Article published online: 2020-12-18

Evaluation of the Reproducibility of Lauge-Hansen, Danis-Weber, and AO Classifications for Ankle Fractures*

Avaliação da reprodutibilidade das classificações Lauge-Hansen, Danis-Weber e AO para fraturas do tornozelo

Lucas Sacramento Ramos[1] Henrique Mansur Gonçalves[2] Anderson Freitas[3] Marcio de Paiva Oliveira[3] Diogo Marcelino Santos Lima[4] Welvis Soares Carmargo[1]

[1] Orthopedics and Traumatology Department, Hospital Regional do Gama, Brasília, DF, Brazil
[2] Department of Medicine, Universidade de Brasília, Brasília, DF, Brazil
[3] Instituto de Pesquisa e Ensino do Hospital Ortopédico e Medicina Especializada, Brasília, DF, Brazil
[4] Department of Medicine, Centro Universitário de Brasília (CEUB), Brasília, DF, Brazil

Address for correspondence Lucas Sacramento Ramos, SQS 214, bloco C, apartamento 605, Asa Sul, Brasília, DF, 70293-030, Brazil (e-mail: lucas.sacramentoramos@gmail.com).

Rev Bras Ortop 2021;56(3):372–378.

Abstract

Objective The present study aims to analyze the intra- and interobserver reproducibility of the Lauge-Hansen, Danis-Weber, and Arbeitsgemeinschaft für Osteosynthesefragen (AO) classifications for ankle fractures, and the influence of evaluators training stage in these assessments.

Methods Anteroposterior (AP), lateral and true AP radiographs from 30 patients with ankle fractures were selected. All images were evaluated by 11 evaluators at different stages of professional training (5 residents and 6 orthopedic surgeons), at 2 different times. Intra- and interobserver agreement was analyzed using the weighted Kappa coefficient. Student t-tests for paired samples were applied to detect significant differences in the degree of interobserver agreement between instruments.

Results Intraobserver analysis alone had a significant agreement in all classifications. Moderate to excellent interobserver agreement was highly significant ($p \leq 0.0001$) for the Danis-Weber classification. The Danis-Weber classification showed, on average, a significantly higher degree of agreement than the remaining classification systems ($p \leq 0.0001$).

Conclusion The Danis-Weber classification presented the highest reproducibility among instruments and the evaluator's little experience had no negative influence on the reproducibility of ankle fracture classifications. *Level of Evidence II, Diagnostic Studies – Investigating a Diagnostic Test.*

Keywords
- ankle fractures
- classification
- reproducibility of results

* Study developed by the Orthopedics and Traumatology Service of the Hospital Regional do Gama, DF, Brazil, and by the Instituto de Pesquisa e Ensino do Hospital Ortopédico e Medicina Especializada (IPE-HOME-DF, in the Portuguese acronym) Brasília, DF, Brazil.

received
March 16, 2020
accepted
July 6, 2020
published online
December 18, 2020

DOI https://doi.org/10.1055/s-0040-1718508.
ISSN 0102-3616.

Thieme Revinter Publicações Ltda., Rua do Matoso 170, Rio de Janeiro, RJ, CEP 20270-135, Brazil

Resumo

Objetivo Avaliar a reprodutibilidade intra- e interobservador das classificações de Lauge-Hansen, Danis-Weber e Arbeitsgemeinschaft für Osteosynthesefragen (AO) para as fraturas de tornozelo, e a influência do estágio de formação dos participantes na avaliação.

Métodos Foram selecionadas radiografias de 30 pacientes com fratura de tornozelo nas incidências anteroposterior (AP), perfil e AP verdadeiro. Todas as imagens foram avaliadas por 11 participantes em diferentes estágios de formação profissional (cinco residentes e seis cirurgiões ortopédicos), em dois momentos distintos. Analisou-se a concordância inter- e intraobservador por meio do coeficiente Kappa ponderado. O teste t de Student para amostras pareadas foi aplicado para verificar se havia diferença significativa no grau de concordância interobservador entre os instrumentos.

Resultado Observou-se que existe concordância significativa em todas as classificações quando da análise intraobservador isolada. Existe concordância interobservador altamente significativa de grau moderado a ótimo na classificação de Danis-Weber ($p \leq 0{,}0001$). A classificação de Danis-Weber apresentou, em média, grau de concordância significativamente maior que as outras classificações ($p \leq 0{,}0001$).

Conclusão A classificação de Danis-Weber se mostrou a mais reprodutiva entre os instrumentos avaliados, e a pouca experiência do avaliador não influencia negativamente a reprodutibilidade das classificações das fraturas do tornozelo. *Nível de Evidência II, Estudos Diagnósticos - Investigação de um Exame para Diagnóstico.*

Palavras-chave

- fraturas do tornozelo
- classificação
- reprodutibilidade dos testes

Introduction

Ankle fractures comprise ~ 10% of all human body fractures; these injuries are more common in women and are associated with obesity and smoking.[1,2] The demographic transition resulted in an approximately 3-fold increase in the incidence of these fractures in elderly patients for the last 30 years.[1–3] At the ankle joint, the talus body fits into the malleolar clamp, functioning as a modified trochlea, and it is stabilized by lateral, medial and syndesmotic ligament complexes.[4] When subjected to deforming forces, mainly rotational, this complex bone-capsule-ligament anatomy suffers a number of injuries that must be studied. Ankle fractures diagnosis is based on clinical history, physical examination, and regional image evaluation, usually with simple ankle radiographs in anteroposterior (AP), lateral and true AP (with 20° internal rotation) views.[1,4]

Classification systems are important tools for prognosis definition and to guide the most appropriate treatment. A good classification system must have simple language and provide reliable information for correct propaedeutics.[5] In addition, it must be feasible, reliable, and reproducible. This latter feature depends on intra- and interobserver agreement[5,6] Reproducibility studies are classical in the literature to assess the quality of a classification system, especially in orthopedics, since they help to define which instrument provides greater agreement and understanding in the scientific community.[7]

The Lauge-Hansen classification for ankle fractures was the most used system for many years. It is based on trauma mechanism and it considers both foot positioning and the deforming force direction (i.e., pronation with abduction, pronation with external rotation, supination with adduction and supination with external rotation). The Danis-Weber classification is mostly anatomical and is based on the topography of the lateral malleolus and line type. Injuries are classified as infrasyndesmotic (A), transsyndesmotic (B) and suprasyndesmotic (C). The Arbeitsgemeinschaft für Osteosynthesefragen (AO) Group classification redefines the three types of Danis-Weber classification by taking into account medial injuries. Therefore, lesions are classified as infrasyndesmotic (isolated [A1], with medial malleolus injury [A2] or with postmedial fracture [A3]), transsyndesmotic (isolated [B1], with medial injury [B2] or with medial and posterolateral injuries [B3]) and suprasyndesmotic (simple fracture [C1], multifragmentary fracture [C2] or proximal fibular fracture [C3]).[1,8–10]

Although there are some studies in the literature evaluating the reproducibility of the various classification systems for ankle fractures, they are controversial and there is no consensus on which one is the most appropriate. In addition, little has been discussed about the relationship between the reproducibility of the instruments and the evaluator's experience.[11,12] Thus, the present study aims to analyze which of the three main classification systems for ankle fracture has the highest intra- and interobserver reproducibility, and whether the training stage of the evaluators influences the assessment. We believe that more complex classification systems present lower reproducibility and that more experienced evaluators will achieve greater agreement rates.

Material and Methods

Patients with ankle fractures in 2018 were selected after approval by the Research Ethics Committee with the opinion number 2.697.068/18. The study met all requirements regarding human rights.

Skeletally mature patients with a diagnosis of ankle fracture and AP, lateral, and true AP (with 20° internal rotation of the ankle) radiographic images were included randomly as they were seen in the hospital emergency room, up to a total of 30 subjects. Patients with no radiographs in the aforementioned views, with tests deemed low-quality by the researchers, or those who did not agree to participate in the study were excluded.

Radiographs were photographed and digitalized in the personal file of the main researcher. All images were inserted into Survey Monkey Canada Inc., Ottawa, Canada, which generated a virtual questionnaire for their classification by evaluators according to the Danis-Weber, Lauge-Hansen, and AO Group systems. The questionnaire also had illustrations of each classification system that the evaluators could consult at any time (►**Figures 1**, **2**, and **3**). The virtual questionnaire was sent to a total of 11 orthopedists in different stages of training, including 6 members of the Sociedade Brasileira de Ortopedia e Traumatologia (SBOT, in the Portuguese acronym), 2 specialists in foot and ankle surgery (from the Associação Brasileira de Medicina e Cirurgia do Tornozelo e Pé [ABTPé, in the Portuguese acronym), 4 non-specialists, and 5 resident physicians, 1 in the 1st year (R1), 2 in the 2nd year (R2) and 2 in the 3rd year (R3) of training to assess interobserver agreement. To assess intraobserver agreement, the same questionnaire was sent to these evaluators to repeat the process after one month.

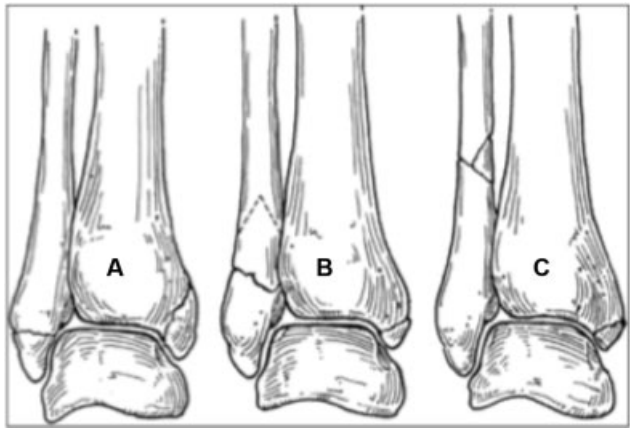

Fig. 1 Weber classification.

Statistical Analysis

The descriptive analysis presented data expressed as frequency (n) and percentage (%) in tables. The inferential analysis was composed by the weighted Kappa coefficient for intra- and interobserver agreement analysis of the Danis-Weber, Lauge-Hansen, and AO classification instruments in two time points. The Student t-tests for paired samples determined whether there was a significant difference in the degree of interobserver agreement between these instruments.

Intra- and interobserver reliability were assessed by the weighted Kappa coefficient, which determined whether there was a significant agreement, on an ordinal scale, for the Danis-Weber (3 levels), Lauge-Hansen (4 levels), and AO (9 levels) classification systems between the 2 time points (M1 and M2, i.e., 1 month after M1) in the sample of 30 radiographic studies. It is known that Kappa coefficients closer to 1 indicate stronger (or perfect) agreement between observers; in this case, observers are similar under the qualitative aspect of the assessment. On the other hand, Kappa coefficients closer to 0 indicate greater disagreement, i.e., there is no reproducibility and observed differences do not happen by chance.

The samples correspond to the Kappa coefficients of the 55 comparisons between evaluators, and there are 55 comparisons in the total sample; the subsample consisting of specialists alone has 15 comparisons, whereas the subsample consisting of residents alone has 10 comparisons, and the subsample of specialists versus residents presents 30 comparisons.

Significance was determined at a 5% level. The statistical analysis was performed using the statistical software SAS System, version 6.11 (SAS Institute, Inc., Cary, NC, USA).

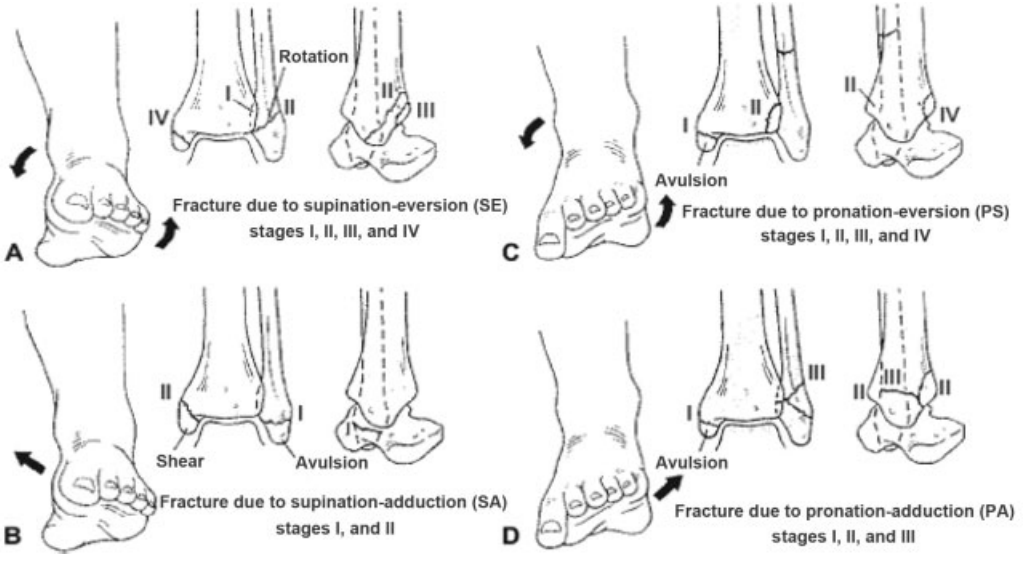

Fig. 2 Lauge-Hansen classification.

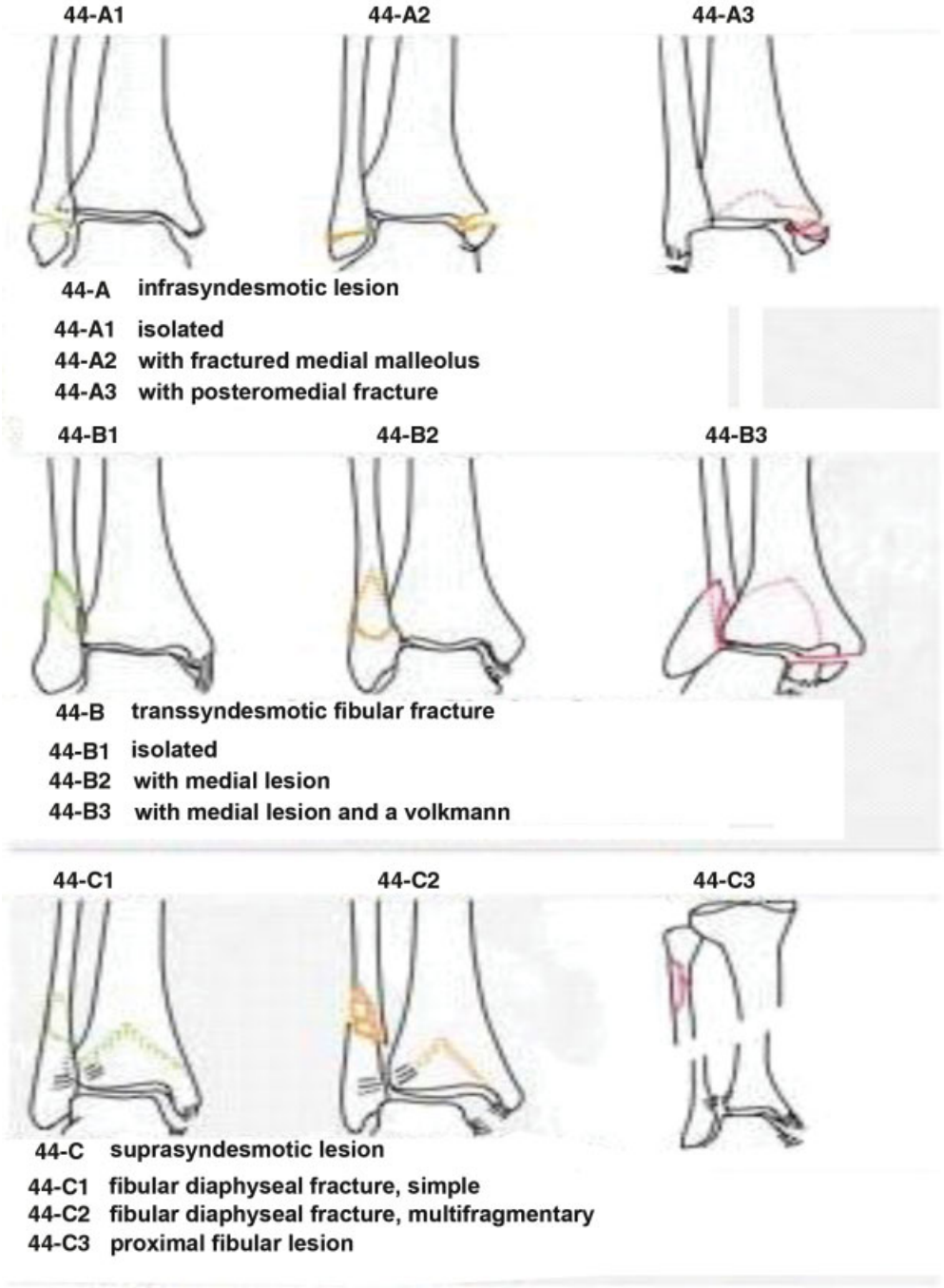

Fig. 3 AO classification for ankle fractures.[7]

Results

There was a significant agreement for intraobserver reproducibility for all 11 evaluators using the Danis-Weber classification ($p \leq 0.0001$); for 9 evaluators using the Lauge-Hansen classification, with 1 specialist ($p = 0.65$) and 1 resident ($p = 0.30$) showing no reproducibility between time points; and for 10 professionals using the AO classification, but with a specialist ($p = 0.071$) with no reproducibility between time points. In general, residents showed better intraobserver agreement than specialists.

There was a highly significant agreement ($p < 0.0001$), of moderate to excellent degree, in interobserver reproducibility for the Danis-Weber classification for all evaluators in both time points. Although there was a highly significant correlation ($p \leq 0.0001$) for most pairs in the Lauge-Hansen classification system, 7 pairs from the M1 time point and 3 pairs from the M2 time point showed no statistical significance. There was a highly significant agreement ($p \leq 0.0001$) for most pairs in the AO classification, but 7 pairs from the M1 time point showed no statistical relevance.

Table 1 Concordance degree among instruments

Sample*	Instrument	M1					M2				
		mean		SD	instrument pairs	*p* value[a]	mean		SD	instrument pairs	*p* value[a]
Total (n = 55)	W (1)	0.81	±	0.11	1 *vs* 2	< 0.0001	0.79	±	0.11	1 *vs* 2	< 0.0001
	LG (2)	0.43	±	0.21	1 *vs* 3	< 0.0001	0.52	±	0.18	1 *vs* 3	< 0.0001
	AO (3)	0.43	±	0.19	2 *vs* 3	0.83	0.48	±	0.21	2 *vs* 3	0.18
only among specialists (n = 15)	W (1)	0.76	±	0.10	1 *vs* 2	< 0.0001	0.79	±	0.13	1 *vs* 2	< 0.0001
	LG (2)	0.29	±	0.15	1 *vs* 3	< 0.0001	0.49	±	0.15	1 *vs* 3	< 0.0001
	AO (3)	0.30	±	0.17	2 *vs* 3	0.82	0.41	±	0.27	2 *vs* 3	0.15
only among residents (n = 10)	W (1)	0.86	±	0.08	1 *vs* 2	0.0008	0.76	±	0.08	1 *vs* 2	0.003
	LG (2)	0.59	±	0.19	1 *vs* 3	0.0001	0.54	±	0.18	1 *vs* 3	0.0005
	AO (3)	0.61	±	0.10	2 *vs* 3	0.85	0.54	±	0.12	2 *vs* 3	0.84
specialists *vs* residents (n = 30)	W (1)	0.82	±	0.11	1 *vs* 2	< 0.0001	0.80	±	0.11	1 *vs* 2	< 0.0001
	LG (2)	0.44	±	0.20	1 *vs* 3	< 0.0001	0.52	±	0.20	1 *vs* 3	< 0.0001
	AO (3)	0.44	±	0.18	2 *vs* 3	0.99	0.50	±	0.21	2 *vs* 3	0.61

Abbreviations: AO, Arbeitsgemeinschaft für Osteosynthesefragen; LG, classificação Lauge-Hansen; SD, standard deviation; W, classificação Danis-Weber.
*Kappa-weighted statistical sample.
[a]Student *t* test for paired samples.

In addition, correlations in the degree of interobserver agreement between instruments in the total sample and subsamples of evaluators were assessed. ►**Table 1** shows mean and standard deviation (SD) values for the degree of interobserver agreement (weighted Kappa coefficient) of the three instruments in the total sample and subsamples of evaluators from M1 and M2 time points. The Danis-Weber classification system presented, on average, a significantly higher degree of agreement than the Lauge-Hansen and AO systems in the total sample and subsamples both in M1 and M2 time points. There was no significant difference, at the 5% level, in the degree of agreement between the Lauge-Hansen and AO classification systems in the total sample and subsamples both in M1 and M2 time points.

Regarding the influence of the training stage of evaluators on reproducibility, it was observed that, in general, residents showed better intraobserver agreement, with values greater than the Kappa for the three classification systems and statistically significant differences for the two evaluated time points ($p < 0.05$).

Discussion

The main findings of the present study are partially consistent with our initial hypotheses. More complex classification systems for ankle fractures presented lower reproducibility. In contrast, however, more experienced evaluators agreed less in their responses at two different times.

In our study, 11 evaluators in different stages of training (residents, orthopedists and specialists in foot and ankle surgery) were asked to classify ankle fractures in 30 radiographic images, and their answers were statistically analyzed using the weighted Kappa method. Audigè et al.[13] carried out a systematic review on reproducibility studies of fracture classification systems and concluded that all of them relied on the Kappa method, but interpretation varied a lot due to confidence intervals (CIs). To avoid this bias, we used the CI defined by Landi et al.[14]

Fonseca et al.[11] evaluated the same classification systems for ankle fractures (namely, Danis-Weber, Lauge-Hansen, and AO), with 6 evaluators and 83 images; however, they considered AP and lateral radiographs alone. This study revealed a greater reproducibility for the Danis-Weber classification ($\kappa = 0.49$), with lower rates for the Lauge-Hansen ($\kappa = 0.32$) and AO ($\kappa = 0.38$) classification systems, which presented low agreement. Similar results were found by Alexandropoulos et al.,[12] who used three evaluators to classify 294 images of ankle fractures. They reported poor agreement for three classification systems ($\kappa = 0.327$–0.408, 0.174–0.476, and 0.397–0.483 for Broos-Bisschop, Lauge-Hansen, and AO, respectively).[12] Our study, in contrast, observed a highly significant degree of interobserver agreement for all classification systems, with values higher compared to previous studies ($\kappa = 0.79$, 0.52 and 0.48 for Danis-Weber, Lauge-Hansen, and AO, respectively). We believe that the high degree of agreement obtained in our study is related to the higher number of radiographic views compared with previous studies.[11,15,16] A most complete radiographic study certainly contributed to a more accurate diagnosis, facilitating lesion classification.

Few studies similar to ours evaluate intraobserver agreement.[11] Tenório et al.[15] reported that intraobserver agreement was moderate to high for the Lauge-Hansen classification ($\kappa = 0.58$) and moderate to almost perfect for the Danis-Weber classification ($\kappa = 0.76$). In our study, with 11 professionals, intraobserver agreement was significant ($p < 0.05$) among all evaluators for the Danis-Weber classification, for 9 evaluators using the Lauge-Hansen classification and for 10 evaluators

using the AO classification. We believe this happened because the questionnaire was large, resulting in less accuracy at the second evaluation. In addition, the greater complexity of the Lauge-Hansen and AO classification systems and the need to understand the fracture trauma mechanism for their correct use decrease their reproducibility.[11,15]

One of the goals of our work was to assess the influence of different stages of knowledge on practical activity. It is expected that as people study and become accustomed to a particular classification system, agreement between them and within their own observations would increase.[5] Fonseca et al.[11] reported that this variable did not influence the reproducibility rates of the studied classifications. However, since the authors only performed an interobserver agreement analysis, their understanding is partially limited. In our study, residents showed better intraobserver agreement, contrary to common sense. We believe that while residents resorted more often to the template illustrations provided for each classification system, most experienced evaluators classified fractures according to memory. This fact highlights the importance of knowing the instruments and their subtypes when using them to classify a fracture, since it often helps in the decision-making of surgical treatment.[17]

Based on our results, we conclude that a complete radiographic study, including AP, lateral and true AP views, is essential to classify ankle fractures, as well as the detailed knowledge of the instrument used and the occasional use of templates. Among the classification systems evaluated, although the Danis-Weber classification has proven to be the most reproducible, it provides insufficient information to guide fracture treatment, requiring an additional assessment of ankle joint stability for proper surgical indication. We believe that there is no ideal radiographic classification for malleolar fractures that presents high reproducibility and, at the same time, enables correct surgical planning. Thus, in more complex fractures, preoperative evaluation using computed tomography (CT) helps to understand the injury, especially trimalleolar lesions with posterior malleolus fragmentation.[18] Black et al. showed that CT plays an important role in fracture-dislocation, trimalleolar and suprasyndesmotic fractures, improving the preoperative study and surgical planning.[19]

We are aware of the limitations of our study. The main one is the number of radiographic images evaluated, which is lower compared to similar works. There are several articles in the literature evaluating the reproducibility of various classification systems for fractures in an attempt to define which one is the best.[10,20] However, there is still no consensus on the ideal methodology, since the number of analyzed images and evaluators influences the agreement on answers.[13,20] Numbers too small or too large decrease agreement.[14] Tenório et al.[15] used a total of 50 radiographs and 8 evaluators, whereas, in another study,[11] 6 evaluators classified 83 radiographs. We increased the number of evaluators to elevate the statistical power of interobserver agreement. In addition, the interval of 1 month between the 2 time points of questionnaire application differs from most previous studies including intraobserver analysis. This fact may have decreased the agreement, since the memory response is impaired. However, we were able to achieve the goal of the present study, which was the evaluation of responses from each evaluator alone.

Conclusion

The Danis-Weber classification was shown to be more reproducible compared with the Lauge-Hansen and AO systems, with a moderate to high degree of both intra- and interobserver agreement. The Lauge-Hansen and AO classification systems, on the other hand, presented similar low to moderate intra- and interobserver agreement.

In addition, residents showed a higher intraobserver agreement in all classifications, demonstrating that the little experience of the evaluator has no negative influence on the reproducibility of ankle fracture classifications.

Financial Support
There was no financial support from public, commercial, or non-profit sources.

Conflict of Interests
The authors have no conflict of interests to declare.

References

1 Marsh JL, Saltzman CL. Ankle Fractures. In: Bucholz RW, Heckman JD, Court-Brown CM, Charles A, editors. Rockwood & Green's Fractures in Adults. Philadelphia: Lippincott Williams & Wilkins; 2006:2147–2247
2 King CM, Hamilton GA, Cobb M, Carpenter D, Ford LA. Association between ankle fractures and obesity. J Foot Ankle Surg 2012;51 (05):543–547
3 Budny AM, Young BA. Analysis of radiographic classifications for rotational ankle fractures. Clin Podiatr Med Surg 2008;25(02): 139–152, v
4 Sinizio H. Xavier. Ortopedia e traumatologia: princípios e práticas. Porto Alegre: Artmed; 2009
5 Belloti JC, Tamaoki MJ, Franciozi CE, et al. Are distal radius fracture classifications reproducible? Intra and interobserver agreement. Sao Paulo Med J 2008;126(03):180–185
6 Alla SR, Deal ND, Dempsey IJ. Current concepts: mallet finger. Hand (N Y) 2014;9(02):138–144
7 Hahn DM, Colton CL. Malleolar fractures. In: Rüedi TP, Murphy WM, editors. AO Principles of fracture management. New York: Thieme Stuttgart; 2001:559–582
8 Tartaglione JP, Rosenbaum AJ, Abousayed M, DiPreta JA. Classifications in brief: Lauge-Hansen classification of ankle fractures. Clin Orthop Relat Res 2015;473(10):3323–3328
9 Okanobo H, Khurana B, Sheehan S, Duran-Mendicuti A, Arianjam A, Ledbetter S. Simplified diagnostic algorithm for Lauge-Hansen classification of ankle injuries. Radiographics 2012;32(02):E71–E84
10 Randsborg PH, Sivertsen EA. Classification of distal radius fractures in children: good inter- and intraobserver reliability, which improves with clinical experience. BMC Musculoskelet Disord 2012;13:6
11 Fonseca LLD, Nunes IG, Nogueira RR, Martins GEV, Mesencio AC, Kobata SI. Reproducibility of the Lauge-Hansen, Danis-Weber, and AO classifications for ankle fractures. Rev Bras Ortop 2017;53 (01):101–106
12 Alexandropoulos C, Tsourvakas S, Papachristos J, Tselios A, Soukouli P. Ankle fracture classification: an evaluation of three

classification systems : Lauge-Hansen, A.O. and Broos-Bisschop. Acta Orthop Belg 2010;76(04):521–525
13 Audigé L, Bhandari M, Kellam J. How reliable are reliability studies of fracture classifications? A systematic review of their methodologies. Acta Orthop Scand 2004;75(02):184–194
14 Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(01):159–174
15 Tenório RB, Mattos CA, Araújo LH, Belangero WD. Análise da reprodutibilidade das classificações de Lauge-Hansen e Danis-Weber para fraturas de tornozelo. Rev Bras Ortop 2001;36(11/12):434–437
16 Martin JS, Marsh JL. Current classification of fractures. Rationale and utility. Radiol Clin North Am 1997;35(03):491–506
17 Brage ME, Rockett M, Vraney R, Anderson R, Toledano A. Ankle fracture classification: a comparison of reliability of three X-ray views versus two. Foot Ankle Int 1998;19(08):555–562
18 Kumar A, Mishra P, Tandon A, Arora R, Chadha M. Effect of CT on Management Plan in Malleolar Ankle Fractures. Foot Ankle Int 2018;39(01):59–66
19 Black EM, Antoci V, Lee JT, et al. Role of preoperative computed tomography scans in operative planning for malleolar ankle fractures. Foot Ankle Int 2013;34(05):697–704
20 Berger AJ, Momeni A, Ladd AL. Intra- and interobserver reliability of the Eaton classification for trapeziometacarpal arthritis: a systematic review. Clin Orthop Relat Res 2014;472(04):1155–1159