International Orthopaedics (SICOT) (2016) 40:1981–1985
DOI 10.1007/s00264-016-3236-x

ORIGINAL PAPER

Post-traumatic subtalar osteoarthritis: which grading system should we use?

Robert-Jan O. de Muinck Keizer [1] · **Manouk Backes** [1] · **Siem A. Dingemans** [1] · **J. Carel Goslings** [1] · **Tim Schepers** [1]

Received: 31 December 2015 / Accepted: 27 May 2016 / Published online: 24 June 2016

Abstract

Purpose To assess and compare post-traumatic osteoarthritis following intra-articular calcaneal fractures, one must have a reliable grading system that consistently grades the post-traumatic changes of the joint. A reliable grading system aids in the communication between treating physicians and improves the interpretation of research. To date, there is no consensus on what grading system to use in the evaluation of post-traumatic subtalar osteoarthritis. The objective of this study was to determine and compare the inter- and intra-rater reliability of two grading systems for post-traumatic subtalar osteoarthritis.

Methods Four observers evaluated 50 calcaneal fractures at least one year after trauma on conventional oblique lateral, internally and externally rotated views, and graded post-traumatic subtalar osteoarthritis using the Kellgren and Lawrence Grading Scale (KLGS) and the Paley Grading System (PGS). Inter- and intra-rater reliability were calculated and compared.

Results The inter-rater reliability showed an intra-class correlation (ICC) of 0.54 (95 % CI 0.40-0.67) for the KLGS and an ICC of 0.41 (95 % CI 0.26 – 0.57) for the PGS. This difference was not statistically significant. The intra-rater reliability showed a mean weighted kappa of 0.62 for both the KLGS and the PGS.

Conclusion There is no statistically significant difference in reliability between the Kellgren and Lawrence Grading System (KLGS) and the Paley Grading System (PGS). The PGS allows for an easy two-step approach making it easy for everyday clinical purposes. For research purposes however, the more detailed and widely used KLGS seems preferable.

Keywords Calcaneus · Classification · Hindfoot · Posttraumatic osteoarthritis · Subtalar joint

Introduction

Displaced intra-articular calcaneal fractures are complex injuries which can lead to longstanding disability. These fractures are notorious for the development of symptomatic osteoarthritis (OA) of the subtalar joint due to post-traumatic intra-articular incongruency [1–3]. Although the calcaneus involves multiple joints, it is mostly the subtalar joint in which OA causes problems [3].

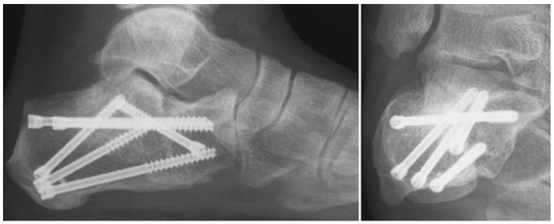
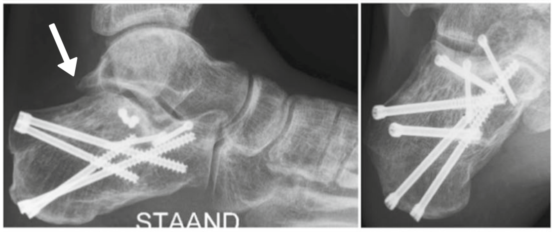
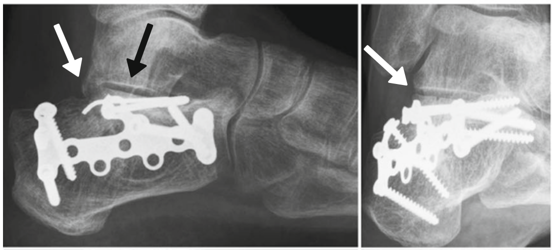
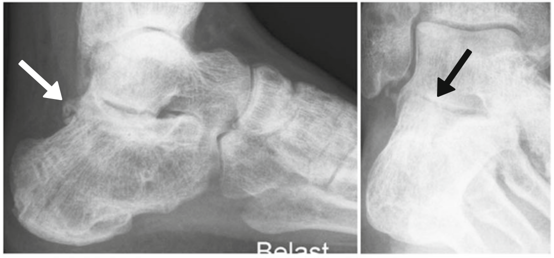
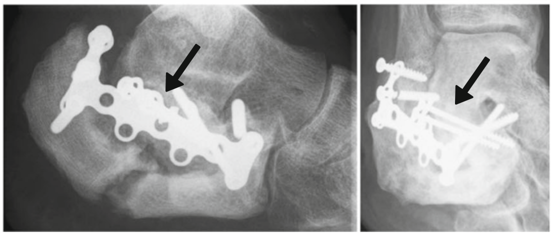
The treatment of intra-articular calcaneal fractures remains subject to discussion [4–8]. In order to adequately assess and compare the different treatment options, one must have a reliable radiological grading system that consistently grades the post-traumatic changes of the joint. Up till now, it is unclear which radiological grading system is best for evaluating post-traumatic subtalar OA. To our knowledge, there is only one systematic review that evaluates the methods of grading foot OA [9]. This study showed that 70 % of studies describing OA in all foot and hindfoot joints use the Kellgren and Lawrence Grading System (KLGS).

The KLGS was originally introduced in 1957 for the evaluation of OA of the hand, wrist, spine, hip, and knee joint [10]. It is a grading scale that reaches from 0 (no radiographic findings of osteoarthritis) to 4 (definite osteophytes with severe joint space narrowing and subchondral sclerosis) (Table 1) [10]. A recent study evaluated the inter- and intra-rater reliability of the system for the subtalar joint in patients following total ankle replacement and found a moderate inter- and intra-

✉ Robert-Jan O. de Muinck Keizer
rjodemuinckkeizer@amc.nl

[1] Trauma Unit, Department of Surgery, Academic Medical Center, Amsterdam, The Netherlands

Table 1 Kellgren and Lawrence (KL) grading system and Paley (P) grading system

KL-0. No radiographic findings of osteoarthritis		**P-0.** A normal joint space, with no evidence of degenerative cysts or subchondral sclerosis
KL-1. Minute osteophytes of doubtful clinical significance (white arrow)		-
KL-2. Definite osteophytes (white arrow) with unimpaired joint space		**P-1.** Subchondral sclerosis (black arrow), osteophytes (white arrow) and cyst formation, without narrowing of the joint space
KL-3. Definite osteophytes (white arrow) with moderate joint space narrowing (black arrow)		**P-2.** Narrowing of the joint space (black arrow), with sclerosis and cyst formation
KL-4. Definite osteophytes with severe joint space narrowing (black arrow) and subchondral sclerosis		**P-3.** Complete loss of the joint space (black arrow)

rater agreement at best (K = 0.37 and K = 0.43 respectively) [11]. Despite its widespread use, the KLGS has never been validated for use in the evaluation of post-traumatic OA of the subtalar joint.

Other systems that assess arthritic changes of the subtalar joint include systems that were developed for cadaveric studies (Drayer-Verhagen) [12], rheumatoid arthritis (Larsen) [13], or use CT imaging to visualize post-traumatic changes (Ogut) [14]. One of the classifications that was specifically introduced to grade subtalar OA after calcaneal fractures, is the grading system by Paley and colleagues in 1993 [3]. This scale reaches from 0 (normal joint space) to 3 (complete destruction of joint space) (Table 1).

A reliable grading tool should not only benefit the assessment of OA in epidemiological and clinical studies, it should also improve the communication between involved clinicians. In order to reach this goal, a grading system needs to show a high inter- and intra-rater reliability. The purpose of this study was therefore to assess the inter- and intra-rater reliability of the most widely used grading system for post-traumatic osteoarthritis of the subtalar joint (Kellgren and Lawrence Grading System) and to compare it with a lesser-known system and less complex system (Paley Grading System).

Materials and methods

Between November 2010 and June 2014 102 patients (aged 18 to 75) with 104 displaced intra-articular calcaneal fractures (Sanders type II-IV) were managed with open reduction and internal fixation through either an extended lateral or sinus tarsi approach. As part of their participation in a large prospective trial (EF3X-trial) [15], these patients underwent radiographic evaluation of post-traumatic osteoarthritis (OA). Approval to use anonymized radiographs was given by the medical ethical board for the EF3X-trial and its successive studies.

A selection of 50 patients representing the full spectrum of OA severity were evaluated by means of one lateral, one internally (Brodén), and one externally rotated view of the subtalar joint. Radiographs were blinded for patient identifiers and numbered randomly. To minimize influence of the statistical challenge often referred to as the "kappa paradox", 50 cases were selected by an independent observer to represent the full spectrum of OA severity [16].

The presence and severity of post-traumatic OA was assessed by four observers: one experienced foot and ankle trauma surgeon and three MD, PhD fellows with calcaneal fractures as the main focus of their research. To reflect clinical practice, radiographs were reviewed on a standard PC monitor. Prior to classifying the OA, the two grading systems were explained to the observers. A reference sheet detailing the grading system was available throughout the task. A standardized data entry sheet was used to record the grading.

The initial read used the KLGS to grade the presence and severity of OA of the subtalar joint. After a minimum of five days, a second set of 25 cases was scored again to evaluate intra-rater variability. This process was repeated with the Paley Grading System. All observers were blinded to the ratings of the other observers.

Inter-rater reliability (IRR) was calculated using intra-class correlations (ICC). Higher ICC values indicate greater IRR, with an ICC estimate of 1 indicating perfect agreement and 0 indicating only random agreement. We used a two-way mixed, single-measures, consistency ICC, which is identical to a weighted kappa strategy but can be used for three or more raters [17]. Cut-offs were used as provided by Cicchetti et al., with IRR being poor for ICC values less than 0.40, fair for values between 0.40 and 0.59, good for values between 0.60 and 0.74, and excellent for values between 0.75 and 1.0 [18]. Inter-rater reliability was computed using IBM SPSS Statistics for Windows, version 22.0 (IBM Corp, Armonk, NY).

To compute intra-rater reliability, we used Lights' kappa strategy [19]. With this technique, we computed a (square) weighted kappa for both observers' sessions separately, yielding four different intra-rater weighted kappa's per grading system. We then used the arithmetic mean of these estimates to provide an overall index of agreement for each grading system. As this mean is in fact a weighted kappa, interpretation was based on the guidelines proposed by Landis and Koch: a kappa less than 0.00 indicates poor agreement, 0.00–0.20 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, and 0.81–1.00 almost perfect agreement [20]. Weighted kappa was computed using R Statistical Software (R-Project for Statistical Computing, Version 3.1.2, Package IRR, Vienna, Austria) followed by manually computing the arithmetic mean.

Results

Each of the four observers graded all the available radiographs. For both grading systems, it took approximately 30–45 minutes to grade the series of 50 sets.

The interrater reliability showed an ICC of 0.54 (95 % CI 0.40-0.67) for the KLGS and an ICC of 0.41 (95 % CI 0.26 – 0.57) for the Paley Grading System (Table 2). This difference was not statistically significant.

The intra-rater reliability showed a mean weighted kappa of 0.62 for both the KLGS and the Paley Grading System (Table 3).

Discussion

We found a fair inter-rater reliability for both the Kellgren and Lawrence (ICC 0.54) and the Paley Grading System (ICC 0.41). Intra-rater reliability was substantial for both systems (kappa 0.62 and 0.62 respectively). There was no statistically significant difference in reliability between the two systems. Although the average measures ICC is substantially higher than the single measures ICC (Table 2), this interpretation is reserved for clinical studies that use multiple observers, which is often not the case.

The lack of comparable studies makes it difficult to interpret our results in the light of existing literature. To our knowledge, this is the first study to assess reliability and compare these grading systems for posttraumatic subtalar joint OA. We did not find comparable studies that evaluate the Paley Grading System.

With regard to the Kellgren and Lawrence Grading System, we found higher reliability than Mayich and collegues, who assessed subtalar osteoarthritis after total ankle replacement and found weighted kappa's of 0.37 ± 0.06 (interrater) and 0.43 ± 0.07 (intrarater) [11]. This is surprising, as in contrast to secondary causes for subtalar osteoarthritis, the fractured subtalar joint is often incongruent and its view more often hampered by implants, potentially lowering reliability. To describe reliability they used both weighted kappa and Fleiss' kappa, which are limited in accommodating more than two observers and handling categorical data respectively. A more appropriate statistical analysis would perhaps have given different and more comparable results. Holzer and colleagues found higher reliability of the KLGS in post-traumatic ankle joints (inter-rater ICC 0.61 and intra-rater ICC 0.75) [21]. Moreover, Moon and colleagues evaluated post-traumatic OA of the ankle using the KLGS and found weighted kappa's of 0.58–0.80 (inter-rater) and 0.51-0.81 (intra-rater) [22]. The complex anatomy of the subtalar joint when compared to the

Table 2 Inter-rater reliability

	Kellgren and Lawrence Grading System ICC (95 % CI)	Paley Grading System ICC (95 % CI)	P-value
Single measures	0.54 (0.40-0.67)	0.41 (0.26 – 0.57)	NS
Average measures	0.82 (0.73 – 0.89)	0.74 (0.58 – 0.84)	NS

ICC: intraclass correlation. *CI*: confidence interval. *NS*: not significant

ankle joint might account for the slighty lower ICC for the KLGS we found in our study.

There are many ways to determine the degree of agreement amongst or within raters. Frequently agreement is reported by the percentage that raters agree in their ratings, often referred to as percentage agreement. However, this measure systematically overestimates the level of agreement by not correcting for agreement that would be expected by chance alone [17]. A more sophisticated analysis that corrects for this overestimation is the kappa-statistic [23]. Cohen's kappa is thought to be a robust measure for inter-rater agreement; however, it is not applicable to ordinal data and does not take into account the distance between two ratings. Cohen's *weighted* kappa can be used for data with an ordinal structure; it has the advantage that the further two raters are apart, the lower the IRR estimate will be [24]. It is limited however by the fact that it can only accommodate two raters. Fleiss' kappa is suitable for three or more raters, but is only available for nominal data and not suitable for fully crossed designs (were all subjects are rated by all raters) [25]. A final solution for larger numbers of raters is using Lights strategy, where kappa's are computed for all coder pairs and then uses the arithmetic mean of these estimates to provide an overall index of agreement [19]. A measure that is suitable for ordinal, interval, and ratio variables is the intraclass correlation (ICC). It is identical to a weighted kappa but has the advantage that it can handle more than two raters [26].

Strengths of this study include that we are the first to report on reliability of grading systems that evaluate post-traumatic OA of the subtalar joint specifically. Additionally, we have not only assessed inter-rater reliability but also evaluated reliability within raters. We used observers with different levels of experience in the assessment of calcaneal fractures in both clinical and research context. Earlier studies have shown that the level of experience of the observers, and the complexity of the classification system, do not usually affect inter-observer reliability [27]. Our study will help guide future researchers in their choice of grading system when reporting on post-traumatic subtalar osteoarthritis, and assist in the comparison of different treatment modalities for calcaneal fractures.

Table 3 Intra-rater reliability

	Kellgren and Lawrence Grading System Weighted kappa	Paley Grading System Weighted kappa
Rater 1	0.480	0.579
Rater 2	0.516	0.434
Rater 3	0.671	0.863
Rater 4	0.813	0.605
Lights' kappa	0.620	0.621

This study is limited in the number of grading systems it compares. However, many available systems are similar or poorly documented. Many systems resemble each other, grading osteophytes, subchondral sclerosis, and narrowing and disappearance of the joint space in various degrees. We chose to compare the most widely used (KLGS) and a lesser-known but more joint-specific and less complex system (PGS). We excluded systems that were not specifically used for the subtalar joint or were developed for cadaveric studies (Drayer-Verhagen) [12], rheumatoid arthritis (Larsen) [13], or were CT-based (Ogut) [14]. An additional limitation is the fact that we did not have a gold standard available to determine the accuracy of both grading systems. To minimize the potential effect of the kappa paradox, we selected fractures with a wide spectrum of OA severity. In published cohorts however, the severity of osteoarthritis leans toward more severe osteoarthritis [28].

Our results suggest that there is no statistically significant difference in reliability between the Kellgren and Lawrence and the Paley Grading Systems. This leaves room for a comparison on different grounds. The Paley grading system describes subchondral sclerosis from grade 1 and higher, while the KLGS only describes this feature in the most severe grade 4. The KLGS leans heavily toward the presence of osteophytes and adds an extra grade to the system by classifying "osteophytes of doubtful clinical significance". While this might improve accuracy of the description of the state of the joint, it is indeed doubtful what its clinical relevance is and whether this justifies a more complex grading system. The Paley Grading System simply acknowledges the presence of 1) secondary characteristics (osteophytes, subchondral sclerosis, and cyst formation) and 2) joint space narrowing, allowing for a two-step approach when grading OA. Since the Paley Grading System is non-inferior to the Kellgren and Lawrence Grading system and less complex to comprehend, this could be a reason to use the Paley system in future clinical settings. However, when it comes to comparing different treatment modalities in research, a more detailed and widely used system (i.e., KLGS) would be more convenient.

Conclusion

Both the Kellgren and Lawrence Grading System (KLGS) and the Paley Grading System (PGS) have a fair inter-rater reliability. Intra-rater reliability is substantial for both systems. There is no statistically significant difference in reliability between the KLGS and the PGS.

Acknowledgments We would like to thank S. van Dieren (PhD, Clinical Epidemiologist) for her contributions to the statistical background of this paper.

Compliance with ethical standards

Conflicts of interest No benefits in any form have been received or will be received from a commercial party related directly or indirectly to the subject of this article. The authors received no financial support for the research, authorship, and/or publication of this article.

References

1. Franke J, Wendl K, Suda A et al (2014) Intraoperative three-dimensional imaging in the treatment of calcaneal fractures. J Bone Joint Surg 96:1–7
2. Thordanson D, Krieger L (1996) Operative vs. nonoperative treatment of intra-articular fractures of the calcaneus: a prospective randomized trial. Foot Ankle Int 17:2–9
3. Paley D, Hall H (1993) Intra-articular fractures of the calcaneus. A critical analysis of results and prognostic factors. J Bone Joint Surg Am 75:342–54
4. Bruce J, Sutherland A (2013) Surgical versus conservative interventions for displaced intra- articular calcaneal fractures (Review).
5. Griffin D, Parsons N, Shaw E et al (2014) Operative versus non-operative treatment for closed, displaced, intra-articular fractures of the calcaneus : randomised controlled trial. BMJ 4483:1–13. doi:10.1136/bmj.g4483
6. Rausch S, Klos K, Wolf U et al (2014) A biomechanical comparison of fixed angle locking compression plate osteosynthesis and cement augmented screw osteosynthesis in the management of intra articular calcaneal fractures. Int Orthop 38:1705–10. doi:10.1007/s00264-014-2334-x
7. Sampath Kumar V, Marimuthu K, Subramani S et al (2014) Prospective randomized trial comparing open reduction and internal fixation with minimally invasive reduction and percutaneous fixation in managing displaced intra-articular calcaneal fractures. Int Orthop 38:2505–12. doi:10.1007/s00264-014-2501-0
8. Simon P, Goldzak M, Eschler A, Mittlmeier T (2015) Reduction and internal fixation of displaced intra-articular calcaneal fractures with a locking nail: a prospective study of sixty nine cases. Int Orthop 39:2061–7. doi:10.1007/s00264-015-2816-5
9. Trivedi B, Marshall M, Belcher J, Roddy E (2010) A systematic review of radiographic definitions of foot osteoarthritis in population-based studies. Osteoarthr Cartil 18:1027–35. doi:10.1016/j.joca.2010.05.005
10. Kellgren JH, Lawrence JS (1957) Radiological assessment of osteo-arthrosis. Ann Rheum Dis 16:494–503
11. Mayich DJ, Pinsker E, Mayich MS et al (2013) An analysis of the use of the Kellgren and Lawrence grading system to evaluate peritalar arthritis following total ankle arthroplasty. Foot Ankle Int 34:1508–15. doi:10.1177/1071100713495379
12. Drayer-Verhagen F (1993) Arthritis of the subtalar joint associated with sustentaculum tali facet configuration. J Anat 183(Pt 3):631–4
13. Larsen A (1977) Radiographic evaluation of rheumatoid arthritis and related conditions by standard reference films. Acta Radiol Diagn 18:481–91
14. Ogut T, Ayhan E, Kantarci F et al (2011) Medial fracture line significance in calcaneus fracture. J Foot Ankle Surg 50:517–21. doi:10.1053/j.jfas.2011.04.018
15. Beerekamp MSH, Ubbink DT, Maas M et al (2011) Fracture surgery of the extremities with the intra-operative use of 3D-RX: a randomized multicenter trial (EF3X-trial). BMC Musculoskelet Disord 12:151. doi:10.1186/1471-2474-12-151
16. Warrens MJ (2010) A formal proof of a paradox associated with Cohen's kappa. J Classif 27:322–32. doi:10.1007/s00357-010-9060-x
17. Hallgren KA (2012) Computing inter-rater reliability for observational data: an overview and tutorial. Tutor Quant Methods Psychol 8:23–34. doi:10.1016/j.biotechadv.2011.08.021.Secreted
18. Cicchetti D (1994) Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess 6:284–90
19. Light R (1971) Measures of response agreement for qualitative data: some generalizations and alternatives. Psychol Bull 76:365–77
20. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–74
21. Holzer N, Salvo D, Marijnissen A (2015) Radiographic evaluation of posttraumatic osteoarthritis of the ankle: the Kellgren–Lawrence scale is reliable and correlates with clinical symptoms. Osteoarthr Cartil 23:363–9
22. Moon JS, Shim JC, Suh JS, Lee WC (2010) Radiographic predictability of cartilage damage in medial ankle osteoarthritis. Clin Orthop Relat Res 468:2188–97. doi:10.1007/s11999-010-1352-2
23. Viera AJ, Garrett JM (2005) Understanding interobserver agreement: the kappa statistic. Fam Med 37:360–3
24. Cohen J (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 70:213–20
25. Fleiss J (1971) Measuring nominal scale agreement among many raters. Psychol Bull 76:378–82
26. Norman G, Streiner D (2008) Biostatistics: the bare essentials. BC Decker, Ontario
27. Humphrey C, Dirschl D, Ellis T (2005) Interobserver reliability of a CT-based fracture classification. J Orthop Trauma 19:616–22
28. Ibrahim T, Rowsell M, Rennie W et al (2007) Displaced intra-articular calcaneal fractures: 15-Year follow-up of a randomised controlled trial of conservative versus operative treatment. Injury 38:848–55. doi:10.1016/j.injury.2007.01.003