Assessing the Interobserver Reliability of the 2018 AO/OTA Fracture Classification System for Long Bone Fractures

Timothy S. Keeley[1], Daniel Theng[1], Tim Wu[1], Daniel Rohde[2], Meir Tibi Marmor[2]
[1]University of California, San Francisco, School of Medicine, San Francisco, CA, [2]University of California San Francisco, Department of Orthopaedic Surgery, San Francisco, CA,
Timothy.keeley@ucsf.edu

Disclosures: N/A

INTRODUCTION: Fracture classification systems provide a standardized framework for describing injury patterns, facilitating effective communication, consistent treatment planning, and enabling comparative outcomes research. Long bone fractures represent a significant portion of the global orthopaedic trauma burden and require reliable classification systems for treatment optimization. The Arbeitsgemeinschaft für Osteosynthesefragen (AO) and the Orthopaedic Trauma Association (OTA) developed a widely accepted alphanumeric system to enhance fracture documentation and communication. This study evaluated the interobserver reliability of the 2018 AO/OTA classification of long bone fractures and agreement across fracture subsets. We hypothesized that agreement would be almost perfect at early classification levels, including bone and location, and progressively decrease at the level of type, with comparable agreement across fracture subsets stratified by characteristics (i.e., bone, location, and type).

METHODS: Radiographic images of acute fractures were extracted from both our institution's de-identified database and the electronic medical records of patients. Long bone fractures, along with patellar and malleolar fractures, sustained by adult patients (age ≥ 18 years) were included. All images were reviewed by an orthopaedic trauma surgeon for quality, with images deemed suboptimal being excluded or supplemented with computed tomography images. Nine fellowship-trained orthopaedic trauma surgeons and members of the OTA Classification & Outcomes Committee assigned classifications using the 2018 AO/OTA Fracture and Dislocation Compendium. Each fracture was assigned an independent classification by two reviewers. Reviewers subsequently developed a consensus classification for each study, which was used as the reference for stratified analysis across fracture subsets. Cohen's kappa analysis and accuracy were calculated for fractures that received classifications by two reviewers at least up to the level of type. These analyses were conducted across individual (i.e., bone only and location only) and cumulative (i.e., combined bone, location, and type) 2018 AO/OTA components. Subset analysis assessed agreement across similar characteristics of bone, location, and type. Bones representing at least 20 fractures of the batch were included in the respective subgroup analysis. Kappa coefficients were interpreted using the Landis and Koch criteria.

RESULTS SECTION: A total of 147 fractures met the inclusion criteria, with fractures of the humerus (n = 26), radius (n = 31), and femur (n = 39) included for subset analysis of the bones. Interobserver agreement (kappa, accuracy) was almost perfect for bone (κ = 1.00, 100%) and location (κ = 0.990, 99.3%) across the entire fracture batch. Agreement decreased to moderate at the level of type (κ = 0.627, 76.9%) and progressively thereafter for group (κ = 0.370, 56.8%), subgroup (κ = 0.342, 53.7%), qualifications (κ = 0.278, 36.0%), and universal modifiers (κ = -0.226, 3.1%). A similar agreement was observed for bone, location, and type in the cumulative kappa analysis; however, combined agreement was increased at the group (κ = 0.536, 54.4%) and subgroup (κ = 0.416, 42.2%) levels compared to the analysis of individual elements. When stratifying the analysis by subsets of bone, location, and type, a similar pattern of agreement was observed. However, at the level of type, radius fractures (κ = 0.750, 83.9%) were classified with higher agreement compared to femur (κ = 0.589, 82.1%) and humerus (κ = 0.615, 77.8%). In contrast, agreement was higher at the level of group for femur (κ = 0.395, 59.0%) and humerus (κ = 0.337, 55.6%) compared to radius (κ = 0.221, 51.6%). Across subsets of all fracture locations (proximal, diaphyseal, and distal), accuracy was similar for bone and location. However, agreement was substantially lower for diaphyseal fractures at the level of type (κ = 0.262, 60.5%). Agreement across Type A, B, and C fractures was comparable until the level of group, where Type A (κ = 0.506, 66.1%) fractures were classified at higher agreement than B (κ = 0.356, 57.1%) and C (κ = 0.124, 38.7%).

DISCUSSION: The interobserver agreement for classifying long bones using the 2018 AO/OTA classification system was almost perfect at the levels of bone and location, but decreased at the level of type and afterward. Similar trends of agreement were observed across subsets of fractures that shared characteristics of bone, location, and type. However, inter-rater agreement varied across fracture subsets at deeper levels of classification. The 2028 modification of the AO/OTA classification should aim to improve reliability at deeper levels of classification to enhance its utility in clinical and research applications.

SIGNIFICANCE/CLINICAL RELEVANCE: Fracture classification systems facilitate effective communication of injury patterns and consistency in fracture treatment. This study highlights the need to improve reliability at deeper AO/OTA classification levels in anticipation of the 2028 AO/OTA modification.

Table 1. Batch Interobserver Agreement and Subset Analysis Assessed through Kappa Analysis & Accuracy*

Category	Full Batch (n=147)	Humerus (n=26)	Radius (n = 31)	Femur (n= 39)	Proximal (n = 47)	Diaphyseal (n = 43)	Distal (n =46)	Type A (n = 74)	Type B (n = 40)	Type C (n = 32)
Bone	1.000 (100%)	1.000 (100%)	1.000 (100%)	1.000 (100%)	1.000 (100%)	1.000 (100%)	1.000 (100%)	1.000 (100%)	1.000 (100%)	1.000 (100%)
Location	0.990 (99.3%)	1.000 (100%)	1.000 (100%)	1.000 (100%)	1.000 (100%)	1.000 (100%)	0.000 (97.8%)	0.980 (98.7%)	1.000 (100%)	1.000 (100%)
Type	0.627 (76.9%)	0.615 (77.8%)	0.750 (83.9%)	0.589 (82.1%)	0.713 (83.3%)	0.262 (60.5%)	0.768 (84.8%)	-0.055 (88.0%)	-0.024 (65.0%)	-0.181 (65.6%)
Group	0.370 (56.8%)	0.337 (55.6%)	0.221 (51.6%)	0.395 (59.0%)	0.499 (67.4%)	0.103 (33.3%)	0.344 (59.5%)	0.506 (66.1%)	0.356 (57.1%)	0.124 (38.7%)
Subgroup	0.342 (53.7%)	0.284 (53.8%)	0.229 (52.2%)	0.321 (53.6%)	0.272 (47.2%)	None	0.412 (61.1%)	0.528 (67.6%)	0.120 (39.1%)	0.281 (48.0%)
Qualifications	0.278 (36.0%)	0.260 (38.1%)	0.011 (20.0%)	0.380 (63.6%)	0.115 (18.8%)	0.434 (58.8%)	0.002 (13.6%)	0.454 (60.0%)	0.209 (28.6%)	0.000 (12.5%)
Universal Modifiers	-0.226 (3.1%)	-0.240 (0.0%)	-0.071 (11.8%)	-0.079 (4.3%)	-0.144 (3.4%)	-0.308 (0.0%)	-0.246 (6.7%)	-0.242 (2.0%)	-0.287 (3.3%)	-0.107 (5.3%)
Combined: Bone, Location	0.993 (99.3%)	1.000 (100%)	1.000 (100%)	1.000 (100%)	1.000 (100%)	1.000 (100%)	0.968 (97.8%)	0.985 (98.7%)	1.000 (100%)	1.000 (100%)
Combined: Bone, Location, Type	0.752 (76.2%)	0.726 (77.8%)	0.783 (83.9%)	0.754 (82.1%)	0.797 (83.3%)	0.562 (60.5%)	0.799 (82.6%)	0.853 (86.7%)	0.629 (65.0%)	0.577 (65.6%)
Combined: Bone, Location, Type, Group	0.536 (54.4%)	0.484 (51.9%)	0.457 (51.6%)	0.515 (56.4%)	0.617 (64.6%)	0.299 (32.6%)	0.582 (60.9%)	0.653 (66.7%)	0.429 (45.0%)	0.332 (37.5%)
Combined: Bone, Location, Type, Group, Subgroup	0.416 (42.2%)	0.371 (40.7%)	0.319 (35.5%)	0.385 (41.0%)	0.419 (43.8%)	0.299 (32.6%)	0.459 (47.8%)	0.563 (57.3%)	0.286 (30.0%)	0.194 (21.9%)
Combined: Bone, Location, Type, Group, Subgroup, Qualifications	0.301 (30.6%)	0.201 (22.2%)	0.233 (25.8%)	0.333 (35.9%)	0.316 (33.3%)	0.213 (23.3%)	0.312 (32.6%)	0.443 (45.3%)	0.190 (20.0%)	0.082 (9.4%)
Combined: Bone, Location, Type, Group, Subgroup, Qualifications, Universal Modifiers	0.099 (10.2%)	0.072 (7.4%)	0.114 (12.9%)	0.137 (15.4%)	0.135 (14.6%)	0.062 (7.0%)	0.101 (10.9%)	0.154 (16.0%)	0.046 (5.0%)	0.026 (3.1%)

*Landis & Koch interpretation: <0 = Poor; 0.00–0.20 = Slight; 0.21–0.40 = Fair; 0.41–0.60 = Moderate; 0.61–0.80 = Substantial; 0.81–1.00 = Almost perfect.