Letter to the Editor

Reliability of Ankle Fracture Classification by Junior Residents and Medical Students in Simulated Clinical Settings

Dear Editor,

Ankle fractures are common[1,2] and account for 9% of all fractures[3] with an annual incidence of 174–248 cases per 100,000 adults,[4,5] and their prognostication and management are guided by imaging-based classification systems for ankle fractures.[6] The AO Foundation/Orthopaedic Trauma Association's (AO/OTA) revised classification[7] and Weber classification[8] are commonly used to classify ankle malleolar fractures (Fig. 1).

The AO/OTA classification is organised around the location and associated characteristics of the fracture. On the other hand, the Weber classification considers the location of the fibular fracture relative to the syndesmosis.[9] Infra-syndesmotic fractures (Weber A) are stable and managed conservatively, while unstable trans-syndesmotic (Weber B) and supra-syndesmotic (Weber C) fractures are managed surgically.[10] Clinical outcomes differ, with good to excellent outcomes reported in 82.7% and 83.8% of Weber A and B fractures, respectively, compared to 70.4% of Weber C fractures.[5] Operative outcomes also vary, with good to excellent results seen in 95.2%, 94.6% and 80.6% of Weber A, B and C fractures, respectively.[11] Various studies have evaluated the inter-observer reliability and intra-observer reproducibility of both classifications, with kappa values ranging from 0.42–0.86 and 0.34–0.93 for AO/OTA and Weber classifications, respectively,[9,12,13] that corresponded with moderate to substantial agreement.[14]

In the busy emergency department and clinic session and examination, clinicians and students have only seconds to examine a radiograph before they advise on the likely management—conservative or surgical—of an ankle fracture, and whether 1 or both tibia and fibula require fixation. In the literature, studies that compared

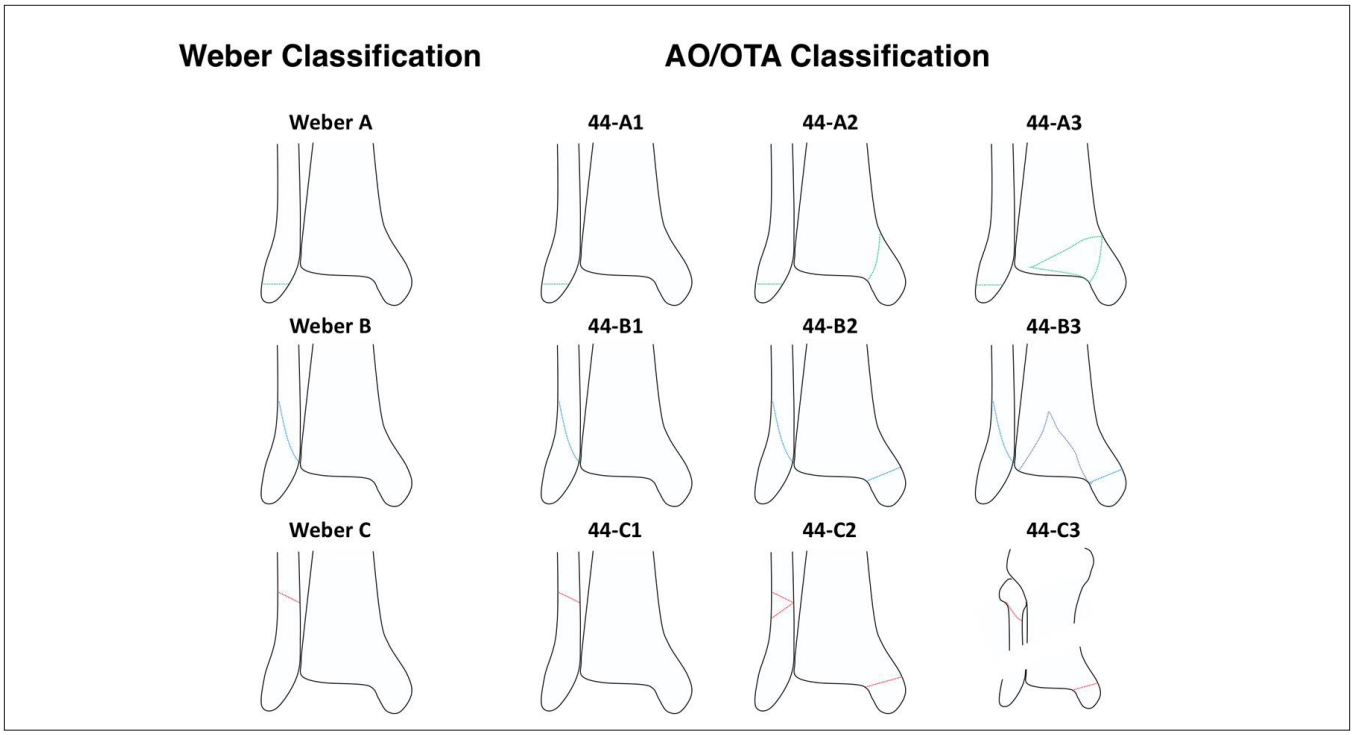

Fig. 1. AO Foundation/Orthopaedic Trauma Association (AO/OTA) and Weber ankle fracture classification systems.

ankle fracture classifications were performed without any consideration of the time taken to do so. However, under the duress of time, the reliability of these classifications may be questioned. Consequently, this study investigated the inter- and intra-observer reproducibility of the Weber and AO/OTA classifications and the differences between inter- and intra-observer reliability, if any, of a clinic session and a medical student examination under time conditions that simulated clinical consultation.

Materials and Methods

A retrospective review of 81 patients who underwent ankle malleolar fracture surgery between 1 January 2015–31 December 2016 was performed. For each patient, a standardised anteroposterior view of the earliest postinjury and presurgery ankle radiograph was extracted. The 3 observers—2 medical students and 1 orthopaedic resident—were blinded to the personal, clinical and operative details of patients.

Two months after radiograph extraction, 80 radiographs were classified according to the AO/OTA and Weber classifications 1 week apart on 4 occasions under timed and untimed conditions (Fig. 2). The 1-week interval was chosen to minimise recall bias. For the AO/OTA classification, radiographs were classified according to group levels in segment 44 (such as 44B2 and 44C1). A software was created in C# in Microsoft Visual Studio with an embedded Microsoft Excel to randomise the radiographs and record the time taken by each observer to classify every radiograph.

To control for chance agreement, inter- and intra-observer reliability was calculated based on Cohen's kappa (for 2 observers) and Fleiss' kappa (for >2 observers) using IBM SPSS Statistics for Windows, Version 23.0 (IBM Corp., Armonk, NY, USA). A kappa value of 1 signified perfect agreement, 0 signified chance agreement and a negative value signified agreement that was worse than chance. Kappa

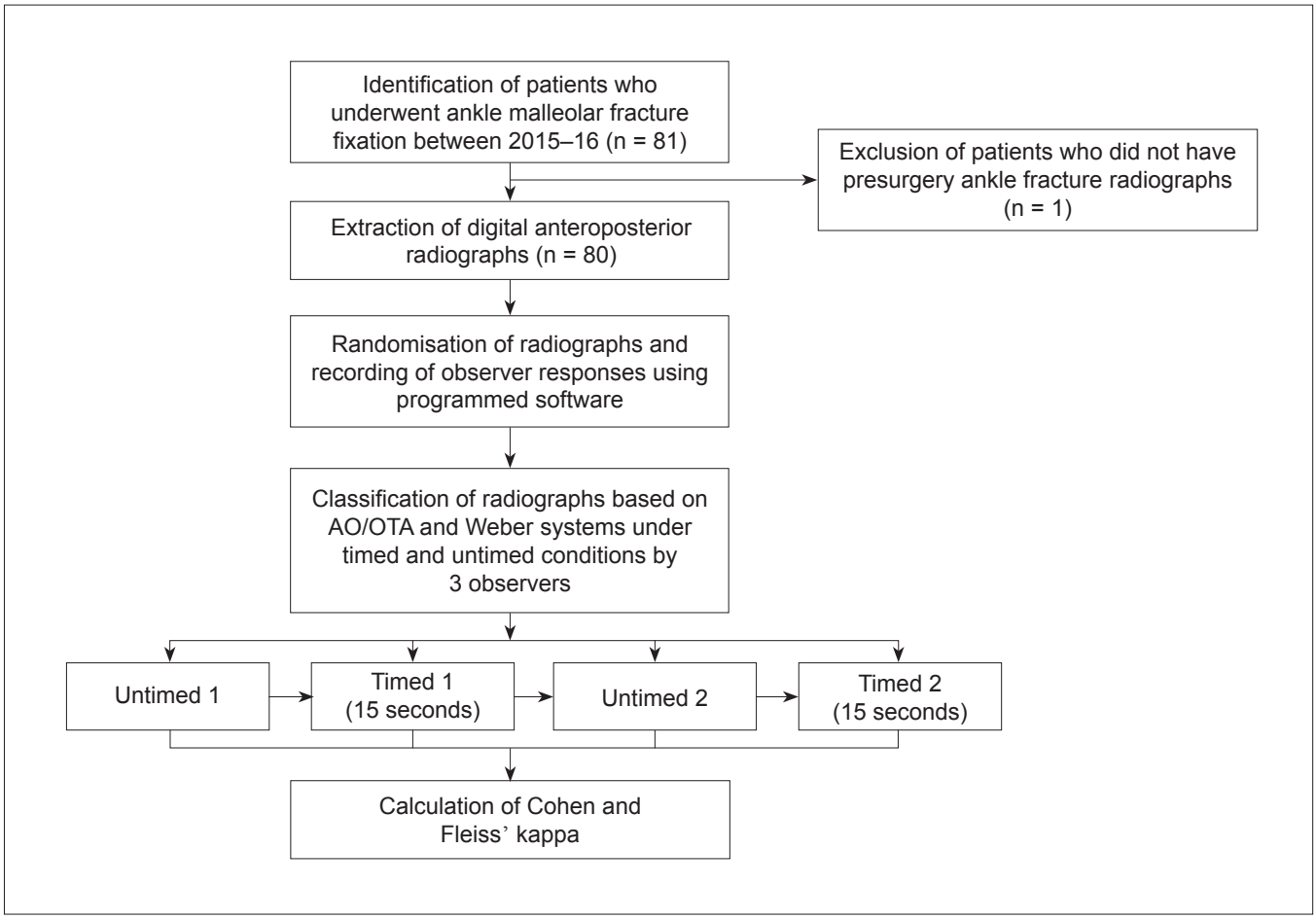

Fig. 2. Flow chart of patient selection process. AO/OTA: AO Foundation/Orthopaedic Trauma Association

values were analysed against the benchmark scale of Landis and Koch: slight (0.00–0.20), fair (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80) and excellent or almost perfect (0.81–1.00).[14]

This study was approved by the Domain Specific Review Board (reference number 2017/01210).

Results

A total of 80 patients were included in the study after 1 patient was excluded when the presurgery radiographs could not be retrieved (Fig. 2). For classifications that were made under timed conditions, the mean time ranged between 9.33–22.2 seconds and 8.87–12.5 seconds for first and second examinations, respectively (Table 1); for classifications that were performed under untimed conditions, it ranged between 7.31–11.4 seconds and 5.88–9.74 seconds for first and second examinations, respectively.

The percentage of agreement in the Weber system ranged between 60.5–82.7% and 69.1–86.4% for timed and untimed classifications, respectively; in the AO/OTA system, it ranged between 43.1–56.8% and 51.9–66.7% for timed and untimed classifications, respectively (Table 2).

Mean inter-observer Cohen's kappa in the Weber system ranged between 0.375–0.747 (fair to substantial agreement) and 0.544–0.802 (moderate to almost perfect agreement[14]) for timed and untimed classifications, respectively; in the AO/OTA system, the values ranged between 0.256–0.416 (fair to moderate agreement) and 0.377–0.563 (fair to moderate agreement) for timed and untimed classifications, respectively.

Fleiss' kappa in the Weber system were 0.519 and 0.551 (moderate agreement) for timed and untimed classifications, respectively; in the AO/OTA system, the values were 0.418 and 0.453 (moderate agreement) for timed and untimed classifications, respectively.

Intra-observer agreement in the Weber system ranged between 72.8–96.3% (kappa 0.689–0.940; substantial to almost perfect agreement) and 76.5–84.0% (kappa 0.608–0.715; substantial agreement) for timed and untimed classifications, respectively (Table 2). In contrast, it ranged between 69.1–76.5% (kappa 0.579–0.661; moderate to substantial agreement) and 63.0–72.8% (kappa 0.463–0.622; moderate to substantial agreement) for timed and untimed classifications in the AO/OTA system, respectively.

Discussion

To the best of our knowledge, no study which replicated the time-sensitive nature of clinical practice has been published in the literature that recommends a minimum duration needed to make an accurate and reliable classification of ankle fractures based on the AO/OTA and Weber systems. An understanding of the variable reliability of both systems under timed and untimed conditions could guide the choice of system that is used in clinical practice and time needed to make a reliable classification.

Previous studies that evaluated the AO/OTA and Weber classification systems had yielded results that ranged from fair to almost perfect agreement. For inter-observer reliability, Juto et al demonstrated substantial to almost perfect agreement (kappa 0.67–0.88) for Weber classification and moderate to substantial agreement (kappa 0.56–0.76) for AO/OTA classification.[12] Malek et al, on the other hand, showed moderate to substantial agreement (kappa

Table 1. Duration to Classify Ankle Fracture Radiographs

Condition	Observer 1		Observer 2		Observer 3	
	Mean Time (in Seconds)	95% CI	Mean Time (in Seconds)	95% CI	Mean Time (in Seconds)	95% CI
Untimed						
1	14.8	4.76 – 34.1	22.2	6.15 – 70.4	9.34	3.49 – 24.8
2	10.7	2.18 – 34.9	8.87	4.25 – 16.4	12.5	4.34 – 34.9
Timed						
1	7.82	3.23 – 15.0	11.4	5.01 – 15.0	7.31	2.59 – 15.0
2	6.94	2.23 – 15.0	9.74	4.60 – 15.0	5.88	2.55 – 13.1

CI: Confidence interval

0.59–0.63) for Weber classification.[13] In terms of intra-observer reproducibility, Juto et al[12] found almost perfect agreement (kappa 0.80–0.93) for Weber classification and substantial to almost perfect agreement (kappa 0.74–0.86) for AO/OTA classification. Malek et al, however, found fair to almost perfect agreement (kappa 0.39–0.86) for Weber classification.[13]

In our study, the same agreement was observed. For inter-observer reliability, there was moderate to almost perfect agreement (kappa 0.544–0.802) and fair to moderate agreement (kappa 0.377–0.563) for Weber and AO/OTA classifications, respectively. For intra-observer reproducibility, there was substantial agreement (kappa 0.608–0.715) and moderate to substantial agreement (kappa 0.463–0.622) for Weber and AO/OTA classifications, respectively.

Table 2. Inter- and Intra-Observer Reliability for AO/OTA and Weber Ankle Fracture Classification Systems Under Timed and Untimed Conditions

Variable	Weber			AO/OTA		
	PA	Cohen's Kappa	95% CI	PA	Cohen's Kappa	95% CI
Inter-Observer						
1 and 2						
Untimed 1	86.4	0.802	0.663 – 0.941	66.7	0.563	0.430 – 0.696
Untimed 2	79.0	0.642	0.456 – 0.828	53.1	0.381	0.244 – 0.518
Timed 1	82.7	0.747	0.588 – 0.906	54.3	0.372	0.252 – 0.492
Timed 2	81.5	0.728	0.556 – 0.900	56.8	0.416	0.283 – 0.549
2 and 3						
Untimed 1	71.6	0.544	0.375 – 0.713	59.4	0.471	0.338 – 0.604
Untimed 2	69.1	0.557	0.385 – 0.730	51.9	0.377	0.248 – 0.506
Timed 1	66.7	0.554	0.403 – 0.705	51.9	0.376	0.253 – 0.499
Timed 2	64.2	0.457	0.308 – 0.606	43.1	0.403	0.278 – 0.528
1 and 3						
Untimed 1	75.3	0.653	0.496 – 0.810	61.7	0.502	0.371 – 0.633
Untimed 2	72.8	0.612	0.440 – 0.784	61.7	0.471	0.346 – 0.596
Timed 1	65.4	0.523	0.356 – 0.690	45.7	0.256	0.144 – 0.367
Timed 2	60.5	0.375	0.218 – 0.532	53.1	0.358	0.236 – 0.480
Intra-Observer						
1						
Untimed	84.0	0.698	0.531 – 0.865	72.8	0.622	0.493 – 0.751
Timed	84.0	0.752	0.580 – 0.924	75.3	0.580	0.447 – 0.713
2						
Untimed	84.0	0.715	0.554 – 0.876	70.4	0.521	0.380 – 0.662
Timed	96.3	0.940	0.858 – 1.00	76.5	0.661	0.534 – 0.788
3						
Untimed	76.5	0.608	0.443 – 0.773	63.0	0.463	0.328 – 0.598
Timed	72.8	0.689	0.544 – 0.834	69.1	0.579	0.457 – 0.701

AO/OTA: AO Foundation/Orthopaedic Trauma Association; CI: Confidence interval; PA: Percentage of agreement

For both timed and untimed classifications, Cohen's kappa between 2 observers showed that the Weber system had a mean of 1 Landis and Koch grade higher than that for the AO/OTA system. Based on Landis and Koch's scale—where a difference of 0.2 between kappa values constitutes a significant difference—there was, however, no significant difference in Cohen's kappa for all observer pairs except for 1 pair (observer 1 vs 2) and Fleiss' kappa. With the exception of 1 classification, no significant difference was observed in all timed and untimed classifications based on the Landis and Koch's scale. Consequently, using the AO/OTA or Weber system, a minimum of 15 seconds are needed to make a reliable classification.

A limitation of this study was that the patient cohort was not representative of the ankle malleolar fracture population. Since they comprised patients who had operative management of ankle fractures, patients who were conservatively managed for Weber A fractures were excluded.[10] However, our study evaluated more complicated Weber B and C fractures that were typically more difficult to classify and had a greater impact on patients' outcomes. Another limitation was recall bias. To minimise it, a 2-month wait was observed after radiograph extraction before classifications commenced, and these were performed 1 week apart. Additionally, none of the observers were orthopaedic surgeons, but this was in line with one of the aims of this study which was to compare classifications undertaken by emergency department staff and medical students who lacked extensive experience in sub-classification and surgical fracture management. Since our findings on percentage agreement and kappa values corroborated those of studies that involved experienced surgeons, the metrics that were derived to describe reliability appeared to be independent of the level of expertise. Although lateral radiographs were not used in our study, the findings on percentage agreement and kappa values were similar to those of studies that had used radiographs with anteroposterior and lateral views.[9,12,13]

Conclusion

Our study showed that inter- and intra-observer reliability of the AO/OTA and Weber classification systems are similar under timed and untimed conditions. Consequently, the utility of either system is similar when they are used by either emergency department staff or medical students. The duration of 15 seconds are also needed to make a reliable classification using either system.

REFERENCES

1. Bauer M, Bengnér U, Johnell O, Redlund-Johnell I. Supination-eversion fractures of the ankle joint: changes in incidence over 30 years. Foot Ankle 1987;8:26–8.
2. Praemer A, Furner S, Rice DP. Musculoskeletal Conditions in the United States. 2nd ed. Park Ridge, Illinois: American Academy of Orthopaedic Surgeons;1999.
3. Court-Brown CM, Caesar B. Epidemiology of adult fractures: a review. Injury 2006;37:691–7.
4. Kannus P, Palvanen M, Niemi S, Parkkari J, Järvinen M. Increasing number and incidence of low-trauma ankle fractures in elderly people: Finnish statistics during 1970–2000 and projections for the future. Bone 2002;31:430–3.
5. Karam E, Shivji FS, Bhattacharya A, Bryson DJ, Forward DP, Scammell BE, et al. A cross-sectional study of the impact of physiotherapy and self directed exercise on the functional outcome of internally fixed isolated unimalleolar Weber B ankle fractures. Injury 2017;48:531–5.
6. Goost H, Wimmer MD, Barg A, Kabir K, Valderrabano V, Burger C. Fractures of the ankle joint: investigation and treatment options. Dtsch Arztebl Int 2014;111:377–88.
7. AO Foundation. Malleoli. Available at: https://www2.aofoundation.org/wps/portal/surgery?showPage=diagnosis&bone=Tibia&segment=Malleoli. Accessed on 31 March 2020.
8. Weber BG. The Injuries of the Upper Ankle. 2nd ed. Berne: Verlag Hans Huber; 1972.
9. Verhage SM, Rhemrev SJ, Keizer SB, Quarles van Ufford HME, Hoogendoorn JM. Interobserver variation in classification of malleolar fractures. Skeletal Radiol 2015;44:1435–9.
10. Bellringer SF, Brogan K, Cassidy L, Gibbs J. Standardised virtual fracture clinic management of radiographically stable Weber B ankle fractures is safe, cost effective and reproducible. Injury 2017;48:1670–3.
11. Low CK, Pang HY, Wong HP, Low YP. A retrospective evaluation of operative treatment of ankle fractures. Ann Acad Med Singapore 1997;26:172–4.
12. Juto H, Möller M, Wennergren D, Edin K, Apelqvist I, Morberg P. Substantial accuracy of fracture classification in the Swedish Fracture Register: evaluation of AO/OTA-classification in 152 ankle fractures. Injury 2016;47:2579–83.
13. Malek IA, Machani B, Mevcha AM, Hyder NH. Inter-observer reliability and intra-observer reproducibility of the Weber classification of ankle fractures. J Bone Joint Surg Br 2006;88:1204–6.
14. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.

Glen ZQ Liau, [1,2]*MBBS, MMed (Ortho), MBA*, Sean KA Phua, [1]*MBBS*, Tian Pei Li, [1]*MBBS*, Yu Han Chee, [1,2]*MBChB, MRCS, FRCS*

[1] Department of Orthopaedic Surgery, National University Hospital, Singapore
[2] University Orthopaedics, Hand and Reconstructive Microsurgery Cluster, National University Health System, Singapore

Address for Correspondence: Dr Glen Liau Zi Qiang, Department of Orthopaedic Surgery, National University Hospital, 5 Lower Kent Ridge Road, Singapore 119074.
Email: glen_liau@nuhs.edu.sg