Acta Orthopaedica et Traumatologica Turcica 50 (2016) 601–605

Contents lists available at ScienceDirect

Acta Orthopaedica et Traumatologica Turcica

journal homepage: https://www.elsevier.com/locate/aott

Is surgeons' experience important on intra- and inter-observer reliability of classifications used for adult femoral neck fracture?

Ali Turgut*, Mert Kumbaracı, Önder Kalenderer, Gökhan İlyas, Tayfun Bacaksız, Levent Karapınar

Tepecik Eğitim ve Araştırma Hastanesi, İzmir, Turkey

ARTICLE INFO

Article history:
Received 11 May 2015
Received in revised form
15 September 2015
Accepted 11 November 2015
Available online 23 November 2016

Keywords:
Femoral neck fracture
Classifications
Reliability
Experience

ABSTRACT

Purpose: To evaluate whether surgeons' experience affect inter- and intra-observer reliability among mostly used classification systems for femoral neck fractures.

Material and methods: A power point presentation was prepared with 107 slides which were anteroposterior radiographs of each femoral neck fracture. Five residents, 5 orthopaedic surgeons and 5 senior orthopaedic surgeons reviewed this presentation and classified the fractures according to Garden, Pauwels and AO classifications. The order of the slides was changed and reviews were repeated after 3 months. Fleiss kappa and intraclass correlation coefficient values were calculated to evaluate inter and intra-observer reliability.

Results: Garden and AO classifications' inter-observer reliabilities were similar and higher than Pauwels classification. Among three experience groups, the inter-observer reliability for Garden classification was highest in senior surgeon group, the interobserver reliability for AO classification was highest in surgeon group, and interobserver reliability of Pauwels classification was highest in low experienced groups (residents and surgeons). Intra-observer reliability was highest for Garden and lowest for Pauwels classifications. Surgical experience was found to be not effective for intraobserver reliability.

Conclusion: Both Garden and AO classifications were more reliable than Pauwels classification. Surgical experience was not significantly important on these three classification systems' evaluation.

Level of Evidence: Level IV, Diagnostic study

Introduction

Femoral neck fractures (FNF) are common and these injuries are rarely managed nonoperatively. Surgical management options include fracture fixation, hemiarthroplasty, and total hip arthroplasty.[1] There are three common classification systems which are used to define FNF which are named as: Garden, Pauwels and Arbeitsgemeinschaft für Osteosynthesefragen (AO).[2–4]

Garden's is probably mostly used classification for FNF which categorizes this injury into four types based upon the displacement on the anteroposterior (AP) hip radiographs.[5] Pauwels et al had classified FNF according to the shearing angle of the fracture line.[3] They suggested that the more vertical the fracture line, the higher incidence of non-union. AO classification is based on fracture level, degree of displacement and the angle of the fracture line.[6]

Fracture classification systems have multiple purposes. They should facilitate common language between treating physicians and assist documentation and research. Classification systems for fractures are useful tools also for making a decision on an adequate method of treatment, on the proper selection of implants and for assessment of the treatment outcomes.[7] So high inter-observer and intra-observer reliability is necessity for useful classification systems. Classifying FNF properly is as important as other fractures. There are several reports which investigate these classification systems' inter- and intra-observer reliability.[6–12] To our knowledge, none of these studies evaluated all of classifications which were mentioned above at the same fracture radiograph also by evaluating the surgeons' experience. The purpose of this study was to evaluate which classification system had better inter or intra-

* Corresponding author.
E-mail address: draliturgutort@yahoo.com.tr (A. Turgut).

Peer review under responsibility of Turkish Association of Orthopaedics and Traumatology.

http://dx.doi.org/10.1016/j.aott.2015.11.004

observer reliability and also whether surgeons' experience affect these reliabilities among mostly used classification systems for FNF.

Materials and methods

Between January 2009 and June 2014, 107 patients (64 male, 43 female) who had acute FNF were treated by closed reduction and percutaneous fixation with 3 cannulated screws. Preoperatively taken antero-posterior (AP) radiographs of fractured hip were saved as "Joint Photographic Experts Group (JPEG)" format from medical record electronic system (Probel®) of the hospital. A Microsoft® Office PowerPoint presentation (PPT) had been prepared by a surgeon who was not included as an observer in the study with JPEG images of each patient. There were totally 107 slides. Totally 15 observers were divided into three groups according to their experiences. Group 1 was consisted of 5 residents who were training more than 2 years in orthopaedics, group 2 was consisted of 5 orthopaedic trauma surgeons in their 2 or 3 years of experience and finally group 3 was consisted of 5 senior orthopaedic trauma surgeons who had more than 7 years of experience. Observers were asked to classify the fractures according to Garden, Pauwels, and AO classification systems. Garden classification was used as 4 grades, Pauwels classification as 3 grades, and AO classification as 3 grades (B1: non displaced to minimally displaced subcapital fracture, B2: transservical fracture through the middle or base of the neck, B3: displaced-non impacted subcapital fracture).

Explanations and figures of the classification systems were distributed to all of the observers which were quoted from a well known textbook of orthopaedic trauma.[13] All of the observers reviewed the PPT individually. After 3 months, order of the PPT slides was randomly changed. New order of the slides was recorded to allow to evaluate intraobserver reliability. Observers were asked to classify the fractures again. At each review 2 weeks were permitted for observers. The results of the second reviews were edited according to the first reviews to enable comparison between observers and groups.

SPSS 17 version for Windows was used for statistical analysis. Percentage agreement and intra-observer reliability was calculated using intraclass correlation coefficient (ICC).[14,15] The ICC values were given with 95% confidence intervals. An ICC value greater than 0.75 was considered as excellent agreement, 0.40 to 0.75 was fair to good and below 0.40 was poor. Inter-observer reliability was evaluated by calculating Fleiss Kappa with the use of Microsoft® Office Excel.[16]

Results

Mean age of the patients was 47.3 ± 11.2 years (18–68). AO classification's Fleiss kappa value was markedly higher than Garden and Pauwels classifications for the first review (0.430, 0.338 and 0.246, respectively). Fleiss kappa values of AO and Garden classifications were similar for the second review (0.342 and 0.344, respectively) (Fig. 1, Table 1). Fleiss kappa values were highest for senior surgeons both in first and second reviews of Garden classification, highest for surgeons both in first and second reviews of Pauwels classification and highest for residents in first review of AO classification and highest for senior surgeons in second review of AO classification (Fig. 1). Overall and the groups' Fleiss kappa values are shown in Table 1. Intra-observer reliability was found to be higher for Garden classification than AO and Pauwels classifications both in each experience groups and overally (overally; 0.759, 0.617 and 0.460, respectively) (Tables 2 and 3).

Discussion

An ideal classification system has to be valid, reliable and reproducible. It should facilitate a standard language for both practicing surgeons and researchers.[17,18] There are several studies which had investigated the inter- and intra-observer reliability of classification systems' for FNF.[6–12] Only one of these studies evaluated all of three classification systems (Garden, Pauwels, and AO) on the same radiographs without assessing the effect of surgical experience.[7] In this current study, these 3 classification systems' intra- and inter-observer reliabilities and also the affect of surgical experience on these reliabilities were evaluated. The major finding of this study was that Garden and AO classifications' inter-observer reliabilities were similar and higher than Pauwels classification. Among three experience groups, the inter-observer reliability for Garden classification was highest in senior surgeon group, the inter-observer reliability for AO classification was highest in surgeon group, and inter-observer reliability of Pauwels classification

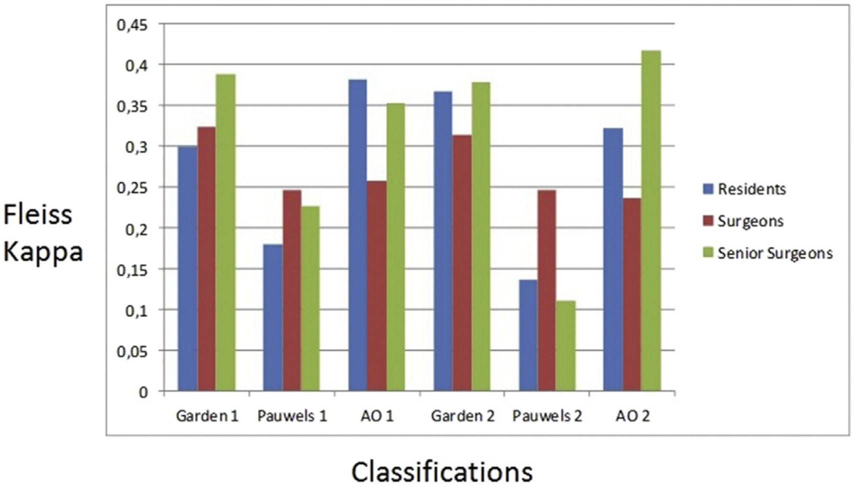

Fig. 1. Inter-observer reliabilities of classifications for both two reviews.

Table 1
Fleiss κ values for inter-observer reliability overally and for each groups of experience and classifications.

	Overall Fleiss Kappa (min–max)	Residents' Fleiss Kappa (min–max)	Surgeons' Fleiss Kappa (min–max)	Senior urgeons' Fleiss Kappa (min–max)
Garden Review 1	0,338 (0,327–0349)	0,299 (0,262–0336)	0,324 (0,287–0360)	0,387 (0,352–0423)
Garden Review 2	0,344 (0,333–0355)	0,367 (0,331–0403)	0,313 (0,276–0350)	0,378 (0,341–0414)
Pauwels Review 1	0,246 (0,231–0261)	0,180 (0,131–0228)	0.230 (0.181–0.280)	0,226 (0,180–0272)
Pauwels Review 2	0,183 (0,168–0197)	0,136 (0,087–0185)	0,245 (0,199–0291)	0,110 (0,063–0156)
AO Review 1	0,430 (0,401–0459)	0,381 (0,337–0424)	0,258 (0,214–0302)	0,352 (0,309–0396)
AO Review 2	0,342 (0,328–0355)	0,322 (0,278–0366)	0,236 (0,192–0280)	0,416 (0,371–0462)

Table 2
κ values for intra-observer reliability.

	Garden Classification	Pauwels Classification	AO Classification
Resident 1	0.790 (0.707–0.852)	0.694 (0.581–0.780)	0.752 (0.657–0.824)
Resident 2	0.815 (0.740–0.870)	0.633 (0.505–0.734)	0.704 (0.594–0.788)
Resident 3	0.690 (0.576–0.777)	0.502 (0.346–0.631)	0.598 (0.461–0.707)
Resident 4	0.598 (0.462–0.707)	0.332 (0.153–0.490)	0.660 (0.538–0.754)
Resident 5	0.869 (0.813–0.908)	0.191 (0.003–0.367)	0.732 (0.630–0.809)
Mean for Resident Group	0.752	0.470	0.689
Surgeon 1	0.736 (0.636–0.812)	0.158 (−0.032–0.337)	0.038 (−0.152–0.225)
Surgeon 2	0.853 (0.792–0.897)	0.745 (0.647–0.818)	0.830 (0.760–0.881)
Surgeon 3	0.706 (0.597–0.789)	0.600 (0.464–0.709)	0.791 (0.707–0.852)
Surgeon 4	0.782 (0.696–0.846)	0.704 (0.595–0.788)	0.605 (0.470–0.712)
Surgeon 5	0.770 (0.680–0.837)	0.190 (0.002–0.366)	0.399 (0.227–0.546)
Mean for Surgeon Group	0.769	0.479	0.532
Senior Surgeon 1	0.823 (0.750–0.875)	0.427 (0.260–0.570)	0.710 (0.602–0.793)
Senior Surgeon 2	0.730 (0.628–0.808)	0.227 (0.040–0.398)	0.655 (0.532–0.751)
Senior Surgeon 3	0.761 (0.668–0.830)	0.528 (0.377–0.652)	0.774 (0.685–0.840)
Senior Surgeon 4	0.810 (0.734–0.867)	0.584 (0.445–0.696)	0.749 (0.652–0.821)
Senior Surgeon 5	0.655 (0.532–0.751)	0.389 (0.216–0.538)	0.266 (0.082–0.433)
Mean for Senior Surgeon Group	0.756	0.431	0.631
For All Groups	0.759	0.460	0.617

Table 3
Intra-observer reliability of classifications for each group.

	Garden			Pauwels			AO		
	Excellent (κ:0.75–1)	Fair to Good (κ:0.40–0.75)	Poor (κ:<0.40)	Excellent (κ:0.75–1)	Fair to Good (κ:0.40–0.75)	Poor	Excellent (κ:0.75–1)	Fair to Good (κ:0.40–0.75)	Poor (κ:<0.40)
Residents	3	2	0	0	3	2	1	4	0
Surgeons	3	2	0	0	3	2	2	1	2
Senior Surgeons	3	2	0	0	3	2	1	3	1

was highest in low experienced groups (residents and surgeons). Intra-observer reliability was highest for Garden and lowest for Pauwels classifications. Surgical experience was not found to be important for intra-observer reliability.

Garden classification is the most commonly used system for FNF.[8] Frandsen et al stated that eight observers were agreed on Garden stage in 22% of fractures, they also reported that the ability to delineate the different stages was poor and Garden classification was superior to Pauwels classification for FNF.[10] In a recent study, Van Embden et al found that the inter-observer reliability of four staged Garden classification was lower than simplified (two stage: non displaced and displaced) classification.[9] Kappa (κ) values were 0.31 and 0.52, respectively.[9] They also stated that intra- and inter-observer reliabilities were similar for both surgeons and residents. Zlowodski et al studied perception of Garden classification with 298 orthopaedic trauma surgeons.[8] They stated that simplified Garden classification should be useful to increase inter-observer reliability as other studies did so.[9,19,20] Thomsen et al found a poor inter-observer agreement (κ:0.39–0.40), while Oakes et al reported a moderate inter-observer agreement (κ:0.42).[21,22] In our study, inter-observer reliability of Garden classification was 0.34 (Fleiss κ) for both two reviews overally (Table 1). Most experienced group's κ values were 0.38 while least experienced groups' were 0.33 on average. κ value for intra-observer reliability was meanly 0.76 (excellent). Our results were compatible with the literature. Although the most experienced group's κ values were highest, we could not state that experience is important for inter-observer reliability of Garden classification. Although our results showed that Garden Classification's inter-observer reliability was highest, it doesn't mean that this classification system is enough to allow surgeons to talk the same language about these fractures yet the overall κ values are not high enough. These findings support the usefulness of simplified version of this classification system since inter-observer reliability of four staged system is quite low.

The Pauwels classification was firstly defined in 1935.[3] Although comparative studies evaluating Garden and Pauwels classifications were available in the literature, inter-observer reliability was studied for the first time in 2011.[11] In the mentioned study, inter-observer agreement was found to be κ:0.31 for all observers, 0.38 for the surgeons and 0.27 for the residents, intra-observer reliability results were not given. In our study; inter-observer agreement of whole observers' was found to be κ:0.24 for the first review and κ:0.18 for the second review of Pauwels classification. κ value for intra-observer reliability was meanly 0.46 (fair to good). Highest κ values were found in the middle experienced group for both two reviews (0.23 and 0.24). According to these data, we could not state that experience is important for inter-observer reliability of Pauwels classification. The reason of such low values about inter- and intra-observer reliability of Pauwels classification system is probably due to difficulty in decision making on preoperatively taken radiographs. Measuring the Pauwels angle can vary on preoperatively taken radiographs because imagined angle between the horizontal and fracture lines can be altered according to the position of the femoral shaft (varus, valgus, internal or external rotation).[23] Since the fractures with more shear angles can result in an increased possibility of loss of reduction, it can be better to classify FNF's in the operating room on fluoroscopic view after performing the reduction. One can than decide how to fix the fracture. In a very recent study Wang et al[23] defined a modified method of measuring the shearing angle of FNF. It was stated that, because of the difficulty in measuring a true line above the femoral head, the angle between the imaginary horizontal line over the femoral neck and the fracture line will vary if the position of the patient changes. The authors used the central line of the femoral shaft as an assisted parameter for the measurement of the shearing angle and as a result the intra- and inter-observer reliability of Pauwels classification was found to be markedly higher than the traditional measurement method (κ:0.32 and 0.68) for intra-observer reliability, κ:0.18 and 0.63 for inter-observer reliability.

The most new one of the FNF classification systems is AO classification.[7] Blundell et al had found κ:0.29 meanly for inter-observer reliability and κ:0.5 meanly for intra-observer reliability[6] when they had simplified the AO classification as undisplaced (B1.1, B1.2, B1.3), basal (B2.1), and displaced (B2.2, B2.3, B3.1, B3.2, B3.3). The κ values found to be significantly increased (meanly inter-observer κ:0.74 and meanly intra-observer κ:0.85). In our study, we found κ:0.43 and 0.34 for inter-observer reliability for the first and second reviews. Intra-observer reliability was found to be fair to good both for overally (κ:0.617) and for all each experience groups (κ:0.689, 0.532 and 0.631 respectively) (Tables 2 and 3). Relatively better values of AO classification's inter-observer reliability can be due to not considering subgroups in our study group. Our results showed similar intra- and inter-observer reliability values as mentioned in the literature for group AO classification (not subgroup).[7] We could not state that experience influences intra- and inter-observer reliability of AO classification for FNF according to our results.

There is only one study that evaluated all three classifications' intra- and inter-observer reliability for FNF on the same patient group which was performed by Gašpar et al.[7] Subgroup AO classification was evaluated additionally in this study but influence of experience on these classifications' reliability and reproducibility was not evaluated. According to Gašpar et al, inter-observer and intra-observer κ values were; 0.41–0.49, 0.19–0.38 and 0.44–0.56 for Garden, Pauwels and group AO classifications respectively.[7] In our study, κ values for the inter-observer reliability were found as 0.34, 0.25, and 0.43 for these classifications respectively in the first review. In the second review, these values were 0.34, 0.18, and 0.34. κ values for intra-observer reliability were found as 0.76, 0.46 and 0.62 respectively. The study of Gašpar et al and our current study state that both Garden and group AO classifications are reliable than Pauwels classification. Our study also shows that surgical experience is not significantly important for evaluation of these three classification systems. There are some differences between these two studies. There are more fractures and observers in our study compared to the study of Gašpar et al (107 fractures and fifteen observers in our study and 77 fractures and 5 observers in Gašpar et al's study). We think that as the number of observers and fractures increase, the validity and reproducibility of the classifications can vary. This conception is quite valid not only for this study, also for all other studies about FNF classifications because there are lower than 10 observers in most of them.

This study has several limitations. First of all, simplified Garden and subgroup AO classifications were not evaluated. Surgeon who is more experienced in a specific classification system can not be objective. As mentioned before, Garden classification is most commonly used one as in our clinic so this reality could affect the results. The strengths of this study are both first and second reviews' results are given for inter-observer reliability, observer number is fifteen and the observers are in three experience groups with equal number. The radiographs of the patients were collected from digital database, so radiography qualities were good and finally observers reviewed the radiographs after three months in a different order.

As a conclusion we could state that both Garden and AO classifications are more reliable than Pauwels classification for FNF. It does not mean that these classifications can be named as successful because their inter-observer reliabilities were not found high enough although their intra-observer reliabilities were found to be better. We think that classifying FNF by using traditional Pauwels angle on preoperatively taken radiographs is unreliable and using it should be avoided, instead it may be more accurate if the measurements are performed on fluoroscopy views by the modified method which is defined by Wang et al.[23] Using three dimensional computed tomography images may provide higher intra- and inter observer reliability values but this method is expensive and it has radiation hazards to the patient. To our knowledge this study is the first one which evaluates surgical experience's affect on these classification systems validity and reproducibility. Our results demonstrate that surgical experience is not significantly important on these three classification systems' evaluation.

Conflict of interest

All of the authors state that there is no conflict of interest about them.

References

1. Jain NB, Losina E, Ward DM, Harris MB, Katz JN. Trends in surgical management of femoral neck fractures in the United States. *Clin Orthop Relat Res*. 2008;466(12):3116–3122.
2. Garden RS. Low-angle fixation in fractures of the femoral neck. *J Bone Jt Surg*. 1961;43:647–663.
3. Pauwels F. Einteilung der Schenkelhalsbruche nach mechanischen Gesichtspunkten. In: Pauwels F, ed. *Der Schenkelhalsbruch, ein mechanisches problem: Grundlagen des Heilungsvoraganges Prognose und Kausal Therapie*. Stuttgard: Ferdinand Enke Verlag; 1935:32–36.
4. Muller ME, Nazarian S, Koch P, Schatzker J. *The AO Classification of Fractures of Long Bones*. Berlin: Springer-Verlag; 1990.
5. Chen W, Li Z, Su Y, Hou Z, Zhang Q, Zhang Y. Garden type I fractures myth or reality? A prospective study comparing CT scans with X-ray findings in Garden type I femoral neck fractures. *Bone*. 2012;51(5):929–932.
6. Blundell CM, Parker MJ, Pryor GA, Hopkinson-Woolley J, Bhonsle SS. Assessment of the AO classification of intracapsular fractures of the proximal femur. *J Bone Jt Surg Br Vol*. 1998;80(4):679–683.
7. Gašpar D, Crnković T, Durović D, Podsednik D, Slišurić F. AO group, AO subgroup, Garden and Pauwels classification systems of femoral neck fractures: are they reliable and reproducible? Medicinski glasnik: official publication of the Medical Association of Zenica-Doboj Canton. *Bosnia Herzeg*. 2012;9(2): 243–247.
8. Zlowodzki M, Bhandari M, Keel M, Hanson BP, Schemitsch E. Perception of Garden's classification for femoral neck fractures: an international survey of 298 orthopaedic trauma surgeons. *Arch Orthop Trauma Surg*. 2005;125(7): 503–505.
9. Van Embden D, Rhemrev SJ, Genelin F, Meylaerts SAG, Roukema GR. The reliability of a simplified Garden classification for intracapsular hip fractures. *Orthop Traumatol Surg Res*. 2012;98(4):405–408.

10. Frandsen PA, Andersen E, Madsen F, Skjodt T. Garden's classification of femoral neck fractures. An assessment of inter-observer variation. Journal of Bone & Joint Surgery. *Br Vol*. 1988;70(4):588–590.
11. Van Embden D, Roukema GR, Rhemrev SJ, Genelin F, Meylaerts SA. The Pauwels classification for intracapsular hip fractures: is it reliable? *Injury*. 2011;42(11): 1238–1240.
12. Parker MJ, Dynan Y. Is Pauwels classification still valid? *Injury*. 1998;29(7): 521–523.
13. Leighton Ross K. Fractures of the neck of the femur. In: Bucholz Robert W, Heckman James D, Court-Brown Charles M, eds. *Rockwood and Green's Fractures in Adults*. 6th ed. Lippincott Williams & Wilkins; 2006:1754–1791.
14. Fleiss JL, Cohen J. The equivalence of weighted kappa and intraclass correlation coefficient as measures of reliability. *Educ Psychol Meas*. 1973;33:613–619.
15. Landis JR, Koch GG. The measurement of observer agreement for categorial data. *Biometrics*. 1977;33:159–174.
16. Fleiss JL. *Statistical Methods for Rates and Proportions*. 2nd ed. New York: John Wiley and Sons; 1981.
17. Belloti JC, Tamaoki MJ, Franciozi CE, et al. Are distal radius fracture classifications reproducible? Intra and interobserver agreement. *Sao Paulo Med J*. 2008;126(3):180–185.
18. Karanicolas PJ, Bhandari M, Kreder H, et al, Collaboration for Outcome Assessment in Surgical Trials (COAST) Musculoskeletal Group. Evaluating agreement: conducting a reliability study. *J Bone Jt Surg Am*. 2009;91(suppl 3): 99–106.
19. Beimers L, Kreder HJ, Berry GK, et al. Subcapital hip fractures: the Garden classification should be replaced, not collapsed. *Can J Surg*. 2002;45:411–414.
20. Scavenius M, Ibsen A, Ronnebech J, Aagaard H. Inter- and intra-observer variation in the assessment of femoral neck fractures according the Garden classification. *Acta Orthop Scand*. 1996;67(suppl 267):29.
21. Thomsen NO, Jensen CM, Skovgaard N, et al. Observer variation in the radiographic classification of fractures of the neck of the femur using Garden's system. *Int Orthop*. 1996;20(5):326–329.
22. Oakes DA, Jackson KR, Davies MR, et al. The impact of the Garden classification on proposed operative treatment. *Clin Orthop Relat Res*. 2003;409:232–240.
23. Wang SH, Yang JJ, Shen HC, Lin LC, Lee MS, Pan RY. Using a modified Pauwels method to predict the outcome of femoral neck fracture in relatively young patients. *Injury*. Oct 2015;46(10):1969–1974.