European Journal of Orthopaedic Surgery & Traumatology (2019) 29:1125–1129
https://doi.org/10.1007/s00590-019-02404-8

ORIGINAL ARTICLE • ANKLE - FRACTURES

Reproducibility of the stability-based classification for ankle fractures

Bjarke Viberg[1,2,3] · Tamim A. Haidari[1] · Jesper Stork-Hansen[1] · Roland Knudsen[1] · Rune Dueholm Bech[1]

Received: 21 November 2018 / Accepted: 19 February 2019 / Published online: 21 February 2019

Abstract

Background Classification of ankle fracture is important when deciding for operative or conservative treatment. This study rates the reproducibility of ankle stability assessment and compares it with the classification by Lauge-Hansen and Arbeitsgemeinschaft für Osteosyntesefragen (AO) in adult patients with primary ankle fractures.

Methods A total of 496 consecutive ankle fractures were included, and the X-ray images were reviewed 2 times by 2 medical students, 2 residents, and 1 consultant in orthopedic traumatology. The raters were blinded to each other and to their own results. Unweighted Kappa statistics were used to assess reproducibility.

Results Overall mean (95% CI) interrater Kappa results were 0.65 (0.64; 0.68) for Lauge-Hansen, 0.62 (0.60; 0.63) for AO and 0.61 (0.57; 0.62) for the stability assessment. The intrarater results ranged from a mean Kappa of 0.64–0.80 for the medical students, 0.65–0.81 for the residents and 0.82–0.84 for the consultant.

Conclusion The stability assessment has substantial to almost-perfect agreement which is comparable to the Lauge-Hansen and AO classifications.

Keywords Reproducibility · Stability · Ankle fracture · Classification

Introduction

Ankle fracture is a very common fracture and represents approximately 9% of all fractures with an incidence of 138-169/100,000/year [1–3]. Many of these fractures are treated operatively, but there are controversies in which fracture types should be offered operative treatment [4]. There is probably a correlation between the severity of the ankle fracture and complications [5], and reliable fracture classification systems are needed for indication of operative treatment.

The two most used classification systems for ankle fractures are Lauge-Hansen and AO [6, 7]. The original work of Lauge-Hansen from 1940 to 1950s describes the mechanism of injury, while AO classifies the ankle fractures according to the position of the lateral malleolus fracture and integrity of the syndesmosis (infrasyndesmotic, transsyndesmotic, suprasyndesmotic) based on X-ray images. These two classification systems do not seem to have good reproducibility with Kappa values ranging from 0.17 to 0.48 [8–10], and the value is therefore limited. A simple and more functional approach to classify an ankle fracture is to assess the stability of the fracture. Pakarinen et al. [11] described a new method to classify ankle fractures as stable or unstable fractures by using X-rays and, if necessary, a dynamic examination. By using the stability-based classification criteria's, the number of operatively treated ankle fractures may decrease and is therefore cost-effective [12]. However, in order to use a new classification system, the reproducibility of the classification system needs to be assessed.

The primary aim of this study was to assess the reproducibility of the ankle fracture classification systems Lauge-Hansen, AO and the new stability assessment in adult patients with primary ankle fractures. Secondarily, our aim was to assess the impact of clinical experience when classifying the ankle fractures.

✉ Bjarke Viberg
Bjarke.viberg@rsyd.dk

1 Department of Orthopaedic Surgery and Traumatology, Odense University Hospital, Odense, Denmark

2 Department of Orthopaedic Surgery and Traumatology, Kolding Hospital – part of Hospital Lillebaelt, Kolding, Denmark

3 Institute for Regional Health Research, University of Southern Denmark, Odense, Denmark

Method

The study was designed as a consecutive cohort design with data from a Level 1 university hospital. The study was approved by the Danish Data Protection Agency (case number 14/38794), and reporting is performed according to GRRAS [13].

Patients

From 2010 to 2011, data were retrieved from the county database using ICD-10 diagnoses codes for ankle fractures (DS824-DS828). The patients' health records and radiographs were reviewed for the inclusion criteria's:

- Radiographs with adequate quality for classification
- Ankle fracture
- Primary fracture

The exclusion criteria's were

- Age under 18
- Other fractures than ankle
- Inadequate radiographs
- Subsequent fractures of the same ankle in the same patient.

During the study period, 692 radiographs were reviewed by two authors and 496 X-rays compromised the final cohort.

Radiograph assessment

Radiographs were evaluated 2 times by 2 medical students (STA/STB), 2 residents (REA/REB) and 1 consultant in orthopedic traumatology (OC) with a minimum of 14-day interval between the assessments. The data were registered using EpiData Software [14]. All radiographs were measured independently, and any patient identifiers were removed. The raters were blinded to each other and to their own results.

Guide sheets illustrating the classifications and subclassifications were provided prior to assessment to all raters. Each ankle fracture was classified according to Lauge-Hansen as either supination type I–II (S I–II), supination eversion type I–IV (SE I–IV), pronation type I–II (P I–II) or pronation eversion type I–IV (PE I–IV) [6]. Classification by AO (either 44 A1-3, 44 B1-3, or 44 C1-3) and stability was classified as either stable, unstable or test for syndesmosis stability [7, 11]. The stability assessment criteria are given in Table 1. Since it was not possible to do a stress test in our study set up, a category for stress test was added for X-rays with a hematoma on the medial side (Fig. 1).

Statistical analysis

The statistical software program Stata 15 was used for the analyses. Unweighted Kappa statistics were used to assess reproducibility, thereby treating all disagreements equally and correcting measurement agreement for chance [15]. Analyses were performed for intrarater and interrater agreement. The interraters' agreement was based on the first measurement for all raters. A sample size of 489 patients was calculated using 0.7 (rater 1) and 0.66 (rater 2) as proportion of a positive rating, 0.85 Kappa detection level and a precision of 0.05 (CI 0.95).

Guidelines proposed by Landis and Koch [16] were used to categorize inter/intrarater Kappa agreement: poor (< 0), slight (0–0.2), fair (0.21–0.4), moderate (0.41–0.6), substantial (0.61–0.8) and almost perfect (0.81–1).

Results

Of the 496 radiographs, 187 fractures (38%) were classified as SE2 according to the Lauge-Hansen classification and 141 fractures (28%) were SE4 which were the two most common fracture types (Table 2). For AO, the most common fracture was the 44-B1 with 183 fractures (37%).

The combined interrater mean (95% confidence interval) Kappa values were 0.65 (0.64; 0.68) for Lauge-Hansen, 0.62 (0.60; 0.63) for AO and 0.61 (0.57; 0.62) for stability assessment (Table 2). The individual interrater Kappa values for all

Table 1 Overview of the stability assessment criteria

	Stable	Unstable
Fracture	Isolated lateral malleolar fracture Isolated non-displaced medial malleolar fractures	High lateral malleolar fracture Displaced medial malleolar fracture Bi- or trimalleolar fracture
Talar shift	No talar shift	Positive talar shift
Medial tenderness and stress test	No medial tenderness Medial tenderness and negative stress test	Medial tenderness and positive stress test

Fig. 1 X-rays of different malleolar fractures: **a** is a stable fracture (LH SE2, AO B1). **b** is similar to **a** but has a medial hematoma and therefore placed in the category for stress test. **c** is an unstable fracture (LH SE4, AO B2), and **d** is also an unstable fracture (PE3, AO C2)

Table 2 Interrater Kappa results for Lauge-Hansen, AO and stability assessment

Lauge-Hansen	*n* (mean)	Kappa	AO	*n* (mean)	Kappa	Stability	*n* (mean)	Kappa
S1	67	0.86	A1	68	0.86	Yes	241	0.66
S2	12	0.26	A2	20	0.63	No	209	0.74
SU1	1	−0.00	A3	1	−0.16	Test	46	0.12
SU2	187	0.73	B1	184	0.71	NC	15	−0.02
SU3	8	0.19	B2	73	0.32			
SU4	142	0.67	B3	87	0.60			
P1	7	0.28	C1	32	0.55			
P2	2	0.20	C2	21	0.58			
P3	7	0.04	C3	5	0.71			
PU1	5	0.14	NC	18	0.27			
PU2	0	−0.00						
PU3	39	0.65						
PU4	15	0.53						
NC	17	0.22						
Total (95% CI)		0.65 (0.64; 0.68)			0.62 (0.60; 0.63)			0.61 (0.57; 0.62)

Number of radiographs in each category is calculated as a mean from the 5 raters

NC not classified, *Test* indication test for syndesmosis stability, *CI* confidence interval

raters ranged from 0.49 to 0.75 for AO, 0.55–0.72 for Lauge-Hansen and 0.51–0.65 for stability assessment.

Intrarater Kappa values were 0.64–0.82 for AO, 0.68–0.84 for Lauge-Hansen and 0.67–0.84 for stability assessment (Table 3). In general, the consultant had slightly higher intrarater Kappa values ranging from 0.82 to 0.84 in all 4 categories, while the students ranged from 0.64 to 0.80 and the residents ranged from 0.65 to 0.81.

Discussion

We have demonstrated that the stability assessment has substantial to almost-perfect agreement for intra- and interrater assessment which is comparable to the Lauge-Hansen and AO classification. By limiting the classifications categories as the stability assessment tool does, we

Table 3 Interrater Kappa values are shown in normal text and intrarater in italic

Raters	STA	STB	REA	REB	OC
AO					
STA	*0.73*	0.49	0.67	0.51	0.61
STB		*0.64*	0.55	0.49	0.52
REA			*0.81*	0.60	0.75
REB				*0.65*	0.60
OC					*0.82*
Lauge-Hansen					
STA	*0.80*	0.55	0.71	0.60	0.62
STB		*0.69*	0.57	0.60	0.55
REA			*0.80*	0.66	0.72
REB				*0.68*	0.63
OC					*0.84*
Stability assessment					
STA	*0.77*	0.64	0.53	0.61	0.65
STB		*0.67*	0.51	0.63	0.62
REA			*0.76*	0.60	0.61
REB				*0.70*	0.65
OC					*0.84*

STA + STB Student A + B, *REA + B* Resident A + B, *OC* orthopedic consultant

expected that the reproducibility would be higher than for the Lauge-Hansen and AO classification. However, we could not demonstrate that. Instead we demonstrated large variations in Kappa values ranging from poor to almost-perfect agreement. This could indicate that some fractures are easier to diagnose than other fractures. The lowest Kappa values (more than 1 event) were 0.04 for Lauge-Hansen P3 and 0.16 for PU1. These were rarely occurring fractures and might therefore be difficult to diagnose, but the AO C3 had a Kappa value of 0.71 and only 5 fractures. The C3 fracture is though very different from the rest and could be expected easier to diagnose. However, we do not find perfect agreement. The C3 fractures are unstable (if displaced) according to the stability assessment and with fewer categories compared to AO and we had therefore hoped for better agreement. The reason lies within the cohort as the majority of fractures are Lauge-Hansen SU2-4 or AO B1-3. The difference between stable and unstable fractures in these fracture types might be very small, and the uncertainty in the stability assessment must be due to these fracture types.

Only Cordova et al. [17] have previously assessed the reproducibility of the stability assessment using 28 X-ray images and 3 army physicians. They demonstrated an interrater agreement of 0.36 for the stability assessment and 0.21 for Lauge-Hansen and 0.51 for AO. Two major drawbacks are the lack of an intrarater assessment. Moreover, the study is underpowered. Thomsen et al. [18] compared Lauge-Hansen with the AO classification and demonstrated interrater Kappa values of 0.49 for Lauge-Hansen (intrarater 0.60–0.70) and 0.58 for AO (intrarater 0.60–0.71). These results are comparable to ours, and we consider our results to be with good external validity.

The consultant in our study had higher intrarater Kappa values in all classification systems compared to the residents and students but not by a large margin. Students have a lower intrarater agreement, but via concrete instructions they can reach an equal level of agreement as more experience raters [19]. The participating students followed a guide sheet for each classification system, and perhaps therefore they had a very small difference compared to the consultant. Hence, young doctors and students can achieve similar agreement compared to consultants by following clear instructional media.

A limitation in our study is the unequal distribution of events due to the consecutive nature. However, this is also a strength since it reflects the everyday clinic. Another limitation is the fact that the raters knew their evaluations would be analyzed and they might have been more thorough in their assessment. Lastly, we used an extra category for stress testing in the stability assessment for medial hematoma thereby not including medial tenderness. However, we do not think that the inclusion of clinical data would change the results. The major strength in our study is the study size which is at least 400 more than in the previous studies. This combined with a sample size makes our results highly reliable, and bias is reduced.

In conclusion, the new stability assessment has substantial to almost-perfect agreement which is comparable to the Lauge-Hansen and AO classifications.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Court-Brown CM, Caesar B (2006) Epidemiology of adult fractures: a review. Injury 37(8):691
2. Court-Brown CM, Biant L, Bugler KE, McQueen MM (2014) Changing epidemiology of adult fractures in Scotland. Scott Med J 59(1):30
3. Elsoe R, Ostgaard SE, Larsen P (2018) Population-based epidemiology of 9767 ankle fractures. Foot Ankle Surg 24(1):34
4. Toth MJ, Yoon RS, Liporace FA, Koval KJ (2017) What's new in ankle fractures. Injury 48(10):2035
5. Broos PL, Bisschop AP (1991) Operative treatment of ankle fractures in adults: correlation between types of fracture and final results. Injury 22(5):403

6. Lauge-Hansen N (1950) Fractures of the ankle. II. Combined experimental-surgical and experimental-roentgenologic investigations. Arch Surg 60(5):957
7. Müller ME, Nazarian S, Koch P, Schatzker J (1990) The comprehensive classification of fractures of long bones. Springer, Berlin
8. Fonseca LLD, Nunes IG, Nogueira RR, Martins GEV, Mesencio AC, Kobata SI (2018) Reproducibility of the Lauge-Hansen, Danis-Weber, and AO classifications for ankle fractures. Rev Bras Ortop 53(1):101
9. Alexandropoulos C, Tsourvakas S, Papachristos J, Tselios A, Soukouli P (2010) Ankle fracture classification: an evaluation of three classification systems: Lauge-Hansen, A.O. and Broos-Bisschop. Acta Orthop Belg 76(4):521
10. Verhage SM, Rhemrev SJ, Keizer SB, Quarles van Ufford HM, Hoogendoorn JM (2015) Interobserver variation in classification of malleolar fractures. Skelet Radiol 44(10):1435
11. Pakarinen HJ, Flinkkil TE, Ohtonen PP, Ristiniemi JY (2011) Stability criteria for nonoperative ankle fracture management. Foot Ankle Int 32(2):141
12. Michelson JD (2013) Using decision analysis to assess comparative clinical efficacy of surgical treatment of unstable ankle fractures. J Orthop Trauma 27(11):642
13. Kottner J, Audige L, Brorson S, Donner A, Gajewski BJ, Hrobjartsson A, Roberts C et al (2011) Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. J Clin Epidemiol 64(1):96
14. EpiData Association. Available at: http://www.epidata.dk/. Accessed 1 Jan 2010
15. Sim J, Wright CC (2005) The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther 85(3):257
16. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33(1):159
17. Cordova CB, Dunn JC, Kusnezov N, Bader JM, Waterman BR, Orr JD (2018) Comparing clinician use of three ankle fracture classifications. JAAPA 31(2):36
18. Thomsen NO, Overgaard S, Olsen LH, Hansen H, Nielsen ST (1991) Observer variation in the radiographic classification of ankle fractures. J Bone Joint Surg Br 73(4):676
19. Viberg B, Bartholin ML, Weber K, Bech RD, Palm H, Schultz-Larsen M (2016) High reliability of a scoring system for implant position in undisplaced femoral neck fractures. J Orthop Trauma 30(8):432