Eklem Hastalıkları ve Cerrahisi
Joint Diseases and Related Surgery

Original Article / ***Özgün Makale***

Eklem Hastalik Cerrahisi
2017;28(3):177-181
doi: 10.5606/ehc.2017.56816

Intraobserver and interobserver reliability assessment of tibial plateau fracture classification systems

Tibia plato kırıkları sınıflandırma sistemlerinin gözlemcilerin kendi içlerinde ve aralarındaki güvenirliğinin değerlendirilmesi

Anıl Taşkesen, MD.,[1] İsmail Demirkale, MD.,[1] Mustafa Caner Okkaoğlu, MD.,[1] Mahmut Özdemir, MD.,[1] Mustafa Gökhan Bilgili, MD.,[2] Murat Altay, MD.[1]

[1]Department of Orthopedics and Traumatology, University of Health Sciences, Keçiören Training and Research Hospital, Ankara, Turkey

[2]Department of Orthopedics and Traumatology, University of Health Sciences, Bakırköy Dr. Sadi Konuk Training and Research Hospital, Istanbul, Turkey

ABSTRACT

Objectives: This study aims to assess the intra- and interobserver reliability of commonly used tibial plateau fracture classification systems.

Patients and methods: This retrospective cohort study included computed tomography (CT) and plain radiographic images (lateral and anteroposterior X-rays) of 60 patients (40 males, 20 females; mean age 45.9 years; range 18 to 80 years) who presented to two orthopaedic clinics between January 2011 and January 2015 with unilateral tibial plateau fractures. All plain X-rays (XR) and CT images were evaluated by four observers on two separate occasions, 1.5 months apart. All fractures were classified according to the Arbeitsgemeinschaft für Osteosynthesefragen-Orthopaedic Trauma Association (AO-OTA), Schatzker, Hohl and Moore, Luo and revised Duparc systems. Intraobserver reliability was measured with Cohen's kappa (κ) coefficient and interobserver reliability with Fleiss' kappa coefficient.

Results: When Schatzker classification was performed, interobserver reliability was in moderate level for (κ=0.51) for XR and in substantial level for CT (κ=0.61). When AO/OTA classification was used, interobserver reliability was in moderate level for both methods of diagnosis (κXR=0.43 and κCT=0.54, respectively). In the Hohl and Moore classification, the interobserver reliability was also moderate for both methods of diagnosis (κXR=0.45 and κCT=0.51, respectively). Revised Duparc classification showed the lowest interobserver reliability ranging from fair to moderate level (κXR=0.27-0.55 and κCT=0.44-0.61). Interobserver reliability for Luo classification was κCT=0.47. Intraobserver reliability for CT in Luo classification was in substantial level for observers 1, 2 and 3 (κCT=0.67-0.71) and in perfect level for observer 4 (κCT=0.84). Intraobserver reliability was in substantial level in Schatzker classification and in moderate level at the other classifications.

Conclusion: Among the classification systems compared in this study, Schatzker was the most reliable particularly when CT was used. On the other hand, revised Duparc classification presented the worse reliability results due to its complexity and different morphological subtypes.

Keywords: Classification; interobserver variation; reliability; tibial plateau fracture.

ÖZ

Amaç: Bu çalışmada sıkça kullanılan tibia plato kırıklarının sınıflandırma sistemlerinin gözlemcilerin kendi içlerinde ve aralarındaki güvenirliği değerlendirildi.

Hastalar ve yöntemler: Bu retrospektif kohort çalışmaya iki ortopedi kliniğine Ocak 2011 - Ocak 2015 tarihleri arasında tek taraflı tibia plato kırığı nedeniyle başvuran 60 hastanın (40 erkek, 20 kadın; ort. yaş 45.9 yıl; dağılım 18-80 yıl) bilgisayarlı tomografi (BT) ve düz radyografik görüntüleri (yan ve ön-arka grafiler) dahil edildi. Tüm düz grafiler (XR) ve BT görüntüleri dört gözlemci tarafından 1.5 ay arayla iki farklı zamanda değerlendirildi. Tüm kırıklar Arbeitsgemeinschaft für Osteosynthesefragen-Orthopaedic Trauma Association (AO-OTA), Schatzker, Hohl ve Moore, Luo ve revize edilmiş Duparc sistemlerine göre sınıflandırıldı. Gözlemcilerin kendi içlerindeki güvenirlik Cohen'in kappa (κ) katsayısı, gözlemcilerin aralarındaki güvenirlik Fleiss kappa katsayısı ile ölçüldü.

Bulgular: Schatzker sınıflandırması yapıldığında gözlemcilerin aralarındaki güvenilirlik XR için orta düzeyde (κ=0.51), BT için tatmin edici düzeyde idi (κ=0.61). AO/OTA sınıflandırması kullanıldığında gözlemcilerin aralarındaki güvenilirlik iki tanı yöntemi için de orta düzeyde idi (sırasıyla, κXR=0.43 ve κBT=0.54). Hohl ve Moore sınıflandırmasında yine gözlemcilerin aralarındaki güvenirlik iki tanı yöntemi için de orta düzeyde idi (sırasıyla, κXR=0.45 ve κBT=0.51). revize edilmiş Duparc sınıflandırması gözlemciler arasında vasat ile orta düzeyde olarak en az güvenirlik gösterdi (sırasıyla, κXR=0.27-0.55 ve κBT=0.44-0.61). Luo sınıflandırması için gözlemcilerin aralarındaki güvenirlik κBT=0.47 idi. Gözlemcilerin kendi içlerindeki güvenirlik BT için Luo sınıflandırmasında 1, 2 ve 3. gözlemciler için tatmin edici düzeyde (κBT=0.67-0.71), 4. gözlemci için mükemmel (κBT=0.84) düzeyde idi. Gözlemcilerin kendi içlerindeki güvenirlik Schatzker sınıflandırmasında tatmin edici düzeyde, diğer sınıflandırmalarda orta düzeyde idi.

Sonuç: Bu çalışmada karşılaştırılan sınıflandırma sistemlerinin arasında özellikle BT kullanıldığında en güvenilir olan Schatzker idi. Revize edilmiş Duparc sınıflandırması ise, kompleksitesi ve farklı morfolojik alt tipleri nedeniyle en kötü güvenirlik sonuçları gösterdi.

Anahtar sözcükler: Sınıflandırma; gözlemcilerin aralarındaki farklılık; güvenirlik; tibia plato kırığı.

- *Received: June 13, 2017 Accepted: September 14, 2017*

- Correspondence: Mustafa Caner Okkaoğlu, MD. Ümit Mah., 2449. Sok., Defne 5 Sit., No: 6/30, 06810 Çankaya, Ankara, Turkey,
  Tel: +90 507 - 958 82 85 e-mail: canerokkaoglu@yahoo.com

The annual incidence of tibial plateau fractures is 10.3 per 100,000 and more than 90% of them are surgical candidates.[1] The most common type is Arbeitsgemeinschaft für Osteosynthesefragen-Orthopaedic Trauma Association (AO-OTA) type 41-B3 and represents 35% of plateau fractures. Most fractures are complex and comminuted, occasionally soft tissue injuries accompany them and treatment outcomes depend on anatomic reduction, stability degree and early active mobilization of the joint. Since management is primarily determined by a proper classification, one would expect that multiple studies should have evaluated the degree of observer agreement for imaging classifications.

When used by either different surgeons or specialists as radiologists, ideally the results must be consistent for a classification system for different cases both to guide management and speak the same language. Furthermore, for research purposes, a classification system has to provide a consensus to accurately compare different cases. In addition, some information can be found in previous studies in this field, and this study provides comprehensive investigation on all classifications regarding tibial plateau fractures. The accuracy and reliability of a classification can be resolved by its reproducibility, that is, by different observers. AO-OTA,[2] Schatzker et al.,[3] Hohl,[4] Luo et al.[5] and revised Duparc[6] systems are widely used in orthopaedic practice, also being extensively cited in the literature.[7] Nevertheless, there is no universal consensus regarding the reproducibility of these classification systems for tibial plateau fractures nor a study that comprehensively reviews the reproducibility of all these classifications en masse. Given this gap in the literature, in this study, we aimed to assess the intra- and interobserver reliability of commonly used tibial plateau fracture classification systems.[8]

PATIENTS AND METHODS

Computed tomography (CT) and lateral and anteroposterior (AP) views of plain radiographic images of 60 consecutive adult patients (40 males, 20 females; mean age 45.9 years; range 18 to 80 years) presenting to two orthopaedic departments between January 2011 and January 2015 with unilateral tibial plateau fractures were included in this retrospective cohort study. The orthopaedic departments are teaching hospitals that accept tertiary referrals for trauma. The observers were unaware of the patient identity and all aspects of clinical care and had no previous exposure to the films or the patients involved. Plain anteroposterior (AP), lateral radiographs (XR) and axial, coronal and sagittal CT views were made at initial referral to the hospitals and consequently used for analysis. The most recent radiographs of each patient were used. The order of the films was varied for the repeat classification assessment to prevent recall bias. The films were screened before inclusion by the non-observer senior authors to ensure that the films seemed to represent a full range of levels of the classification systems. The observers were not involved in this screening process. Four observers reviewed all XR and CT scans independently on two separate occasions, 1.5 months apart. Two observers were senior trauma surgeons with minimum 10 years of experience. The other two observers were staff surgeons with less than five years of experience who had been given detailed instructions on the relevant surgical anatomy and classification systems and had reviewed more than 10 different films by way of a learning curve before the present study. They were given instructions separately to discourage rote application of instructions from any senior author when classifying the injuries. All fractures were classified according to the AO-OTA,[2] Schatzker et al.,[3] Hohl,[4] Luo et al.[5] and revised Duparc[6] systems. Comprehensive information was available to the observers regarding each of the classification systems in the forms. No time limit was used. A written informed consent was obtained from each patient. The study was conducted in accordance with the principles of the Declaration of Helsinki.

Statistical analysis

Statistical analysis was performed using the SPSS, version 13.0 (SPSS Inc., Chicago, IL, USA): Cohen's kappa (κ) for intraobserver reliability and Fleiss kappa for interobserver reliability. The Landis and Koch interpretation of kappa's values ($\kappa \geq 0.81$ equals almost perfect, κ=0.61–0.8 as substantial, κ=0.41-0.6 as moderate, κ=0.21–0.4 as fair, and $\kappa \leq$ 0.2 as slight correlation) were used. Although this interpretation is arbitrary, they have been well accepted and widely used in the orthopaedic literature. Evaluation of statistical differences between kappa values was calculated with 95% confidence interval and considered significant when the upper and lower boundaries did not overlap.

RESULTS

Interobserver reliability when Schatzker classification was used showed moderate agreement (κ=0.51) for XR and substantial agreement for CT (κ=0.61) (Tables I and II). Intraobserver reliability when using XR indicated moderate agreement for observers 1, 2 and 3 (κ_{XR}=0.54-0.6) and substantial agreement

TABLE I

Kappa coefficients for interobserver reliability when using X-ray

	Observers							
Classification	1/2	1/3	1/4	2/3	2/4	3/4	1-2-3-4	Level of agreement
Schatzker	0.49	0.52	0.61	0.37	0.5	0.54	0.51	Moderate
AO/OTA	0.38	0.49	0.55	0.26	0.43	0.49	0.43	Moderate
Hohl and Moore	0.46	0.47	0.41	0.34	0.44	0.55	0.45	Moderate
Duparc	0.38	0.39	0.55	0.27	0.35	0.41	0.39	Fair

AO/OTA: Arbeitsgemeinschaft für Osteosynthesefragen-Orthopaedic Trauma Association.

TABLE II

Kappa coefficients for interobserver reliability when using computed tomography

	Observers							
Classification	1/2	1/3	1/4	2/3	2/4	3/4	1-2-3-4	Level of agreement
Schatzker	0.6	0.59	0.68	0.55	0.61	0.61	0.61	Substantial
AO/OTA	0.53	0.54	0.55	0.48	0.55	0.57	0.54	Moderate
Hohl and Moore	0.52	0.54	0.46	0.48	0.51	0.55	0.51	Moderate
Duparc	0.5	0.61	0.58	0.44	0.45	0.54	0.52	Moderate
Luo	0.5	0.46	0.44	0.55	0.54	0.35	0.47	Moderate

AO/OTA: Arbeitsgemeinschaft für Osteosynthesefragen-Orthopaedic Trauma Association.

for observer 4 (κ_{XR}=0.64) (Table III). When using CT, intraobserver reliability showed moderate (observer 1: κ_{CT}=0.53) to substantial agreement (observers 2, 3 and 4: κ_{CT}=0.63-0.77).

When AO/OTA classification was used, interobserver reliability between the four observers showed also moderate agreement for both methods of diagnosis (κ_{XR}=0.43 and κ_{CT}=0.54, respectively) (Tables I and II). Intraobserver reliability ranged from fair (observer 1: κ_{XR}=0.4 and observer 3: κ_{XR}=0.39) to moderate (observer 2: κ_{XR}=0.48) agreement and only observer 4 had substantial agreement (κ_{XR}=0.67) when using XR (Table III). However, the intraobserver reliability showed moderate agreement (κ_{CT}=0.49 to 0.6) for the first three observers and substantial agreement for the last observer (κ_{CT}=0.76) when using CT.

In the Hohl and Moore classification, the interobserver agreement was also moderate for both methods (κ_{XR}=0.45 and κ_{CT}=0.51, respectively) (Tables I and II). Intraobserver reliability was moderate (observers 1, 2 and 3: κ_{XR}=0.52) and substantial (observer 4: κ_{XR}=0.72) when using XR and was moderate (observers 1 and 3: κ_{CT}=0.52 and 0.58), substantial (observer 2: κ_{CT}=0.66) or almost perfect (observer 4: κ_{CT}=0.88) when using CT (Table III).

The revised Duparc classification showed the lowest interobserver reliability ranging from fair to

TABLE III

Kappa coefficients for intraobserver reliability

	Intraobserver reliability κ_{XR}/κ_{CT}			
Classification	Observer 1	Observer 2	Observer 3	Observer 4
Schatzker	0.56/0.53	0.6/0.72	0.54/0.63	0.64/0.77
AO/OTA	0.4/0.49	0.48/0.6	0.39/0.51	0.67/0.76
Hohl and Moore	0.51/0.52	0.52/0.66	0.53/0.58	0.72/0.88
Duparc	0.45/0.44	0.52/0.69	0.45/0.57	0.62/0.75
Luo	---/0.68	---/0.71	---/0.67	---/0.84

κ: Kappa; XR: X-ray: CT: Computed tomography; AO/OTA: Arbeitsgemeinschaft für Osteosynthesefragen-Orthopaedic Trauma Association.

moderate agreement (κ_{XR}=0.27-0.55 and κ_{CT}=0.44-0.61, respectively) (Tables I and II). Overall, the interobserver reliability was fair when using XR and moderate when using CT (κ_{XR}=0.39 and κ_{CT}=0.52, respectively). Also, the intraobserver reliability was moderate (κ_{XR}=0.45-0.6) when using XR, but substantial for observers 2 and 4 (κ=0.69 and κ=0.75, respectively) and moderate for observers 1 and 3 (κ_{CT}=0.44 and κ_{CT}=0.57, respectively) when using CT (Table III).

Interobserver reliability among observers was κ_{CT}=0.47. The intraobserver reliability yielded substantial agreement for observers 1, 2 and 3 (κ_{CT}=0.67-0.71) and almost perfect agreement was obtained by observer 4 (κ_{CT}=0.84) (Table III).

DISCUSSION

To our knowledge, this is the first study assessing the intraobserver and interobserver reliability of all classification systems for tibial plateau fractures. Although these classification systems have been described for plateau fractures, none has been universally accepted. Our study demonstrated that although difference in κ value was too small to be clinically relevant, the Schatzker classification had moderate interobserver agreement when XR was used and substantial interobserver agreement when CT was used. Also, the AO-OTA and Hohl and Moore classifications demonstrated moderate rate of agreement. Finally, a fair concordance was found for revised Duparc classification.

First, the κ statistics achieved for the revised Duparc classification in the present study were similar for interobserver reliability (κ_{XR}=0.39; κ_{CT}=0.52) to a previously reported study that reported a mean κ_{XR}=0.365; κ_{CT}=0.474).[8] However, in this study, the lowest κ statistics were achieved for the revised Duparc classification among classifications studied especially when XR was used. The likely reason for this fair agreement is that the revised Duparc classification is the most comprehensive classification including five main types and 16 subtypes with associated fractures. Particularly, the spinocondylar fractures are hard to be classified, which represent relatively more complicated injuries on XRs. Further evaluation with sequential sagittal three-dimensional (3D)-CT images strengthened the level of agreement up to other classification systems making this classification significantly more reproducible. This is a finding of importance to the practicing surgeon. However, to recognize an isolated posteromedial fracture with associated injury patterns or spinocondylar fractures by XRs are difficult albeit 3D-CT evaluation for this classification did not add any superiority among other classifications. In addition to this, complex plateau fractures that can be described as Schatzker types V and VI, AO-OTA type C, revised Duparc bicondylar, spinocondylar or posteromedial fractures or Hohl and Moore type V fractures necessitate special attention when management with dual plating. Recently, Luo et al.[5] described three column fixations for these kinds of fractures and Mellema et al.[9] compared Luo classification with Schatzker. Although they found fair agreement between observers for Schatzker and Luo classifications, we found moderate interobserver and substantial intraobserver agreement for Luo classification.

Kappa statistics are most widely available in the published data for the Schatzker classification system. Some variability exists among published studies. Our study yielded comparable interobserver (κ_{XR}=0.51; κ_{CT}=0.61) and intraobserver (κ_{XR}= 0.58; κ_{CT}=0.66) results to those of Brunner et al.[10] (κ_{XR}=0.418; κ_{CT}=0.755). This was slightly better than the quoted mean κ_{CT} of 0.61 from several other studies[8,11,12] but worse than findings of Brunner et al.[10] and Hu et al.[13]

The clinical experience has been suggested to affect the interobserver reliability between observers with different level of experience;[10,14] however, our study findings contradicted with this, with no significant differences found among the observers despite differences in experience.

Overall, the Schatzker, AO-OTA and Hohl and Moore classification systems yielded similar κ statistics (0.45 to 0.51) with moderate level of agreement for interobserver reliability when using XR; however, only Schatzker classification reached substantial agreement level after screening by CT. Only the revised Duparc classification achieved fair agreement level by use of XR and the observers noted that its complexity compared with that of the other systems required more frequent rote application of instructions to make a classification and as a consequence took more time to apply than did the other systems. In addition, this system differs from Schatzker, Hohl and Moore and AO-OTA systems in which similar but basic anatomical and morphological characteristics of the fracture lines represent similar subtypes making them easy to memorize. Briefly stated, when one decides on the type of the Schatzker type, to decide on the subtype in other classification is easier than that of revised Duparc classification. With no improvement in reliability compared with other classification systems, and because of its complexity and the more time demanded to apply it, the revised Duparc classification did not find a widespread use in clinical practice or as a research tool.

All available classification systems for tibial plateau fractures are limited by their reliability and reproducibility, although in the present study, the Schatzker classification was more reliable than other systems if assessed using XRs as well as sagittal and coronal 3D-CT sections. Comminuted fracture configurations with multiple fracture lines induce a classification challenge albeit 3D-CT sections. In our practice, no classification system aids sufficiently in decision-making for treatment. This is because these systems do not provide detailed information for mechanism of injury, associated injuries or patient factors that may affect management.

A potential limitation of the present study was choosing basic fracture subtypes of classification systems for practical purposes. Another limitation was that only the intra- and interobserver reliability were assessed. No assessment was made of the validity, responsiveness, or internal consistency of the classification systems. Last, the number of the patients included in this study seems to be low; however, slightly higher than that of previously reported studies.

In conclusion, among the classification systems compared in this study, Schatzker was the one with the best agreement when CT was used. The AO-OTA and Hohl and Moore classifications presented similar kappa statistics with moderate level of agreement. The revised Duparc classification presented worse agreement among specialists, probably because of its complexity and different morphological subtypes. Especially in clinical context, no classification systems in this research are universally accepted and they raise questions for suitability of their usage. Development of a comprehensive but simple system that enables surgeon to choose the optimal treatment method as well as to obtain the best prognosis is required by future studies in orthopaedic field.

Acknowledgement

We thank Volkan Orhan, PhD, for his valuable contribution in statistical analysis.

Declaration of conflicting interests

The authors declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding

The authors received no financial support for the research and/or authorship of this article.

REFERENCES

1. Elsoe R, Larsen P, Nielsen NP, Swenne J, Rasmussen S, Ostgaard SE. Population-Based Epidemiology of Tibial Plateau Fractures. Orthopedics 2015;38:e780-6.
2. Müller ME, Nazarian S, Koch P, Schatzker J. The comprehensive classification of fractures in long bones. 1st ed. Berlin: Springer-Verlag; 1990. p. 148-91.
3. Schatzker J, McBroom R, Bruce D. The tibial plateau fracture. The Toronto experience 1968--1975. Clin Orthop Relat Res 1979;138:94-104.
4. Hohl M. Fractures of the proximal tibia and fibula. In: Rockwood C, Green D, Buckolz R, editors. Fractures In Adults. 3rd ed. Philadelphia: J.B Lippincott; 1991. p. 1725-61.
5. Luo CF, Sun H, Zhang B, Zeng BF. Three-column fixation for complex tibial plateau fractures. J Orthop Trauma 2010;24:683-92.
6. Gicquel T, Najihi N, Vendeuvre T, Teyssedou S, Gayet LE, Huten D. Tibial plateau fractures: Reproducibility of three classifications (Schatzker, AO, Duparc) and a revised Duparc classification. Orthop Traumatol Surg Res 2013;99:805-16.
7. Patange Subba Rao SP, Lewis J, Haddad Z, Paringe V, Mohanty K. Three-column classification and Schatzker classification: a three- and two-dimensional computed tomography characterisation and analysis of tibial plateau fractures. Eur J Orthop Surg Traumatol 2014;24:1263-70.
8. Atik OŞ. Do we surgeons perform surgery only? Eklem Hastalik Cerrahisi 2016;27:123-4.
9. Mellema JJ, Doornberg JN, Molenaars RJ, Ring D, Kloen P. Interobserver reliability of the Schatzker and Luo classification systems for tibial plateau fractures. Injury 2016;47:944-9.
10. Brunner A, Horisberger M, Ulmar B, Hoffmann A, Babst R. Classification systems for tibial plateau fractures; does computed tomography scanning improve their reliability? Injury 2010;41:173-8.
11. Doornberg JN, Rademakers MV, van den Bekerom MP, Kerkhoffs GM, Ahn J, Steller EP, et al. Two-dimensional and three-dimensional computed tomography for the classification and characterisation of tibial plateau fractures. Injury 2011;42:1416-25.
12. te Stroet MA, Holla M, Biert J, van Kampen A. The value of a CT scan compared to plain radiographs for the classification and treatment plan in tibial plateau fractures. Emerg Radiol 2011;18:279-83.
13. Hu YL, Ye FG, Ji AY, Qiao GX, Liu HF. Three-dimensional computed tomography imaging increases the reliability of classification systems for tibial plateau fractures. Injury 2009;40:1282-5.
14. Rasmussen S, Madsen PV, Bennicke K. Observer variation in the Lauge-Hansen classification of ankle fractures. Precision improved by instruction. Acta Orthop Scand 1993;64:693-4.