.67 (range: 0.31-0.85) Overall without subtypes: Subgroup 1: Mean = 0.83 (range: 0.53-1.00) Subgroup 2: Mean = 0.81 (range: 0.32-1.00)			Agreement with predefined gold standard: First assessment: Whole group: 26% of observers had ≥80% agreement Subgroup 1: 32%-96% 30.0% of observers had ≥80% agreement

(continued)

Table 2. (continued)

	Findings				
	Intraobserver			Interobserver	
First Author (Year)	Time Interval	Reliability (Kappa Values)	Measurement Error (% Agreement)	Reliability (Kappa Values)	Measurement Error (% Agreement)
		Subgroup 3: Mean = 0.79 (range: 0.51-1.00)			Subgroup 2: 40%-92% 31.8% of observers had ≥80% agreement Subgroup 3: 32%-88% 11.5% of observers had ≥80% agreement. Second assessment: Compared with first assessment: range of agreements was comparable, all subgroups had smaller proportion of observers with ≥80% agreement.
Schroeder (2015)[25]					Agreement with predefined gold standard: A3: 59.3% A4: 30.0% A3 and A4: 44.7%
Urrutia (2015) [26]	6 weeks	Overall: 0.71 (95% CI: 0.67-0.76) Overall without subtypes: Whole group: 0.77 (95% CI: 0.72-0.83) Surgeons: 0.81 (95% CI: 0.74-0.87) Residents: 0.74 (95% CI: 0.65-0.82)	Overall: 75.71% Overall without subtypes: 85.95%	Overall: 0.55 (95% CI: 0.52-0.57) Overall without subtypes: 0.62 (95% CI: 0.57-0.66) Type A: 0.61 (95% CI: 0.55-0.67) Type B: 0.57 (95% CI: 0.51-0.63) Type C: 0.69 (95% CI: 0.63-0.75) Sub-types: 0.18 (B1) to 0.94 (A0)	Overall: 30% Overall without subtypes: 54.28%
Yacoub (2015)[27]	8 weeks	Type A: 0.75 Type B: 0.7 Type C: 0.85	Overall: Surgeon 1: 85% Surgeon 2: 75% Overall without subtypes: 88% Type A: 0.87% Type B: 0.89% Type C: 0.94%	Subtypes: 0 (A2) to 0.85 (C)	Overall: 55.5% Overall without subtypes (4 assessments): 64.8% Subtypes: 0% (A2) to 98% (B3)

Abbreviations: CI, confidence interval; SE, standard error.

Table 3. Summary of Evidence on Intraobserver Reliability of the AOSpine Thoracolumbar Injury Classification System.

		Quality of Findings[a]						
		Morphologic Classification						
First Author (Year)	Study Quality	Overall	Overall Without Subtypes	Type A	Type B	Type C	Neurologic Injury	Modifiers
Vaccaro (2013)[3]	Fair	+	+	+	−	0	0	0
Kepler (2016)[23]	Fair	−	+	−	−	0	0	0
Azimi (2015)[20]	Fair	0	0	+	+	+	0	0
Kaul (2016)[22]	Fair	−	−	0	0	0	+	0
Sadiqi (2015)[24]	Fair	−	+	0	0	0	0	0
Urrutia (2015)[26]	Fair	+	+	0	0	0	0	0
Yacoub (2015)[27]	Fair	0	0	+	+	+	0	0
Overall quality of findings[a]		Conflicting	+	Conflicting	Conflicting	+	+	0
Overall quality of evidence[b]		Low	Moderate	Low	Low	Moderate	Low	Unknown

[a] +, positive rating; ?, indeterminate rating; −, negative rating; 0, not reported.[18]
[b] According to the criteria by Prinsen et al[19] (Appendix D).

Table 4. Summary of Evidence on Interobserver Reliability of the AOSpine Thoracolumbar Injury Classification System.

		Quality of Findings[a]						
		Morphologic Classification						
First Author (Year)	Study Quality	Overall	Overall Without Subtypes	Type A	Type B	Type C	Neurologic Classification	Modifiers
Vaccaro (2013)[3]	Fair	−	+	+	−	+	0	0
Kepler (2016)[23]	Fair	−	+	+	−	+	0	0
Azimi (2015)[20]	Fair	0	0	+	+	+	0	0
Barcelos (2016)[21] (2 assessments)	Fair	0	−	−/+[b]	−	−	0	0
Kaul (2016)[22]	Fair	−	−	−	−	+	+	0
Urrutia (2015)[26]	Fair	−	−	−	−	−	0	0
Yacoub (2015)[27]	Fair	0	0	0	0	0	0	0
Overall quality of findings[a]		−	Conflicting	Conflicting	−	Conflicting	+	0
Overall quality of evidence[c]		Moderate	Low	Low	Moderate	Low	Low	Unknown

[a] +, positive rating; ?, indeterminate rating; −, negative rating; 0, not reported.[18]
[b] Two assessments had conflicting findings.
[c] According to the criteria by Prinsen et al[19] (Appendix D).

range: 0.79 and 0.83), showing good intraobserver reliability of the classification containing only the main injury types. Furthermore, they evaluated the agreement between observers and a predefined gold standard of the morphologic classification, which is considered as interobserver agreement. Considering the whole group of observers, the percentage agreement ranged between 32% and 96%, and only 26% of all observers reached high agreement, defined as agreement in at least 80% of cases. Regarding the subgroups, the proportion of observers with a high percentage of agreement was 30% in surgeons with low experience, 31.8% in those with moderate experience, and as low as 11.5% in highly experienced observers. In their second assessment of interobserver agreement 1 month later, although the range of percentage agreements was comparable with the first assessment, a smaller number of observers agreed in all subgroups. The authors concluded that the ATLICS yielded similar intraobserver reliability when used by observers with different levels of experience, although they did not statistically compare the findings between groups.

In another study, Schroeder et al[25] analyzed a subsample of A3 and A4 cases from the study done by Kepler et al,[23] focusing on the effect of observers' culture and experience on the classification of burst fractures. In their study, although the percentage of agreement for classification of A3 injuries was significantly higher than that of A4 in all world regions ($P < .01$) and all subgroups with different levels of experience ($P < .01$), they found no statistically significant difference in classification of burst fractures between observers with different levels of experience or from different regions (Table 2).

The effect of the level of training on the reliability of ATLICS was further evaluated in the study of Urrutia et al,[26] by comparing the kappa values of orthopedic surgery residents and spine surgeons. They found insignificant differences regarding the inter- and intraobserver reliability of the overall morphologic classification and main types of injury.

Discussion

Current review is an evaluation of the methodology and findings of studies on measurement properties of ATLICS, rather than the classification itself. Based on the findings of this study, the evidence was insufficient for making conclusions about the measurement properties of ATLICS. Reliability and validity of a measurement instrument constitute one of the most important aspects of its overall quality, and measures with poor measurement properties may become a major source of bias.[14,28] Therefore, following the development of a new measure, critical appraisal of the evidence on its measurement properties is of utmost importance.

There was no study on content and construct validity of ATLICS. Meanwhile, good content validity is considered as one of the most important criteria that should be met during the selection of measures.[19,29] It has been suggested that measures with unknown content validity should be avoided in the first place.[19] Therefore, evaluation of the content validity of ATLICS by an independent expert panel is crucial.

There was no study on reliability of the modifiers and only limited evidence on reliability of the neurologic injury classification. Considering the reliability of the overall morphologic classification, percentages of agreement were highly variable and intraobserver studies showed conflicting findings. Moreover, there was moderate evidence for insufficient interobserver reliability of the overall morphologic classification. However, with exclusion of subtypes, reliability of the morphologic classification improved in most studies. The underlying reasons for the insufficient findings may stem from the sources of variation that influence the reliability of measurements, such as patients, observers, and the measurement instruments.[30,31] It has been demonstrated that observers' experience and cultural background do not affect the reliability of ATLICS,[24-26] which is also a major advantage of this classification. Meanwhile, in all studies the same static images were used for repeated assessments. Therefore, it is more likely that the reliability of ATLICS might be negatively affected by inherent characteristics of this classification, such as extensive complexity, excessive number of subtypes or lengthy training requirement. Nonetheless, the exact source of negative results remains unknown and these results might be biased due to low quality of the studies.

The generalizability of the findings of reliability studies was limited. Generalizability relies on many factors, including the characteristics of the patients and observers.[31] In all included studies, the observers were exclusively spine surgeons. Therefore, it is not clear if their findings are generalizable to observers from other disciplines, including the radiologists. Furthermore, some studies used a purposive sampling method, aiming to include all injury categories. However, in practice, minor spine injuries are usually more prevalent and sample populations are more skewed toward lower injury intensities.[23,24] Therefore, the purposive sampling approach further limits the generalizability of the findings. Besides, when evaluating the reliability of a measure in a heterogeneous sample, the kappa values are theoretically higher compared to homogeneous samples.[31] Therefore, studies with purposive samples may have overestimated the reliability of the ATLICS.

The most common methodological limitation of the studies was their statistical method, particularly due to the use of unweighted kappa. In contrast to the weighted kappa, simple kappa method fails to discriminate different levels of disagreement.[32] Using simple kappa is problematic, particularly in situations in which different disagreements have different consequences.[32] For example, in the context of spine injuries, misclassification of a less severe injury to a distant injury subtype may result in unnecessary surgical treatment, while misclassification to an adjacent subtype may result in less serious consequences. Therefore, it has been recommended to use the weighted kappa method for evaluation of the reliability of ordinal measures.[30] Meanwhile, this issue might be due to poor reporting, that is, the authors may have used the weighted kappa but failed to report it properly. Although quadratic weighting is the most common scheme, linear weighting seems more appropriate for ATLICS, since it has been shown that surgeons consider an equal progression of injury severity between almost all adjacent pairs of morphologic subtypes.[6]

Conclusions

The exclusion of the morphologic subtypes improved the reliability of ATLICS classification in most studies, resulting in acceptable interobserver reliability of the morphologic classification, and suggesting the simplification of ATLICS as an option for improvement of its reliability. However, it was difficult to draw a clear conclusion since validity studies were missing and reliability studies included in this review had inconsistent findings and methodological limitations. Since majority of the studies were performed by ATLICS developers, further assessments by independent investigators are recommended. ATLICS is an improvement over its predecessors as it includes their strengths and is likely to be increasingly used in future research. Furthermore, the extensive development process of ATLICS indicates a promising framework. Therefore, high-quality studies are warranted to reveal the advantages of this novel classification system.

Appendix A

Search Strategies

A. PubMed

1. AOSpine OR AO-Spine
2. Injur* OR Injury[MeSH Terms]
3. Fracture* OR Fracture[MeSH Terms]
4. classification OR classification[MeSH Terms]
5. 2 OR 3
6. 1 AND 4 AND 5

B. Embase

1. 'aospine' OR 'ao-spine'
2. 'injury'/exp OR 'injur*'
3. 'fracture'/exp OR 'fracture*'
4. 'classification'/exp OR 'classification*'
5. 2 OR 3
6. 1 AND 4 AND 5

Appendix B

Adapted COSMIN Checklist for Evaluation of the Methodological Quality of Studies on Reliability and Measurement Error of Ordinal Classification Systems.

Revised COSMIN Checklist—Reliability and Measurement Error

		Excellent	Good	Fair	Poor	Not Applicable
Design requirements						
1	Were at least 2 measurements available?	At least 2 measurements			Only 1 measurement	
2	Were the administrations independent?	Independent measurements	Assumable that the measurements were independent	Doubtful whether the measurements were independent	measurements NOT independent	
3	Was the time interval stated?	Time interval stated		Time interval NOT stated		*
4	Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable	*
5	Were observers stable in the interim period?	Observers were stable (evidence provided)	Assumable that observers were stable	Unclear if observers were stable	Observers were NOT stable, eg, received additional training	*
6	Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate	*
7	Were the test conditions similar for both measurements? For example, type of administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar	
8	Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study, e.g. lack of blinding regarding the clinical information	Other important methodological flaws in the design or execution of the study	
Statistical methods						
9	Reliability studies: Was kappa calculated?	Kappa calculated			Kappa not calculated	
10	Reliability studies: Was a weighted kappa calculated?	Weighted Kappa calculated		Unweighted kappa calculated		
11	Reliability studies: Was the weighting scheme described? For example, linear, quadratic	Weighting scheme described	Weighting scheme NOT described			*
12	Measurement error studies: Was percentage agreement calculated?	Percentage agreement calculated			Percentage agreement not calculated	

Adapted from Terwee et al[17] under a Creative Commons Attribution–Noncommercial (http://creativecommons.org/licenses/by-nc/4.0/).

Appendix C

Criteria for Evaluation of the Quality of Results.

Measurement Property	Rating[a]	Criteria
Content validity (including face validity)	+	All items refer to relevant aspects of the construct to be measured AND are relevant for the target population AND are relevant for the context of use AND together comprehensively reflect the construct to be measured.
	?	Not all information for "+" reported.
	−	Criteria for "+" not met.
Reliability	+	ICC or weighted kappa $\geq$0.70.
	?	ICC or weighted kappa not reported.
	−	Criteria for "+" not met.
Measurement error	+	SDC or LoA < MIC.
	?	MIC not defined.
	−	Criteria for "+" not met.
Construct validity	+	At least 75% of the results are in accordance with the hypotheses.
	?	No correlations with instrument(s) measuring related construct(s) AND no differences between relevant groups reported.
	−	Criteria for "+" not met.

Adapted from Prinsen et al[19] (as modified from Terwee et al[18]) under a Creative Commons Attribution 4.0 (http://creativecommons.org/licenses/by/4.0/).
Abbreviations: ICC, intraclass correlation coefficient; SDC, smallest detectable change; LoA, limits of agreement; MIC, minimal important change.
[a] +, positive rating; ?, indeterminate rating; −, negative rating.

Appendix D

Criteria for Evaluation of the Quality of Evidence.

Quality Rating	Criteria
High	Consistent findings in multiple studies of at least good quality OR one study of excellent quality AND a total sample size of $\geq$100 patients
Moderate	Conflicting findings in multiple studies of at least good quality OR consistent findings in multiple studies of at least fair quality OR one study of good quality AND a total sample size of $\geq$50 patients
Low	Conflicting findings in multiple studies of at least fair quality OR one study of fair quality AND a total sample size of $\geq$30 patients
Very low	Only studies of poor quality OR a total sample size of <30 patients
Unknown	No studies

Reused from Prinsen et al. [19] under a Creative Commons Attribution 4.0 (http://creativecommons.org/licenses/by/4.0/).

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Disclosures outside of submitted work: ZB-consultancy: Xenco Medical, AO Spine; Research Support: SeaSpine (paid to the institution) JCW-JCW – Royalties – Biomet, Seaspine, Amedica, DePuy Synthes; Investments/Options – Fziomed, Promethean, Paradigm Spine, Benvenue, Nexgen, Vertiflex, Electrocore, Surgitech, Expanding Orthopedics, Osprey, Bone Biologics, Pearldiver; Board of Directors - North American Spine Society, North American Spine Foundation, AO Foundation, Cervical Spine Research Society; Fellowship Funding (paid to institution): AO Foundation.

Funding

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

References

1. Watson-Jones R. The results of postural reduction of fractures of the spine. *J Bone Joint Surg Am*. 1938;20:567-586.
2. Schroeder GD, Harrop JS, Vaccaro AR. Thoracolumbar trauma classification. *Neurosurg Clin N Am*. 2017;28:23-29.
3. Vaccaro AR, Oner C, Kepler CK, et al; AOSpine Spinal Cord Injury & Trauma Knowledge Forum. AOSpine thoracolumbar spine injury classification system: fracture description, neurological status, and key modifiers. *Spine (Phila Pa 1976)*. 2013;38:2028-2037.
4. Magerl F, Aebi M, Gertzbein SD, Harms J, Nazarian S. A comprehensive classification of thoracic and lumbar injuries. *Eur Spine J*. 1994;3:184-201.
5. Vaccaro AR, Lehman RA Jr, Hurlbert RJ, et al. A new classification of thoracolumbar injuries: the importance of injury morphology, the integrity of the posterior ligamentous complex, and neurologic status. *Spine (Phila Pa 1976)*. 2005;30:2325-2333.
6. Kepler CK, Vaccaro AR, Schroeder GD, et al. The thoracolumbar AOSpine Injury Score. *Global Spine J*. 2016;6:329-334.
7. Schroeder GD, Kepler CK, Koerner JD, et al. Can a thoracolumbar injury severity score be uniformly applied from T1 to L5 or are modifications necessary? *Global Spine J*. 2015;5:339-345.
8. Schroeder GD, Vaccaro AR, Kepler CK, et al. Establishing the injury severity of thoracolumbar trauma: confirmation of the hierarchical structure of the AOSpine Thoracolumbar Spine Injury Classification System. *Spine (Phila Pa 1976)*. 2015;40:E498-E503.
9. Vaccaro AR, Schroeder GD, Kepler CK, et al. The surgical algorithm for the AOSpine thoracolumbar spine injury classification system. *Eur Spine J*. 2016;25:1087-1094.
10. Prinsen CAC, Mokkink LB, Bouter LM, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res*. 2018;27:1147-1157. doi:10.1007/s11136-018-1798-3.
11. Mokkink LB, Prinsen CA, Bouter LM, Vet HC, Terwee CB. The COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) and how to select an outcome measurement instrument. *Braz J Phys Ther*. 2016;20:105-113.
12. Moher D, Shamseer L, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev*. 2015;4:1.

13. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol*. 2010;63:737-745.
14. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual Life Res*. 2010;19:539-549.
15. Lucas NP, Macaskill P, Irwig L, Bogduk N. The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). *J Clin Epidemiol*. 2010;63:854-861.
16. Audige L, Bhandari M, Kellam J. How reliable are reliability studies of fracture classifications? A systematic review of their methodologies. *Acta Orthop Scand*. 2004;75:184-194.
17. Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Qual Life Res*. 2012;21:651-657.
18. Terwee CB, Bot SD, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol*. 2007;60:34-42.
19. Prinsen CA, Vohra S, Rose MR, et al. How to select outcome measurement instruments for outcomes included in a "Core Outcome Set"—a practical guideline. *Trials*. 2016;17:449.
20. Azimi P, Mohammadi HR, Azhari S, Alizadeh P, Montazeri A. The AOSpine thoracolumbar spine injury classification system: A reliability and agreement study. *Asian J Neurosurg*. 2015;10: 282-285.
21. Barcelos AC, Joaquim AF, Botelho RV. Reliability of the evaluation of posterior ligamentous complex injury in thoracolumbar spine trauma with the use of computed tomography scan. *Eur Spine J*. 2016;25:1135-1143.
22. Kaul R, Chhabra HS, Vaccaro AR, et al. Reliability assessment of AOSpine thoracolumbar spine injury classification system and Thoracolumbar Injury Classification and Severity Score (TLICS) for thoracolumbar spine injuries: results of a multicentre study. *Eur Spine J*. 2017;26:1470-1476.
23. Kepler CK, Vaccaro AR, Koerner JD, et al. Reliability analysis of the AOSpine thoracolumbar spine injury classification system by a worldwide group of naive spinal surgeons. *Eur Spine J*. 2016; 25:1082-1086.
24. Sadiqi S, Oner FC, Dvorak MF, Aarabi B, Schroeder GD, Vaccaro AR. The influence of spine surgeons' experience on the classification and intraobserver reliability of the novel AOSpine Thoracolumbar Spine Injury Classification System—an international study. *Spine (Phila Pa 1976)*. 2015;40: E1250-E1256.
25. Schroeder GD, Kepler CK, Koerner JD, et al. Is there a regional difference in morphology interpretation of A3 and A4 fractures among different cultures? *J Neurosurg Spine*. 2015:1-8.
26. Urrutia J, Zamora T, Yurac R, et al. An independent interobserver reliability and intraobserver reproducibility evaluation of the new AOSpine Thoracolumbar Spine Injury Classification System. *Spine (Phila Pa 1976)*. 2015;40:E54-E58.
27. Yacoub AR, Joaquim AF, Ghizoni E, Tedeschi H, Patel AA. Evaluation of the safety and reliability of the newly-proposed AO spine injury classification system. *J Spinal Cord Med*. 2017;40:70-75.
28. Audige L, Bhandari M, Hanson B, Kellam J. A concept for the validation of fracture classifications. *J Orthop Trauma*. 2005;19: 401-406.
29. van Middendorp JJ, Audige 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.
30. Wright JG, Feinstein AR. Improving the reliability of orthopaedic measurements. *J Bone Joint Surg Br*. 1992;74:287-291.
31. De Vet HCW, Terwee CB, Mokkink LB, Knol DL. Reliability. In: *Measurement in Medicine: A Practical Guide*. New York, NY: Cambridge University Press; 2011:96-149.
32. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull*. 1968; 70:213-220.
33. Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *J Clin Epidemiol*. 2009;62: 1006-1012.