The intra- and interobserver reliability of five commonly used intertrochanteric femur fracture classification systems

Cem Yıldırım, MD[1] , Osman Görkem Muratoğlu, MD[2] , Kaya Turan, MD[2] , Tugrul Ergün, MD[3] ,
Abdulhamit Mısır, MD[4] , Mahmud Aydın, MD[5]

[1]Department of Orthopedics and Traumatology, Cam and Sakura City Hospital, Istanbul, Turkey
[2]Department of Orthopedics and Traumatology, Istinye University, Istanbul, Turkey
[3]Department of Orthopedics and Traumatology, Bahçeşehir Liv Hospital, Istanbul, Turkey
[4]Department of Orthopedics and Traumatology, Medicana International Hospital, Istanbul, Turkey
[5]Department of Orthopedics and Traumatology, Haseki Training and Research Hospital, Istanbul, Turkey

Almost a half of the hip fractures are extracapsular and are subclassified as intertrochanteric and subtrochanteric.[1] Fracture stability or fracture classification systems are used for the recommendation of treatment in intertrochanteric fractures. Such classifications are also used to recommend proper implant or surgical techniques.

The ideal classification system allows interaction between physicians, guides the planning, predicts the treatment outcome, and is applicable for clinical practice and research. Examination of the fracture evaluation by the same physician and different physicians should yield the same result each time (intraobserver and interobserver reliability).

ABSTRACT

Objectives: This study aims to evaluate the effect of surgical experience on reliability for Boyd-Griffin, Evans/Jensen, Evans, Orthopaedic Trauma Association (main and subgroups), and Tronzo classification systems.

Patients and methods: Between January 2013 and December 2014, radiological images of a total of 60 patients (13 males, 47 females; mean age: 78.9±21.9 years; range, 61 to 96 years) with the diagnosis of intertrochanteric femur fracture were analyzed. Radiographs were evaluated and classified by five residents and five orthopedics and traumatology surgeons according to the Evans, Boyd-Griffin, Evans/Jensen, OTA, and Tronzo classification systems. Intra- and interobserver reliability were calculated using the kappa statistics.

Results: The worst intraobserver compatibility among the residents was the classification system with OTA subgroups (κ=0.516), while the classification system with the best intraobserver fit was found to be OTA main groups (κ=0.744). The worst agreement among surgeons was in the Evans classification system (κ=0.456). However, the best intraobserver agreement was in the OTA main groups (κ=0.741). The best interobserver agreement was observed regarding the OTA main groups (κ=0.699).

Conclusion: The classification that has the best harmony both among residents and surgeons, and between residents and surgeons is the OTA main group classification.

Keywords: Classification, fracture, interobserver, intertrochanteric, intraobserver.

Several classification systems are used for the classification of extra-capsular hip fractures.[2-5] Most utilized is the Evans classification system modified by Jensen and Michaelsen.[4] Recently, Arbeitsgemeinschaft für Osteosynthesefragen (AO) classification system has been introduced. Despite the

widespread use of these systems and the thousands of publications regarding hip fracture, few studies have evaluated the reliability of classification systems and even fewer studies have investigated the reliability of experienced physicians using the classification systems.[6,7]

Evans[2] described an anatomical classification based on the number of fragments and whether the lesser trochanter is split off as a separate fragment. The Jensen modification of Evans' classification consists of five subtypes regarding displacement, the number of fracture fragments, and posteromedial and medial support.[5] The Orthopaedic Trauma Association (OTA) classification for trochanteric femur fractures is built up by three groups of possible types of fractures and then according to increasing fracture severity divided into the subgroups A, B, or C.[8] Tronzo[9] subdivided these fractures into five types according to stability, posteromedial comminution, and fracture line extension. Boyd and Griffin[2] described another classification according to more or less fracture line extension, comminution, subtrochanteric involvement, and extension to the shaft.

In the literature, there has been no comprehensive study evaluating intra- and interobserver reliability and the effect of surgeon experience for the five most used intertrochanteric femur fracture classification systems. In this study, therefore, we hypothesized that the interobserver reliability between senior residents and surgeons for intertrochanteric femur fracture classification systems was moderate and intraobserver reliability of the AO/OTA-main group was better than the other classification system. We aimed to compare inter- and intraobserver reliability of five different intertrochanteric fracture classification systems (Evans, Boyd-Griffin, Evans/Jensen, AO, Tronzo) by two groups of physicians with different ranges of experience.

PATIENTS AND METHODS

This retrospective study was conducted at Haseki Training and Research Hospital, Department of Orthopedics and Traumatology between January 2013 and December 2014. Preoperative anteroposterior and lateral radiographs of intertrochanteric femur fractures treated surgically were screened. Anteroposterior and lateral radiographs of patients showing femoral fractures, which were obtained randomly and retrospectively from hospital data, were selected in the study. Radiographs were carefully chosen by the investigator who was blinded to the study protocol.

Radiological imaging of a total of 60 patients (13 males, 47 females; mean age: 78.9±21.9 years; range, 61 to 96 years) was performed. Patients' data were not shared with the participants. Radiographs eligible for participation in the study were of adequate radiological quality to allow the investigator to classify the fracture and demonstrate extracapsular hip fracture. There was only anteroposterior and lateral view of fractured hip (Figure 1). Selected radiographs were evaluated by five senior residents and five orthopedic surgeons, each with more than five years of experience in orthopedic trauma. Each observer was given brief information on the original illustrations of the Evans, Boyd-Griffin, Evans/Jensen, AO, and Tronzo classification systems. A written informed consent was obtained from each patient. The study protocol was approved by the Istinye University Ethics Committee (date/no: 24.11.2021-2/2021.K-88). The study was conducted in accordance with the principles of the Declaration of Helsinki.

With no clinical information, observers blindly evaluated the radiographs according to the separately defined classification systems. In similar studies, the three-month period was considered appropriate to avoid recall bias.[10,11] Therefore, three months after the first evaluation, the researchers re-evaluated the same radiographs shown in a different order than on the first occasion. During this three-month
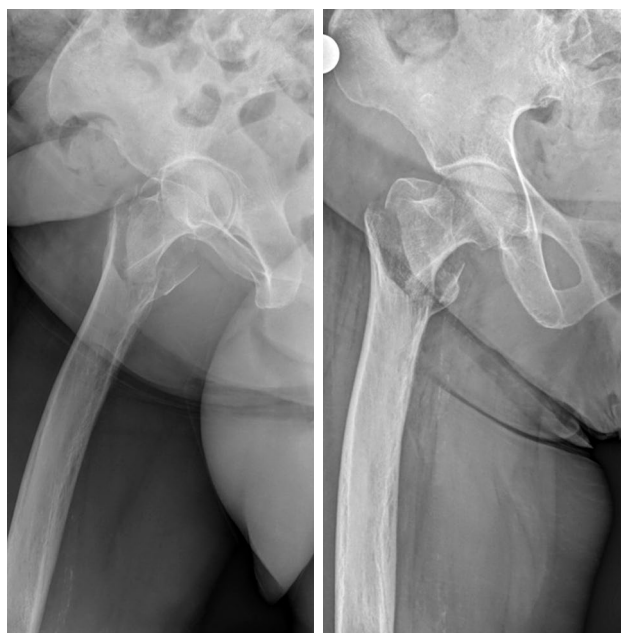

FIGURE 1. Anteroposterior and lateral X-ray views of a 62-year-old female patient.

	Resident group intraobserver (95% CI)	Surgeon group intraobserver (95% CI)	Resident-surgeon group interobserver (95% CI)
TABLE I Intraobserver kappa values of resident group and surgeon group, and interobserver kappa value of resident-surgeon evaluations			
Boyd-Griffin classification	0.660 (0.550-0.770)	0.658 (0.550-0.770)	0.572 (0.532-0.616)
Evans-Jensen classification	0.625 (0.600-0.655)	0.484 (0.434-0.542)	0.498 (0.450-0.553)
Evans classification	0.557 (0.519-0.595)	0.456 (0.409-0.053)	0.438 (0.400-0.481)
AO/OTA-main group	0.744 (0.708-0.785)	0.741 (0.696-0.797)	0.699 (0.649-0.750)
AO/OTA-subgroup	0.516 (0.498-0.540)	0.488 (0.418-0.558)	0.444 (0.418-0.470)
Tronzo classification	0.528 (0.501-0.562)	0.529 (0.489-0.569)	0.554 (0.506-0.614)
CI: Confidence interval; AO/OTA: Arbeitsgemeinschaft für Osteosynthesefragen/Orthopaedic Trauma Association.			

period, 60 designated radiographs were withheld from researchers.

Statistical analysis

For intraobserver reliability, Cohen's kappa value (κ) was obtained using the IBM SPSS for Windows version 21.0 software (IBM Corp., Armonk, NY, USA). Landis and Koch[12] defined values exceeding 0.80 as almost perfect compliance; values between 0.61 and 0.80 as substantial; values between 0.41 and 0.60 as moderate; values between 0.21 and 0.40 as fair; and values between zero and 0.21 as low. Descriptive data were expressed in mean ± standard deviation (SD), median (min-max) or number and frequency. A p value of <0.05 was considered statistically significant.

RESULTS

Interobserver and intraobserver agreement in all classification systems were not significantly different between experienced surgeons and senior residents (p>0.05).

Interobserver agreement

Interobserver agreement for the Boyd-Griffin classification system was moderate for experienced orthopedic surgeons and senior residents (κ=0.572; 95% confidence interval [CI]: 0.532-0.616). Interobserver agreement for Evans/Jensen classification for experienced surgeons and senior residents (κ=0.498 95% CI: 0.450-0.553) and Evans classification (κ=0.438 95% CI: 0.400-0.481) was moderate. In AO/OTA, subgroup agreement was moderate (κ=0.444 95% CI: 0.418-0.470); however, in AO/OTA, the main group agreement was substantial for surgeons and late-term residents (κ=0.699 95% CI: 0.649-0.750). Tronzo classification also showed a moderate agreement for surgeons and late-term residents (κ=0.554 95% CI: 0.506-0.614).

Intraobserver agreement

In the repeated evaluation three months after the first assessment, for the Boyd and Griffin classification experienced surgeons (κ=0.658; 95% CI: 0.550-0.770) obtained a substantial intraobserver agreement similar to senior residents (κ=0.66; 95% CI: 0.550-0.770). When we evaluated intraobserver agreement according to the Evans/Jensen classification system, orthopedic surgeons achieved a moderate agreement (κ=0.484; 95% CI: 0.434-0.542), while senior residents achieved a substantial agreement (κ=0.625; 95% CI: 0.600-0.655). In Evans classification, the resident group and surgeon group showed a moderate agreement (κ=0.557; 95% CI: 0.519-0.595/κ=0.456; 95% CI: 0.409-0.503, respectively). When AO/OTA classification was evaluated, the agreement between the senior residents and the experienced surgeons was moderate in the AO/OTA subgroup (κ=0.516; 95% CI: 0.498-0.540/κ=0.488; 95% CI: 0.418-0.558, respectively) and substantial in the AO/OTA main group (κ=0.744; 95% CI: 0.708-0.785/ κ=0.741; 95% CI: 0.696-0.797, respectively). In Tronzo classification, senior residents and experienced surgeons also showed a moderate agreement (κ=0.528; 95% CI: 0.501-0.562/κ=0.529; 95% CI: 0.489-0.569, respectively) (Table I).

DISCUSSION

An ideal classification system creates a platform for universal communication among surgeons regarding common scenarios and methods for treatment. Classification systems should be both easy to understand and have good interobserver and intraobserver compatibility. This is the first study evaluating five established classification systems with the same surgeons and residents together.

The Evans/Jensen and AO/OTA classifications are the most used intertrochanteric fracture classifications. Although the AO/OTA classification

provided higher agreement than the Evans/Jensen classification, Fung et al.[6] reported that it was insufficient for compatibility. Pervez et al.[10] showed that the AO/OTA main groups had better compatibility than the Evans/Jensen and subgroups together with the AO/OTA classification. Schipper et al.[13] found that the AO/OTA main groups had a better agreement than the classification with subgroups. Zarie et al.[7] reported that the agreement of AO/OTA classification system was weak. In line with the literature, our study showed that, although the best interobserver and intraobserver compliance was found in the AO/OTA main groups, the compliance was substantial even among experienced surgeons (κ=0.744/0.741 – intraobserver, κ=0.699-intraobserver).

In many studies investigating the reliability of fracture classification systems, the images were also evaluated by the residents and experience was shown to increase the reliability of the classifications used in the studies.[13,14] The low agreement on fracture classification among residents used in the studies may be explained by their lack of surgical experience. Gehrchen et al.[15] evaluated intertrochanteric femur fracture in 52 radiographs according to Evan/Jensen classification by two senior residents and two junior residents and did not detect a significant difference in agreement with increasing experience. Behrendt et al.[9] compared Tronzo and AO/OTA classifications, reporting that, similar to our study, experience had no effect on the results, but that the AO/OTA main groups were more compatible than Tronzo. To avoid the limitations described in previous studies, professionals with a wider difference in experience level were included in the study, and there were five assessors in each category, more than similar studies. In our study, there was no significant statistical difference between the interobserver agreement between surgeons and residents. The reason for this is that, despite the difference in experience among the two groups, we believe that the senior residents in our study had sufficient experience in intertrochanteric femur fractures.

In their study, Jin et al.[16] showed that AO/OTA main groups were more concordant than other classifications, but the concordance was much lower when evaluated with subgroups. Our study showed that the compatibility of the AO/OTA classification decreased when evaluated together with the subgroups. Klaber et al.[17] evaluated the compatibility of the new AO/OTA classification defined in 2018 and showed that the new AO/OTA classification system had better interobserver and intraobserver agreement than the classical system.

In complex patterns of intertrochanteric fractures, a better radiological evaluation may help to evaluate the treatment plan and more reliable fracture classification. Computed tomography (CT) and plain radiography have been compared in recent studies for different types of fractures with complicated fracture patterns, such as tibial plateau or calcaneal fractures, and CT has proven superior.[18-20] Cavaignac et al.[21] evaluated the effect of CT on AO/OTA and Evans/Jensen classification systems and showed that it provided a clearer understanding of the fracture, but did not increase interobserver agreement on the classification systems. Another study evaluating the effects of three-dimensional CT examinations on fracture classification systems showed three-dimensional CT to succeed in determining the stability and, thus, implant options, but this study obtained similar results with Cavaignac et al.'s[21] study regarding compliance.[22] Our study did not include CT in fractures due to the additional cost and radiation exposure to participants.

An ideal fracture classification system should provide information on fracture stability and have a high degree of reproducibility. The common philosophy of classification systems designed for intertrochanteric femur fractures is whether the fracture is stable or not. A study by van Embden et al.,[11] which compared the AO/OTA and Jensen classification systems for intertrochanteric femur fractures, showed low agreement among participants on the assessment of a trochanteric fracture as either stable or unstable. Intertrochanteric fractures with four-part, reverse oblique and medial cortical discontinuities are usually considered unstable. However, there is not enough evidence in the literature on this subject.[23,24] Several articles in the literature failed to provide a consensus on fracture stability, although some studies have suggested that medial structural continuity is vital,[2] while Palm et al.[25] and Gotfried[26] reported that an intact lateral wall played a key role in stabilization and fixation of intertrochanteric fractures. In particular, in the Tronzo classification, it may be difficult to interpret the stability of the fracture, as although the discontinuity of the posteromedial wall indicates instability, the lesser trochanter may also be fractured in Tronzo type 2 which is stable.[27] Confusion in the concept of stable fracture probably explains the low agreement in the Tronzo, OTA-subgroups, and Evans/Jensen classifications in our study.

Some of the intertrochanteric classification systems are similar and intertwined: the AO/OTA group A1 fractures are displaced or nondisplaced two-part trochanteric fractures, equivalent to Jensen classification 1,2 and Tronzo 1,2. Group AO/OTA A2 intertrochanteric fractures are comminuted and unstable and are equivalent to Jensen types III, IV, V, and Tronzo types III, IV. Group AO/OTA A3 intertrochanteric fractures are at the level of the lesser trochanter and may be reversed, transverse or oblique. In reverse oblique fractures, the fracture line extends from medial to lateral, from proximal to distal. This fracture group was classified as type V in Tronzo classification, type III in Boyd and Griffin classification, type II in Evans classification, and included in other groups in Jensen's modification. Clinical studies showing an increased risk of fixation failure for reverse oblique fractures, and intramedullary fixation has been recommended in these studies.[28] The reason why current classification systems are inconsistent with the complexity of intertrochanteric fractures may be that these classification systems focus on well-known fracture features such as four-part fractures, reverse oblique fractures, disruption of medial cortical continuity and do not consider less important fracture features as intact lateral wall. By revising the current classification systems and using CT, which can provide three-dimensional imaging, instead of direct radiography, fracture stability, treatment options, and agreement in fracture classification systems can be increased.

In our study, all the physicians who evaluated the classification systems were working in the same clinic. Physicians working in the same clinic are likely to have the same approach in intertrochanteric fracture classification as in all fracture types.[29] Additionally, our study does not focus on implant selection based on determining whether an intertrochanteric fracture is stable or unstable.

In conclusion, none of the commonly used classification systems for trochanteric fractures accurately describe intertrochanteric fractures. To improve current fracture management, existing classification systems should be revised by learning more about fracture characteristics, biomechanical properties of fractures, and their understanding, as well as the definition of successful fracture reduction. The use of CT, which shows the fracture in more detail, would facilitate the understanding of the fracture. Despite all these, we believe that the current AO/OTA classification for intertrochanteric fractures, using the three main types, allows for a common language among treating physicians.

REFERENCES

1. Karagas MR, Lu-Yao GL, Barrett JA, Beach ML, Baron JA. Heterogeneity of hip fracture: Age, race, sex, and geographic patterns of femoral neck and trochanteric fractures among the US elderly. Am J Epidemiol 1996;143:677-82.

2. Evans EM. The treatment of trochanteric fractures of the femur. J Bone Joint Surg Br 1949;31B:190-203.

3. Boyd HB, Griffin LL. Classification and treatment of trochanteric fractures. Arch Surg 1949;58:853-66.

4. Jensen JS, Michaelsen M. Trochanteric femoral fractures treated with McLaughlin osteosynthesis. Acta Orthop Scand 1975;46:795-803.

5. Jensen JS. Classification of trochanteric fractures. Acta Orthop Scand 1980;51:803-10.

6. Fung W, Jonsson A, Buhren V, Bhandari M. Classifying intertrochanteric fractures of the proximal femur: Does experience matter? Med Princ Pract 2007;16:198-202.

7. Zarie M, Mohamoud MF, Farhoud AR, Bagheri N, Khan FMY, Heshmatifar M, et al. Evaluation of the inter and intra-observer reliability of the AO classification of intertrochanteric fractures and the device choice (DHS, PFNA, and DCS) of Fixations. Ethiop J Health Sci 2020;30:755-60.

8. Müller ME, Koch P, Nazarian S, Schatzker J. The comprehensive classification of fractures of long bones. Berlin, Heidelberg: Springer Berlin Heidelberg; 1990.

9. Behrendt C, Faleiro TB, Schulz Rda S, Silva BO, Paula Filho EQ. Repruducibility of Tronzo and AO/ASIF classifications for transtrochanteric fractures. Acta Ortop Bras 2014;22:275-7.

10. Pervez H, Parker MJ, Pryor GA, Lutchman L, Chirodian N. Classification of trochanteric fracture of the proximal femur: A study of the reliability of current systems. Injury 2002;33:713-5.

11. van Embden D, Rhemrev SJ, Meylaerts SA, Roukema GR. The comparison of two classifications for trochanteric femur fractures: the AO/ASIF classification and the Jensen classification. Injury 2010;41:377-81.

12. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.

13. Schipper IB, Steyerberg EW, Castelein RM, van Vugt AB. Reliability of the AO/ASIF classification for pertrochanteric femoral fractures. Acta Orthop Scand 2001;72:36-41.

14. Kreder HJ, Hanel DP, McKee M, Jupiter J, McGillivary G, Swiontkowski MF. Consistency of AO fracture classification for the distal radius. J Bone Joint Surg [Br] 1996;78:726-31.

15. Gehrchen PM, Nielsen JO, Olesen B. Poor reproducibility of Evans' classification of the trochanteric fracture. Assessment of 4 observers in 52 cases. Acta Orthop Scand 1993;64:71-2.

16. Jin WJ, Dai LY, Cui YM, Zhou Q, Jiang LS, Lu H. Reliability of classification systems for intertrochanteric fractures of the proximal femur in experienced orthopaedic surgeons. Injury 2005;36:858-61.

17. Klaber I, Besa P, Sandoval F, Lobos D, Zamora T, Schweitzer D, et al. The new AO classification system for intertrochanteric fractures allows better agreement than the original AO classification. An inter- and intra-observer agreement evaluation. Injury 2021;52:102-5.

18. Humphrey CA, Dirschl DR, Ellis TJ. Interobserver reliability of a CT-based fracture classification system. J Orthop Trauma 2005;19:616-22.

19. Beaulé PE, Dorey FJ, Matta JM. Letournel classification for acetabular fractures. Assessment of interobserver and intraobserver reliability. J Bone Joint Surg [Am] 2003;85:1704-9.

20. Mustonen AO, Koskinen SK, Kiuru MJ. Acute knee trauma: Analysis of multidetector computed tomography findings and comparison with conventional radiography. Acta Radiol 2005;46:866-74.

21. Cavaignac E, Lecoq M, Ponsot A, Moine A, Bonnevialle N, Mansat P, et al. CT scan does not improve the reproducibility of trochanteric fracture classification: A prospective observational study of 53 cases. Orthop Traumatol Surg Res 2013;99:46-51.

22. Cho YC, Lee PY, Lee CH, Chen CH, Lin YM. Three-dimensional CT improves the reproducibility of stability evaluation for intertrochanteric fractures. Orthop Surg 2018;10:212-7.

23. Sarmiento A, Williams EM. The unstable intertrochanteric fracture: Treatment with a valgus osteotomy and I-beam nail-plate. A preliminary report of one hundred cases. J Bone Joint Surg [Am] 1970;52:1309-18.

24. Dimon JH, Hughston JC. Unstable intertrochanteric fractures of the hip. J Bone Joint Surg Am 1967;49:440-50.

25. Palm H, Jacobsen S, Sonne-Holm S, Gebuhr P; Hip Fracture Study Group. Integrity of the lateral femoral wall in intertrochanteric hip fractures: An important predictor of a reoperation. J Bone Joint Surg [Am] 2007;89:470-5.

26. Gotfried Y. The lateral trochanteric wall: A key element in the reconstruction of unstable pertrochanteric hip fractures. Clin Orthop Relat Res 2004;(425):82-6.

27. Tronzo RG. Use of an extramedullary guide pin for fractures of the upper end of the femur. Orthop Clin North Am 1974;5:525-7.

28. Haidukewych GJ, Israel TA, Berry DJ. Reverse obliquity fractures of the intertrochanteric region of the femur. J Bone Joint Surg [Am] 2001;83:643-50.

29. Atik OŞ. What are the expectations of an editor from a scientific article? Jt Dis Relat Surg 2020;31:597-8.