Original Article Clinics in Orthopedic Surgery 2024;16:688-693 • https://doi.org/10.4055/cios23292

Reliability of the 2018 Revised Version of AO/OTA Classification for Femoral Shaft Fractures

Jung-Wee Park, MD*,†, Woo-Lam Jo, MD‡, Byung Kyu Park, MD*, Jong Jin Go, MD*, Minji Han, MPH§, Sungha Chun, MD*, Young-Kyun Lee, MD*,†

*Department of Orthopedic Surgery, Seoul National University Bundang Hospital, Seongnam,
†Department of Orthopedic Surgery, Seoul National University College of Medicine, Seoul,
‡Department of Orthopaedic Surgery, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul,
§Department of Health Science and Technology, Graduate School of Convergence Science and Technology, Seoul National University, Seoul, Korea

Background: The Arbeitsgemeinschaft für Osteosynthesefragen (AO) and the Orthopaedic Trauma Association (OTA) classification system for diaphyseal fracture has been recently revised to refine and enhance the accuracy of fracture categorization. This study aimed to investigate the interobserver reliability of the new AO/OTA classification and to compare it with the older version in femoral shaft fractures.

Methods: We retrospectively analyzed 139 patients (mean age, 43.8 ± 19.5 years; 92 men and 47 women) with femoral shaft fractures who were treated from 2003 to 2017. Four well-trained observers independently classified each fracture following the previous and revised AO/OTA classification system. We calculated the Fleiss kappa for the interobserver reliability.

Results: The previous classification showed the kappa value of 0.580 (95% confidence interval [CI], 0.547–0.613), and the revised version showed 0.528 (95% CI, 0.504–0.552). Both the old and the revised versions showed moderate reliability.

Conclusions: Our study highlights the moderate interobserver reliability of both the previous and new AO/OTA classification systems for diaphyseal femur fractures. These findings emphasize the importance of standardized systems in clinical decision-making and underscore the need for ongoing education and collaboration to enhance fracture classification.

Keywords: *Femoral fracture, Classification, Validation study, Interobserver variability*

Diaphyseal fractures of the femur represent a significant clinical challenge, demanding accurate classification systems to guide treatment decisions and ensure consistent communication among healthcare professionals.[1-4]

Fracture classification is a fundamental component of orthopedic practice, allowing for the categorization of injuries based on specific characteristics such as fracture location, morphology, and associated soft-tissue injuries. The Arbeitsgemeinschaft für Osteosynthesefragen (AO) and the Orthopaedic Trauma Association (OTA) have historically been at the forefront of providing standardized classifications for fractures, distinguishing specific characteristics of each fracture pattern and facilitating communication between surgeons since 1986.[5,6] Recently, an updated version of the AO/OTA classification system for diaphyseal fractures of the femur has been introduced, aiming to refine the existing framework and address potential limitations.[7] The introduction of the new AO/OTA classification system for diaphyseal femur fractures has prompted the need for a rigorous assessment of its interobserver reliability, a key determinant of its real-world applicability.

Interobserver reliability pertains to the consistency of classification when performed by different observers,

Received September 22, 2023; Revised January 22, 2024;
Accepted February 14, 2024
Correspondence to: Young-Kyun Lee, MD
Department of Orthopedic Surgery, Seoul National University Bundang Hospital, 82 Gumi-ro 173beon-gil, Bundang-gu, Seongnam 13620, Korea
Tel: +82-31-787-7204, Fax: +82-31-787-4056
E-mail: ykleemd@gmail.com
Jung-Wee Park and Woo-Lam Jo contributed equally to this work as co-first authors.

such as orthopedic surgeons with varying levels of experience. This reliability is vital to maintain the effectiveness of the classification system across diverse clinical settings, contributing to the accurate communication of fracture patterns and the appropriate selection of treatment modalities.[5] By evaluating the interobserver reliability of the updated AO/OTA classification system against the older version, we can ascertain whether the revisions have successfully improved agreement among orthopedic surgeons in classifying diaphyseal femur fractures. However, the revised version of the AO/OTA classification system has been rarely analyzed in terms of reproducibility and interobserver concordance.[5]

The purpose of the study was to evaluate the interobserver reliability of the new AO/OTA classification compared with the older version for diaphyseal fractures of the femur.

METHODS

This retrospective study was approved by the Institutional Review Board of Seoul National University Bundang Hospital (IRB No. B-2311-864-103), which waived informed consent.

To evaluate the reliability of AO/OTA classification for long bone fractures, we identified the patients, who were treated for diaphyseal fractures of femur at our hospital. The study inclusion criteria were as follows: all patients who were treated for diaphyseal fractures of the femur at our institute from 2003 to 2017, the availability of 2 views (anteroposterior and lateral views of the entire femur), and the age of ≥ 18 years. The femoral shaft was defined as the area between 2 cm below the lesser trochanter and the area immediately above the supracondylar ridge.[8] We excluded patients with a pathologic fracture, a history of previous surgery using metallic implant, and low-energy trauma.

A total of 139 patients representing the full spectrum of femoral shaft fractures were selected by a clinical investigator (SC). The investigator was adequately trained, had sufficient experience to select the radiographs of femoral shaft fractures, and was not involved as an observer. The mean age of the patients was 43.8 ± 19.5 years (range, 18–87 years). There were 92 men and 47 women.

Four observers who were orthopedic surgeons with 6 (WLJ), 5 (HKK), and 4 (BSL and HJW) years of experience and were familiar with the previous AO/OTA classification system participated in the analysis. All 4 observers were fellowship-trained in the same high-volume tertiary academic medical institution. They have performed clinical management of femoral shaft fractures. As the new AO/OTA classification system was published in 2018, 2 surgeons (WLJ and HKK) practiced as specialists using the previous AO/OTA classification system for 2 years and 1 year, respectively, while the other 2 observers (BSL and HJW) were only educated and trained during residency with the previous system.

In the previous AO/OTA classification, number 3 stands for the femur and number 2 for the diaphyseal segment. As shown in Fig. 1, 3 types of fractures are defined and coded with letters: type A consists of simple fractures; type B, wedge-type fractures; and type C, complex fractures. Each of these 3 types can be further subdivided into groups 1, 2, or 3. Overall, the AO/OTA classification system for femoral shaft fractures has 9 groups (32-A1/2/3, 32-B1/2/3, and 32-C1/2/3).

The 4 observers independently classified each fracture in accordance with the previous AO/OTA classification system. Radiographs had no identifying information. All radiographs were provided in random order, and the observers were given as much time as needed for accurate assessment. Each examiner was blinded to other measurements and clinical information and were not allowed to discuss their observations with other investigators.

All observers were familiar with the previous AO/

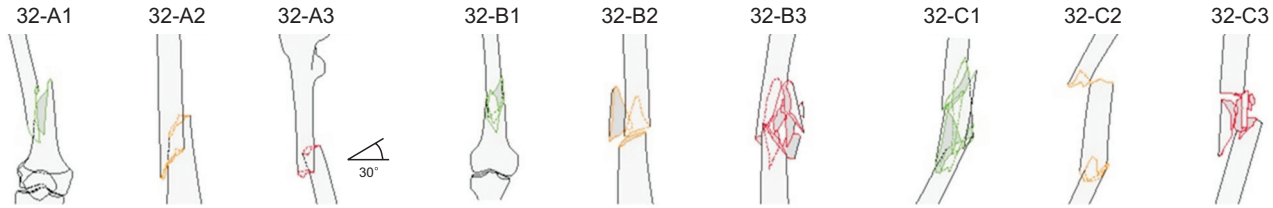

Fig. 1. Previous version of AO Foundation/Orthopaedic Trauma Association (AO/OTA) classification for diaphyseal fractures. Adapted from AO Foundation, Switzerland.[4]

OTA classification system used in this study. To ensure unambiguous application of the fracture classification system, an overview of the classification system was available to the surgeons during the classification. All data were collected by a research assistant (JSO) who did not participate in the reliability sessions. After 12 weeks, 3 consensus building sessions were held before measuring new AO/OTA classification by the 4 orthopedic surgeons.

In the revised AO/OTA classification, number 3 stands for the femur and number 2 for the diaphyseal segment. As shown in Fig. 2, 3 types of fractures are defined and coded with letters: type A consists of simple fractures; type B, wedge-type fractures; and type C, multifragmentary fractures. Differently from the previous classification, types B and C can be further subdivided into groups 2 and 3. Overall, the AO/OTA classification system for femoral shaft fractures has 7 groups (32-A1/2/3, 32-B2/3, and 32-C2/3). Each examiner analyzed each fracture according to the new revised AO/OTA classification system in the same manner. On the second occasion, images were provided in a different random order.

Statistics

Interobserver reliability is the degree of agreement when 2 or more independent observers classify the same fracture. The interobserver reliability of each radiographic measurement was evaluated with a Fleiss' kappa coefficient for each of the 4 observers (WLJ, HKK, BSL, and HJW).[9] The Fleiss' kappa coefficient represents the agreement between raters of more than 3 when assigning categorical ratings to a number of items or classifying items.[9] Interpretation of the values was carried out according to the guidelines of Gisev, which suggest that values < 0 represent poor reliability; 0.00–0.20, slight agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, substantial agreement; and 0.81–1.00, almost perfect agreement.[9,10] Data were analyzed using IBM SPSS version 21 or higher (IBM Corp.).

RESULTS

The previous classification showed the Fleiss kappa value of 0.580 (95% confidence interval [CI], 0.547–0.613), which was classified moderate reliability. Also, the new classification showed the Fleiss kappa value of 0.528 (95% CI, 0.504–0.552), which means moderate reliability. That is, both the old and the new classifications showed moderate reliability among the 4 observers.

The interobserver reliability was assessed by calculating the kappa coefficient, as proposed by Fleiss. Fig. 3 shows the interobserver concordance of each category of classification, considering the 4 observers. The highest value was A1 with a kappa of 0.734 among the previous classification; however, the highest value was C2 with a kappa of 0.791 among the new classification. The B1 and C1 categories that remained in the previous classification showed moderate agreement (0.531) and slight agreement (0.109), respectively.

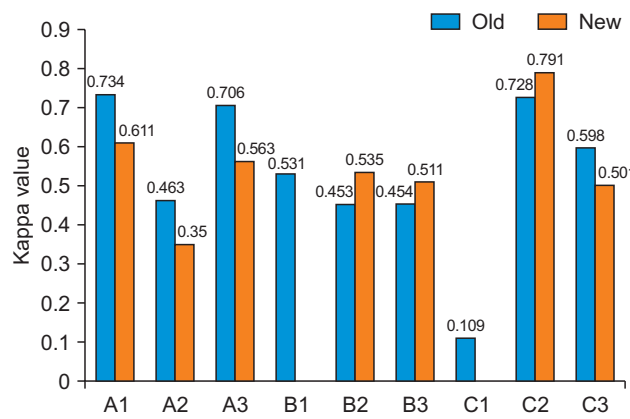

Fig. 3. Interobserver concordance of the 4 evaluators.

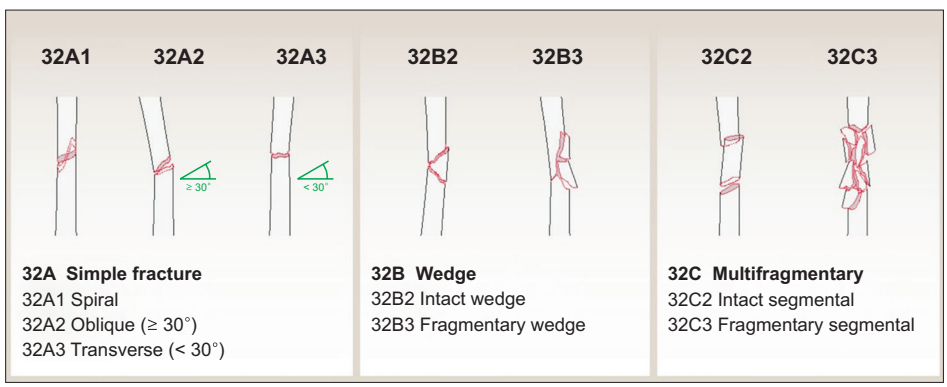

Fig. 2. The 2018 revised version of AO Foundation/Orthopaedic Trauma Association (AO/OTA) classification for diaphyseal fractures.[7,11] Adapted from AO Foundation, Switzerland.[4]

DISCUSSION

In this study analyzing the interobserver reliability among 4 experts, the previous and new classification systems showed moderate reliability with the kappa of 0.580 and 0.528, respectively. The revised AO/OTA classification system on diaphyseal fractures of the femur was different from the previous system, specifically in type B and C fractures.[7] In type B fractures, previous distinction between B1 spiral wedge and B2 bending wedge was inconsistent, and therefore changed into only B2 intact wedge and B3 fragmentary wedge.[7] Similarly, type C fractures previously consisted of C1 spiral, C2 segmental, and C3 irregular but changed into C2 intact segmental and C3 fragmentary. The naming of type C fracture was previously "complex," which was nonspecific and thus replaced by "multifragmentary."[7] Although the experts revised the classification system with aforementioned intentions, the reliability of the revised system has not been previously assessed.

Our results (kappa = 0.580) on the interobserver reliability for the previous AO/OTA classification systems are consistent with previous studies, where reliability for diaphyseal fracture classifications ranged from 0.539 to 0.82 (Table 1).[12-15] In the broader context of orthopedic classification systems, such levels of reliability are comparable to other well-known systems. For example, a study on the reliability of different classification systems for lateral clavicle fractures found mean kappa values, indicating fair to moderate agreement among specialists.[16] The OTA system, the Neer system, and the Jäger/Breitner system showed mean kappa values of 0.338, 0.278, and 0.330, respectively.[16] Similarly, another study on the Sanders and Crosby-Fitzgibbons classification systems for intra-articular calcaneal fractures reported kappa values in the moderate to substantial range for interobserver and intraobserver reliability.[17] These examples illustrate that moderate kappa values, like those found in our study, are not uncommon in orthopedic classifications, reflecting the inherent complexity and subjective aspects of fracture classification. Our findings emphasize the need for standardized training and continual refinement of classification systems to improve their reliability and clinical applicability.

Fracture classification systems play a pivotal role in orthopedic practice, guiding treatment decisions, facilitating communication, and enabling research comparability.[7] The evolution of classification systems, driven by advancements in clinical knowledge and surgical techniques, necessitates rigorous evaluation to ensure their reliability and relevance.[7] Moreover, the observed moderate interobserver reliability for both the previous and new AO/OTA classification systems underscores the challenges inherent in fracture classification, even among a panel of experienced orthopedic experts.

In this study, we aimed to assess the interobserver reliability of the new AO/OTA classification system in comparison to the older version for diaphyseal fractures of the femur, and the results offer valuable insights into the clinical applicability of these systems. The Fleiss' kappa coefficients of 0.580 and 0. 528 for the older and new systems, respectively, indicate moderate agreement among the observers, but also highlight the potential for variability in interpretation. These results suggest that while both classification systems exhibited moderate interobserver reliability, the older system displayed slightly stronger agreement. The treatment of most of the diaphyseal fractures of the femur is intramedullary nailing and treatment decisions might not vary according to the classification. However, precise classification ensures optimal management even with subtle differences in fracture patterns and can potentially improve the outcomes and reduce complications by influencing the implant selection.[18-20] Therefore, evaluating the reliability of the revised classification

Table 1. Reliability of AO/OTA Classification for Diaphyseal Fracture of Long Bone

Study	Location of fracture	Number of observers	Number of patients	Kappa of previous version	Kappa of 2018 revised version
Pignataro et al.[12]	Humerus shaft	6	60	0.539	NA
Mahabier et al.[13]	Humerus shaft	30	90	0.60 (0.59–0.61)*	NA
Slongo et al.[14]	Pediatric long bone	5	267	0.82	NA
Meling et al.[15]	Adult long bone	26	949	0.68 (0.62–0.72)*	NA
This study	Femur shaft	4	139	0.580 (0.547–0.613)*	0.528 (0.504–0.552)*

AO/OTA: AO Foundation/Orthopaedic Trauma Association, NA: not applicable.
*Kappa value (95% confidence interval).

system is very crucial not only in terms of communication and collaboration between healthcare professionals but also in specific surgical planning.

Several factors may contribute to the moderate reliability observed in this study. First, the complexity of femoral diaphyseal fractures, which can exhibit variations in fracture pattern, location, and associated soft-tissue injuries, can introduce challenges in classification agreement. Second, individual observer bias and experience can influence the interpretation of fracture characteristics, potentially leading to discrepancies in classification. Additionally, the transition from the older to the new classification system may require a period of adaptation, affecting the initial interobserver agreement with the newer system. The difference in kappa values between the 2 systems might be attributed to factors such as the familiarity of the experts with the older system due to its longer history of use.

It is noteworthy that while the moderate reliability of the new classification system might be seen as a limitation, it is also indicative of the challenges inherent in devising a classification system that captures the diverse spectrum of fracture patterns accurately.[7] Classification systems must strike a balance between being comprehensive and user-friendly to gain widespread adoption.[7] The new AO/OTA classification system aims to improve accuracy and clinical utility, and while its initial interobserver reliability is comparable to the older version, continuous education and experience may contribute to improved agreement over time.

The sample size of 139 patients in our study, determined by the eligible cases treated at our institution during the study period, reflects the natural incidence and treatment frequency of femoral shaft fractures. This number represents the maximum feasible sample given the inclusion and exclusion criteria, ensuring that our study encompassed a comprehensive range of fracture patterns within the constraints of the clinical setting. While similar studies in orthopedic fracture classification, like those by Howells et al.[21] and Lauder et al.,[17] have employed comparable sample sizes, we recognize that our sample size was contingent upon the clinical caseload rather than a pre-strategized selection. This factor underscores the real-world applicability of our findings, although we acknowledge that a larger, multicenter study could provide a broader generalization of the results.

The difference in kappa values between the 2 classification systems might reflect the familiarity of the experts with the older version due to its historical usage. This factor highlights the need for thorough training and familiarization when adopting new classification systems to mitigate potential inconsistencies during the transition phase. Continuous training and structured discussions among orthopedic surgeons can contribute to refining their understanding of classification criteria, thus enhancing interobserver reliability.[12] Another possibility is that the observers' level of experience between the 2 classification systems could be different and this could have influenced the kappa values. However, the difference between the previous and the new AO/OTA classifications is specifically confined to the subclassification in B and C fractures, and all observers had consistent training in the same tertiary academic institution.

In conclusion, to our knowledge, this is the first study to evaluate the interobserver reliability of the revised 2018 AO/OTA classification, and our study contributes to the understanding of interobserver reliability by comparing the new AO/OTA classification system with the older version for diaphyseal fractures of the femur. Our study's findings, indicating moderate interobserver reliability for both the previous and revised AO/OTA classification systems, have significant implications for clinical practice. They underscore the necessity for standardized training and continuous education among healthcare professionals to enhance consistency in fracture classification, which is pivotal for guiding treatment decisions and improving patient outcomes. Furthermore, these results highlight the need for collaborative decision-making in complex cases to ensure optimal treatment strategies, potentially influencing patient recovery trajectories and reducing the likelihood of complications.

CONFLICT OF INTEREST

No potential conflict of interest relevant to this article was reported.

ACKNOWLEDGEMENTS

The authors wish to thank Hyung Kook Kim, Beomseok Lee, and Heejae Won for participating in measurement of AO/OTA classification and Joo Ohk Sohn for assisting in data collection.

ORCID

Jung-Wee Park https://orcid.org/0000-0002-4515-1895
Woo-Lam Jo https://orcid.org/0000-0001-7021-9348
Byung Kyu Park https://orcid.org/0000-0003-2441-1472
Jong Jin Go https://orcid.org/0000-0002-4743-4434

Minji Han https://orcid.org/0000-0002-2634-3656
Sungha Chun https://orcid.org/0000-0003-0975-395X
Young-Kyun Lee https://orcid.org/0000-0001-6564-4294

REFERENCES

1. Chen YH, Liao HJ, Lin SM, Chang CH, Rwei SP, Lan TY. Radiographic outcomes of the treatment of complex femoral shaft fractures (AO/OTA 32-C) with intramedullary nailing: a retrospective analysis of different techniques. J Int Med Res. 2022;50(6):3000605221103974.
2. Court-Brown CM, Caesar B. Epidemiology of adult fractures: a review. Injury. 2006;37(8):691-7.
3. Salminen ST, Pihlajamaki HK, Avikainen VJ, Bostman OM. Population based epidemiologic and morphologic study of femoral shaft fractures. Clin Orthop Relat Res. 2000;(372):241-9.
4. Muller ME, Nazarian S, Koch P. The comprehensive classification of fractures of long bones. Springer; 1990.
5. Swiontkowski MF, Agel J, McAndrew MP, Burgess AR, MacKenzie EJ. Outcome validation of the AO/OTA fracture classification system. J Orthop Trauma. 2000;14(8):534-41.
6. Fracture and dislocation compendium. Orthopaedic Trauma Association Committee for Coding and Classification. J Orthop Trauma. 1996;10 Suppl 1:v-ix, 1-154.
7. Meinberg EG, Agel J, Roberts CS, Karam MD, Kellam JF. Fracture and dislocation classification compendium: 2018. J Orthop Trauma. 2018;32 Suppl 1:S1-170.
8. Shane E, Burr D, Abrahamsen B, et al. Atypical subtrochanteric and diaphyseal femoral fractures: second report of a task force of the American Society for Bone and Mineral Research. J Bone Miner Res. 2014;29(1):1-23.
9. Gisev N, Bell JS, Chen TF. Interrater agreement and interrater reliability: key concepts, approaches, and applications. Res Social Adm Pharm. 2013;9(3):330-8.
10. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74.
11. AO Foundation. AO/OTA Fracture and Dislocation Classification Compendium—2018 [Internet]. AO Foundation; 2024 [cited 2024 May 30]. Available from: https://www.aofoundation.org/trauma/clinical-library-and-tools/journals-and-publications/classification
12. Pignataro GS, Junqueira AE, Matsunaga FT, Matsumoto MH, Belloti JC, Tamaoki MJ. Evaluation of the reproducibility of the AO/ASIF classification for humeral
13. Mahabier KC, Van Lieshout EM, Van Der Schaaf BC, et al. Reliability and reproducibility of the OTA/AO classification for humeral shaft fractures. J Orthop Trauma. 2017;31(3):e75-80.
14. Slongo T, Audige L, Lutz N, et al. Documentation of fracture severity with the AO classification of pediatric long-bone fractures. Acta Orthop. 2007;78(2):247-53.
15. Meling T, Harboe K, Enoksen CH, Aarflot M, Arthursson AJ, Soreide K. How reliable and accurate is the AO/OTA comprehensive classification for adult long-bone fractures? J Trauma Acute Care Surg. 2012;73(1):224-31.
16. Rauer T, Boos M, Neuhaus V, et al. Inter- and intraobserver agreement of three classification systems for lateral clavicle fractures: reliability comparison between two specialist groups. Patient Saf Surg. 2020;14:4.
17. Lauder AJ, Inda DJ, Bott AM, Clare MP, Fitzgibbons TC, Mormino MA. Interobserver and intraobserver reliability of two classification systems for intra-articular calcaneal fractures. Foot Ankle Int. 2006;27(4):251-5.
18. Ostrum RF, Agarwal A, Lakatos R, Poka A. Prospective comparison of retrograde and antegrade femoral intramedullary nailing. J Orthop Trauma. 2000;14(7):496-501.
19. Ryu HG, Shin DW, Han BS, Kim SM. Risk factors associated with fixation failure in intertrochanteric fracture treated with cephalomedullary nail. Hip Pelvis. 2023;35(3):193-9.
20. Song SH. Radiologic outcomes of intramedullary nailing in infraisthmal femur-shaft fracture with or without poller screws. Biomed Res Int. 2019;2019:9412379.
21. Howells NR, Hughes AW, Jackson M, Atkins RM, Livingstone JA. Interobserver and intraobserver reliability assessment of calcaneal fracture classification systems. J Foot Ankle Surg. 2014;53(1):47-51.