Trauma

Journal of Orthopaedic Surgery

Journal of Orthopaedic Surgery
29(2) 1–7

DOI: 10.1177/23094990211010520
journals.sagepub.com/home/osj
SAGE

'Absolute' inter-observer classifications agreement for proximal humeral fractures with a single shoulder anteroposterior X-ray

Rocío Martínez-Sola[1], Vicente J León-Muñoz[2], Antoine Nicolas Najem-Rizk[1], Beatriz Soler-Vasco[1], Carlos J Arrieta-Martínez[1], Eva López-Sorroche[1], Encarnación Cárdenas-Grande[1], Guillermo Salmerón-Vélez[1], José Ángel Ruiz-Molina[1], Francisco Martínez-Martínez[2] and Fernando Santonja-Medina[2]

Abstract

Purpose: Several studies have been carried out, and there is no classification for proximal humeral fractures (PHF) exempted from variability in interpretation and with questioned reliability. In the present study, we investigated the 'absolute diagnostic reliability' of the most currently used classifications for PHFs on a single anterior-posterior X-ray shoulder image. **Methods:** Six orthopaedic surgeons, with varying levels of experience in shoulder pathology, evaluated radiographs from 30 proximal humeral fractures, according to the 'absolute reliability' criteria. Each of the observers rated each fracture according to Neer, Müller/AO and Codman-Hertel's classification systems. **Results:** The overall inter-observer agreement (κ) has been 0.297 (CI95% 0.280 to 0.314) for the Neer's classification system, 0.206 (CI95% 0.193 to 0.218) for the Müller/AO classification system, and 0.315 (CI95% 0.334 to 0.368) for the Codman-Hertel classification system. We found loss of agreement in Neer's classification as the study progressed, low agreement in the AO classification, and stable values in the different evaluations with the best degree of agreement for Codman-Hertel classification, with a moderate agreement in the second evaluation among the six evaluators. **Conclusion:** The Neer, AO, and Hertel-Codman classification systems for PHF with a single radiographic projection have a difficult interpretation for orthopaedic surgeons of varying levels of experience, and therefore substantial agreements are not obtained.

Keywords

classification, inter-observer and intra-observer reliability, proximal humeral fracture

Date received: 23 December 2020; Received revised 14 March 2021; accepted: 30 March 2021

Introduction

Proximal humeral fractures (PHFs) comprise 6% of all fractures in adults with an overall incidence of 73 fractures per 100,000 inhabitants and has significantly increased over the last decades (PHFs have tripled over the previous 30 years). The expectations continue to rise as the population ages.[1–7] Despite several publications (randomized trials,[8] observational studies,[9] and systematic reviews[10–12] of the PHFs have been published), it remains difficult to interpret their results, to perform prognostic studies, and to

[1] Department of Orthopaedic Surgery and Traumatology, Hospital Universitario Torrecárdenas, Almería, Spain
[2] Department of Orthopaedic Surgery and Traumatology, Hospital Clínico Universitario Virgen de la Arrixaca, El Palmar, Murcia, Spain

Corresponding author:
Vicente J León-Muñoz, Department of Orthopaedic Surgery and Traumatology, Hospital Clínico Universitario Virgen de la Arrixaca, Ctra. Madrid-Cartagena, s/n, 30120 El Palmar, Murcia, Spain.
Email: vleonmd@gmail.com

obtain consensus on treatment recommendations when concise definitions and a standard 'fracture language' are lacking.[9,13,14] Published results for treatment and evidence-based recommendations are inconclusive.[10–12] The most widely used PHFs classifications are the system created by Charles Neer 2nd in 1970,[15] updated in 2002,[16] the AO/OTA classification, based on the Müller classification from 1990[17] and updated in 2007,[18] and the Codman-Hertel binary fracture description system,[19] updated after the findings of low reliability in 1993 by Siebenrock and Gerber.[20]

Neer's classification system defines PHFs based on the number of fracture fragments (parts) and their displacement, which is defined as having a separation of more than 1 cm or an angle greater than or equal to 45 degrees.[16] The Müller/AO classification system was created for standardization and with defined terms of fracture description. Each bone and bone segment are classified into three categories (A, B, C), which are subdivided into three groups and each group into three subgroups. The type A fracture is extra-articular and unifocal, the type B is partially intra-articular and bifocal, and type C refers to intra-articular trace fractures, establishing 27 classification subgroups. Müller/AO classification system is more complicated than Neer's classification.[21] The Codman-Lego system was developed by Hertel et al. in 2004[19] and graphically represents the four parts of the proximal humerus (head, major and minor tuberosities, and diaphysis). The absence of a union between any of the four parts represents a fracture trace, making 12 different patterns possible, labelled with the numbers from 1 to 12. Thinking in terms of fracture planes rather than fracture fragments represented the paradigm shift based on the vascularization studies of the humeral head by Hertel et al.[19] and has the highest agreement rates. However, this system does not differentiate between varus and valgus displacement, which is crucial for the reduction and fixation of this type of fractures.[19]

Over the past decades, the reliability of the different classification systems has been questioned. Multiple studies with different imaging modalities (X-ray, computed tomography (CT) scan, and 3D reconstructions) have reported low agreement among observers when attempting to classify PHFs.[21–27] To the best of our knowledge, no study of the 'absolute reliability' of the PHF classifications with X-ray images has been published. 'Absolute reliability' is considered to be the analysis of a minimum of 30 cases, by at least six blinded observers and a minimum of three to five separate evaluations every 2 weeks in time by each observer.[28,29] Our study aimed to evaluate the absolute diagnostic reliability of the most currently used classifications for PHFs (Neer, AO, and Hertel) on a single anterior-posterior (AP) X-ray shoulder image among orthopaedic surgeons with different levels of experience. We hypothesize that there is significant variability in the reliability of the classifications described, with better results among more experienced evaluators.

Figure 1. Example of X-ray images provided to observers. An ID number was assigned to each X-ray, and any signs were removed from identification. As an example, this fracture was classified by most evaluators as IVA/IVB Neer, B1 AO, and 3/7 Hertel.

Materials and methods

We have prospectively analysed standard AP projection X-ray studies of patients between 50 and 80 years of age with PHF archived on the Picture Archiving and Communication System (PACS) of a secondary-level hospital. It was calculated (significance level of 5% and a power of 80%) that a sample size of 30 cases would be sufficient to detect a minimum variability of 10% between the groups of evaluators. X-rays of patients treated by PHF in a consecutive series over 1 year were evaluated. In some patients, nonoperative treatment was chosen and in others, surgical treatment. Out of 45 cases, the 15 worst quality images were discarded. We also excluded pathological fractures and previous fractures in the same location. Each of the 30 selected radiographs was assigned an ID number, and any signs were removed from identification (Figure 1). The images were randomly arranged for evaluation. No other projection or CT scan images were provided to the evaluators; the study's primary objective was to assess the absolute reliability of the three classifications studied on a single anterior-posterior radiograph. The study was approved by the Institutional Review Board and the Ethical Committee.

We designed a 'absolute' reliability study, according to Hopkins' criteria (a minimum of 30 cases assessed by a minimum of six assessors, with a minimum of three assessments and with a minimum interval between each assessment of 2 weeks).[29] Three evaluations of 30 AP shoulder X-ray were performed, each assessment separated by 1 month, with each of the six evaluators rating them according to the three systems (Neer, AO, and Codman-Hertel), independently and blindly. In each of the

re-evaluations, the order of the X-rays was changed to ensure the blinded evaluation.

All six evaluators were orthopaedic surgeons with varying levels of training and experience in shoulder pathology. The first group consisted of two shoulder surgery specialists with more than 10 years of experience (observer 1 and 2). The second group was made up of two orthopaedic consultants with over 10 years of experience who were not exclusively dedicated to shoulder pathology (observer 3 and 4). The third group was made up of two orthopaedic surgery resident physicians (postgraduate year-2) (observer 5 and 6). The selection of a minimum of six evaluators was a methodological requirement to analyse 'absolute reliability'. The experience profile followed conventional criteria whereby a senior surgeon is considered a surgeon with 10 or more years of experience. In a homogenization session, before the study start, the criteria for the Neer, AO, and Codman-Hertel classifications were reviewed with all six evaluators. In this training session, the three classifications' criteria and their differences were thoroughly reviewed with the evaluators. The necessary documentation was provided, and the evaluators could practice with examples of cases different from those in the study.

Statistical analysis was performed using the Statistical Package for the Social Sciences (SPSS), version 25 for Windows (SPSS, Inc., Chicago, Illinois, USA). We have used the kappa statistics to determine interrater reliability. Given the limitations of Cohen's kappa analysis (agreement measurement limited to two observers), we have also performed Fleiss' kappa, (κ) to determine the level of agreement between the observers of variables measured on a categorical scale.[30] We have reported the 95% confidence interval for Fleiss' kappa. We have assessed the level of agreement among observers according to the criteria by Landis and Koch (<0 indicate no agreement, 0.00 to 0.20 indicate slight agreement, 0.21 to 0.40 indicate fair agreement, 0.41 to 0.60 indicate moderate agreement, 0.61 to 0.80 indicate substantial agreement, and 0.81 to 1.0 indicate almost perfect or perfect agreement).[31]

Results

We recorded a total of 1620 observations among the six evaluators. The intra-observer agreement between the different evaluations is shown in Table 1.

The result of intra-observer variability has been erratic. However, we can observe the tendency to decrease the intra-observer agreement, as the time between one evaluation and the other increases, in Neer's classification and, on the contrary, the tendency to improve the degree of intra-observer agreement in the Codman-Hertel classification. The overall inter-observer agreement (Fleiss' kappa) has been 0.297 (CI95% 0.280 to 0.314) for the Neer's classification system, 0.206 (CI95% 0.193 to 0.218) for the Müller/AO classification system, and 0.315 (CI95% 0.334 to 0.368) for the Codman-Hertel classification system. Besides, when analysing the degree of agreement among the six evaluators in the three different evaluations with the Fleiss kappa statistician, we found loss of agreement in Neer's classification as the study progressed (0.383 for the first evaluation, 0.282 for the second, and 0.163 for the third), low agreement in the AO classification (0.196, 0.2, and 0.179), and stable values in the different evaluations with the best degree of agreement for Codman's classification, (0.239, 0.451, and 0.336), with the moderate agreement[30] in the second evaluation among the six evaluators. The differences in the agreement level among the more expert observers (observers 1 and 2) compared to the general agreement of the six evaluators in the three different evaluations are shown in Table 2. Table 3 shows the inter-observer agreement subdivided by fracture type.

Discussion

The extreme variability and complexity of PHF hinder a univocal definition of fracture patterns. There is quite a consensus on the difficulty of categorization of PHF according to different classification systems and in low reliability between and among observers on various imaging modalities.[20–27,32] The most important aspect that our study contributes to this topic is the methodological application of the 'absolute reliability' criteria proposed by Hopkins.[29] To the best of our knowledge, no study of the absolute reliability of the PHF classifications with X-ray images has been previously published.

Majed et al.[21] evaluated several classification systems (Neer, AO, Codman-Hertel and a prototype classification system by Resch et al.[33]) with three-dimensional printed models. They hypothesized that current PHF classification systems, regardless of imaging methods, are not sufficiently reliable to aid clinical management of these injuries. The κ coefficient values for the inter-observer reliability (four independent senior observers, experts in proximal humeral fracture management) of this study were 0.33 for Neer, 0.11 for AO, and 0.44 for Codman-Hertel classification system. Sukthankar et al.[34] assessed the intra-observer and inter-observer reliability of the Codman's description by Hertel et al.[19] and compared it with the AO and Neer systems. PHF were examined with anteroposterior, lateral, and axillary radiographs. The authors conclude that the Codman-Hertel classification system provided a more reliable description of proximal humeral fractures than the Neer and AO systems and they argue that this is due to the descriptive approach of Codman's system, which better defines the varieties of PHFs. In addition, they claim that the reliability of these systems can be improved by training in radiographic interpretation and correct measurement of fragment displacement.[34] Gracitelli et al.[24] aimed to evaluate the inter-observer and intra-observer reliability of different radiographic parameters, classifications, and surgical indication in PHFs among 10 orthopaedic surgeons with different levels of experience, who evaluated radiographs in three views from

Table 1. Intra-observer agreement between the different evaluations.

		Observer 1	Observer 2	Observer 3	Observer 4	Observer 5	Observer 6
Neer	1 & 2	0.686 (0.429 to 0.943)*	0.588 (0.367 to 0.809)*	0.595 (0.381 to 0.809)*	0.383 (0.150 to 0.616)*	0.298 (0.124 to 0.472) $p = 0.001$	0.731 (0.521 to 0.941)*
	1 & 3	0.517 (0.290 to 0.744)*	0.514 (0.298 to 0.730)*	0.529 (0.325 to 0.733)*	0.250 (0.046 to 0.454) $p = 0.006$	0.361 (0.124 to 0.472)*	0.386 (0.178 to 0.594)*
	2 & 3	0.585 (0.365 to 0.805)*	0.577 (0.326 to 0.828)*	0.411 (0.184 to 0.638)*	0.623 (0.409 to 0.837)*	0.820 (0.632 to 1.008)*	0.379 (0.159 to 0.599) $p = 0.001$
AO	1 & 2	0.501 (0.289 to 0.713)*	0.320 (0.108 to 0.532)*	0.388 (0.186 to 0.590)*	0.264 (0.072 to 0.456)*	0.365 (0.145 to 0.585)*	0.327 (0.149 to 0.505)*
	1 & 3	0.321 (0.094–0.548)*	0.366 (0.170–0.562)*	0.391 (0.193–0.589)*	0.147 (−0.018 to 0.312) $p = 0.033$	0.38 (0.164 to 0.596)*	0.499 (0.293 to 0.705)*
	2 & 3	0.286 (0.051 to 0.521) $p = 0.001$	0.293 (0.103 to 0.483)*	0.496 (0.296 to 0.696)*	0.537 (0.345 to 0.729)*	0.236 (−0.001 to 0.473) $p = 0.009$	0.375 (0.181 to 0.569)*
Hertel	1 & 2	0.335 (0.100 to 0.570)*	0.518 (0.285 to 0.751)*	0.536 (0.320 to 0.752)*	0.440 (0.242 to 0.638)*	0.468 (0.254 to 0.682)*	0.382 (0.172 to 0.592)*
	1 & 3	0.446 (0.223 to 0.669) $p = 0.002$	0.511 (0.301 to 0.721)*	0.494 (0.274 to 0.714)*	0.278 (0.098 to 0.458) $p = 0.001$	0.461 (0.241 to 0.681)*	0.364 (0.144 to 0.584)*
	2 & 3	0.662 (0.450 to 0.874)*	0.432 (0.216 to 0.648)*	0.695 (0.459 to 0.859)*	0.454 (0.209 to 0.699)*	0.823 (0.664 to 0.982)*	0.566 (0.350 to 0.782)*

Note: Results are shown as Cohen's kappa value (and 95% confidence interval). Neer: Neer's classification system. AO: Müller/AO classification system. Hertel: Codman-Hertel classification system. 1 & 2: Cohen's kappa value between the first and second evaluation. 1 & 3: Cohen's kappa value between the first and third evaluation. 2 & 3: Cohen's kappa value between the second and third evaluation. Observer 1 and 2: two shoulder surgery specialists. Observer 3 and 4: two experienced orthopaedic consultants. Observer 5 and 6: two orthopaedic surgery resident physicians (postgraduate year-2). Level of agreement according to the criteria by Landis and Koch[31] (<0 indicate no agreement, 0.00 to 0.20 indicate slight agreement, 0.21 to 0.40 indicate fair agreement, 0.41 to 0.60 indicate moderate agreement, 0.61 to 0.80 indicate substantial agreement, and 0.81 to 1.0 indicate almost perfect or perfect agreement).

*$p = 0.0005$.

Table 2. Inter-observer agreement among the six evaluators and among the most expert observers in the three different evaluations.

	First evaluation		Second evaluation		Third evaluation	
	Overall κ (95% CI)	Observers 1 and 2 κ (95% CI)	Overall κ (95% CI)	Observers 1 and 2 κ (95% CI)	Overall κ (95% CI)	Observers 1 and 2 κ (95% CI)
Neer	0.383 (0.330 to 0.437)	0.580 (0.378 to 0.782)	0.282 (0.224 to 0.340)	0.406 (0.222 to 0.590)	0.163 (0.109 to 0.216)	0.201 (−0.001 to 0.404)
AO	0.196 (0.156 to 0.236)	0.235 (0.063 to 0.407)	0.200 (0.160 to 0.240)	0.143 (0.008 to 0.278)	0.179 (0.140 to 0.218)	0.834 (0.664 to 1.005)
Hertel	0.239 (0.190 to 0.288)	0.407 (0.234 to 0.580)	0.451 (0.394 to 0.507)	0.361 (0.170 to 0.551)	0.336 (0.281 to 0.391)	0.334 (0.138 to 0.529)

Note: Results are shown as Fleiss kappa statistician (κ) (and 95% confidence interval). Neer: Neer's classification system. AO: Müller/AO classification system. Hertel: Codman-Hertel classification system. Level of agreement according to the criteria by Landis and Koch[31] (<0 indicate no agreement, 0.00 to 0.20 indicate slight agreement, 0.21 to 0.40 indicate fair agreement, 0.41 to 0.60 indicate moderate agreement, 0.61 to 0.80 indicate substantial agreement, and 0.81 to 1.0 indicate almost perfect or perfect agreement).

Table 3. Inter-observer agreement subdivided by fracture type.

Fracture type	Fleiss Kappa for Individual Categories	95% Confidence Interval
Neer group I	0.440	0.411 to 0.469
Neer group II	−0.004	−0.033 to 0.025
Neer group III	0.337	0.308 to 0.366
Neer group IV	0.341	0.312 to 0.370
Neer group V	0.129	0.100 to 0.157
AO A1	0.308	0.279 to 0.337
AO A2	0.258	0.229 to 0.287
AO A3	0.317	0.288 to 0.346
AO B1	0.128	0.099 to 0.157
AO B2	0.142	0.113 to 0.171
AO B3	−0.002	−0.031 to 0.027
AO C1	0.188	0.159 to 0.217
AO C2	0.192	0.163 to 0.221
AO C3	0.072	0.043 to 0.101
Hertel 1	0.454	0.425 to 0.483
Hertel 2	−0.004	−0.033 to 0.025
Hertel 3	0.199	0.170 to 0.227
Hertel 4	0.128	0.099 to 0.156
Hertel 5	−0.002	−0.031 to 0.027
Hertel 6	−0.002	−0.031 to 0.027
Hertel 7	0.329	0.300 to 0.358
Hertel 8	0.103	0.074 to 0.132
Hertel 9	−0.009	−0.038 to 0.020
Hertel 10	0.014	−0.015 to 0.043
Hertel 11	–	–
Hertel 12	0.420	0.391 to 0.449

Note: Level of agreement according to the criteria by Landis and Koch[31] (<0 indicate no agreement, 0.00 to 0.20 indicate slight agreement, 0.21 to 0.40 indicate fair agreement, 0.41 to 0.60 indicate moderate agreement, 0.61 to 0.80 indicate substantial agreement, and 0.81 to 1.0 indicate almost perfect or perfect agreement). In no case was the fracture classified as Codman-Hertel type 11.

40 PHF. They conclude that the pathomorphological classification[33] has higher reliability ($\kappa = 0.504$) than the Neer classification ($\kappa = 0.298$), and has been the factor that most influenced the surgical decision. Also, the results were influenced by the observer's experience. In the study published by LaMartina et al.,[11] three experienced shoulder surgeons agreed unanimously on treatment in only 51% of 274 cases. Furthermore, among the cases where the unanimous agreement was reached, only 63.5% of the patients underwent the selected treatment. The authors conclude that there will always be some degree of uncertainty in treating displaced PHF, that surgical decision making is difficult and that it may be prudent to involve experienced shoulder surgeons in deciding the best patients with displaced PHF management.[11]

In our analysis, the most considerable degree of agreement among different observers has occurred when classifying PHFs using Codman's system. In our study, the intra-observer agreement for Neer and AO classifications decreases as we temporarily move away from the start of the study (brief review of the classification systems and criteria homogenization session). We have also observed this loss of agreement effect when we have analysed the

inter-observer variability. In contrast, this tendency to lose intra and inter-observer agreement has not been as noticeable with Hertel's classification. This decline in the agreement may be due to the progressive loss of attention or interest from observers. The level of agreement has been higher among more expert observers, similar to that published in other studies,[24,32] except when the classification system used has been that of Codman-Hertel. This fact has two possible interpretations. On the one hand, the Codman-Hertel classification may be the one that best reproduces fracture patterns, regardless of the experience of the observer. On the other hand, and, perhaps in our study, being a classification less used in usual clinical practice, the training before the evaluations for the Hertel classification may have been similar among the observers, so the experience variable has had less influence on the outcome. It is essential to consider different fracture-related characteristics (not assessed by the Neer, AO or Codman-Hertel classifications) that may influence functional outcomes[24]: medial metaphyseal communication,[35] displacements in the coronal and sagittal planes, and bone loss on impaction.[10,21,36] The importance of having a system for classifying PHFs with low inter-observer variability goes beyond the academic realm. As indicated by LaMartina et al.,[11] successful management of PHF requires deciding between nonoperative or surgical treatment, deciding on the optimal surgical option for each case, and the technical ability to perform this surgical treatment. It is evident that without a univocal language of fracture, these objectives are difficult to achieve. Regardless of the imaging system used, we do not know the circumstances of the excessive inter-observer variability. It will be necessary to identify them to reduce it and improve reliability.

There are some limitations to our study. Firstly, our study bases its originality on the method applied, since the variability in the interpretation of the different classifications, widely published, is not a novelty. Secondly, we lack a 'gold standard' to compare the answers given by each of our evaluators and thus know their degree of accuracy (sensitivity and specificity), so our study is limited to assessing the diagnostic reliability of the three classification systems analysed. Thirdly, all the observers come from the same hospital centre, reducing the evaluation's variability and external validity.[24] Fourthly, we have exclusively used X-ray images (and a single standard anterior-posterior projection) for the study, which could decrease the intra-observer and inter-observer reliability. Although not without controversy, it is common to complete the information on the X-rays with 2D or 3D computer tomography (CT) images.[37–39] In a recent comparison of the agreement of the Neer's classification system among alone plain radiographs (AP and outlet view), CT images and 3D-reconstructed images, Torrens et al.[40] conclude that the different imaging techniques do not improve the agreement or concordance of the Neer's classification system. Furthermore, CT images are not routinely used in all hospital settings, so using only X-rays in the study may help increase external validity. Moreover, the study aimed to determine whether a single AP radiological image was sufficient for adequate diagnostic matching between different observers. These limitations notwithstanding, the authors believe that the study's outcomes are valuable because there are no published studies, to our knowledge, of the 'absolute reliability' of the PHF classifications with X-ray images.

Conclusions

The Neer, AO, and Hertel-Codman classification systems for PHF have a difficult interpretation for orthopaedic surgeons of varying levels of experience, and therefore substantial agreements are not obtained. According to our results, the system with the least variability in the classification has been that of Codman-Hertel.

Acknowledgement

The authors thank Manme Olvera (FIBAO Hospital Torrecárdenas) for its invaluable assistance with the statistical analysis.

Declaration of conflicting interests

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Vicente J León-Muñoz https://orcid.org/0000-0002-0429-2579

References

1. Court-Brown CM and Caesar B. Epidemiology of adult fractures: a review. *Injury* 2006; 37: 691–697.
2. Donaldson LJ, Cook A and Thomson RG. Incidence of fractures in a geographically defined population. *J Epidemiol Community Heal* 1990; 44: 241–245.
3. Kannus P, Niemi S, Sievänen H, et al. Stabilized incidence in proximal humeral fractures of elderly women: nationwide statistics from Finland in 1970–2015. *J Gerontol A* 2017; 72: 1390–1393.
4. Misra A, Kapur R and Maffulli N. Complex proximal humeral fractures in adults – a systematic review of management. *Injury* 2001; 32: 363–372.
5. Murray IR, Amin AK, White TO, et al. Proximal humeral fractures. *J Bone Joint Surg Br* 2011; 93-B: 1–11.
6. Palvanen M, Kannus P, Niemi S, et al. Update in the epidemiology of proximal humeral fractures. *Clin Orthop Relat Res* 2006; 442: 87–92.
7. Schumaier A and Grawe B. Proximal humerus fractures: evaluation and management in the elderly patient. *Geriatr Orthop Surg Rehabil* 2018; 9: 215145851775051.
8. Rangan A, Handoll H, Brealey S, et al. Surgical vs nonsurgical treatment of adults with displaced fractures of the proximal humerus. *JAMA* 2015; 313: 1037.

9. Beks RB, Ochen Y, Frima H, et al. Operative versus nonoperative treatment of proximal humeral fractures: a systematic review, meta-analysis, and comparison of observational studies and randomized controlled trials. *J Shoulder Elb Surg* 2018; 27: 1526–1534.
10. Foruria AM, de Gracia MM, Larson DR, et al. The pattern of the fracture and displacement of the fragments predict the outcome in proximal humeral fractures. *J Bone Joint Surg Br* 2011; 93-B: 378–386.
11. LaMartina J, Christmas KN, Simon P, et al. Difficulty in decision making in the treatment of displaced proximal humerus fractures: the effect of uncertainty on surgical outcomes. *J Shoulder Elb Surg* 2018; 27: 470–477.
12. Lanting B, MacDermid J, Drosdowech D, et al. Proximal humeral fractures: a systematic review of treatment modalities. *J Shoulder Elb Surg* 2008; 17: 42–54.
13. Brorson S, Alispahic N, Bahrs C, et al. Complications after non-surgical management of proximal humeral fractures: a systematic review of terms and definitions. *BMC Musculoskelet Disord* 2019; 20: 91.
14. Brorson S, Eckardt H, Audigé L, et al. Translation between the Neer- and the AO/OTA-classification for proximal humeral fractures: do we need to be bilingual to interpret the scientific literature? *BMC Res Notes* 2013; 6: 69.
15. Neer CS. Displaced proximal humeral fractures. I. Classification and evaluation. *J Bone Joint Surg Am* 1970; 52: 1077–1089.
16. Neer CS. Four-segment classification of proximal humeral fractures: purpose and reliable use. *J Shoulder Elb Surg* 2002; 11: 389–400.
17. Müller M.E., Koch P and Nazarian S. SJ. Humerus = 1. In: Müller ME, Koch P, Nazarian S and Schatzker J (eds) *The comprehensive classification of fractures of long bones*. Berlin: Springer, 1990, pp. 54–85.
18. Marsh JL, Slongo TF, Agel J, et al. Fracture and dislocation classification compendium – 2007. *J Orthop Trauma* 2007; 21: S1–S6.
19. Hertel R, Hempfing A, Stiehler M, et al. Predictors of humeral head ischemia after intracapsular fracture of the proximal humerus. *J Shoulder Elb Surg* 2004; 13: 427–433.
20. Siebenrock KA and Gerber C. The reproducibility of classification of fractures of the proximal end of the humerus. *J Bone Joint Surg* 1993; 75: 1751–1755.
21. Majed A, Macleod I, Bull AMJ, et al. Proximal humeral fracture classification systems revisited. *J Shoulder Elb Surg* 2011; 20: 1125–1132.
22. Bernstein J, Adler LM, Blank JE, et al. Evaluation of the Neer system of classification of proximal humeral fractures with computerized tomographic scans and plain radiographs. *J Bone Joint Surg* 1996; 78: 1371–1375.
23. Brunner A, Honigmann P, Treumann T, et al. The impact of stereo-visualisation of three-dimensional CT datasets on the inter- and intraobserver reliability of the AO/OTA and Neer classifications in the assessment of fractures of the proximal humerus. *J Bone Joint Surg Br* 2009; 91-B: 766–771.
24. Gracitelli MEC, Dotta TAG, Assunção JH, et al. Intraobserver and interobserver agreement in the classification and treatment of proximal humeral fractures. *J Shoulder Elb Surg* 2017; 26: 1097–1102.
25. Sallay PI, Pedowitz RA, Mallon WJ, et al. Reliability and reproducibility of radiographic interpretation of proximal humeral fracture pathoanatomy. *J Shoulder Elb Surg* 1997; 6: 60–69.
26. Sjödén GOJ, Movin T, Aspelin P, et al. 3D-radiographic analysis does not improve the Neer and AO classifications of proximal humeral fractures. *Acta Orthop Scand* 1999; 70: 325–328.
27. Sjödén GOJ, Movin T, Güntner P, et al. Poor reproducibility of classification of proximal humeral fractures: additional CT of minor value. *Acta Orthop Scand* 1997; 68: 239–242.
28. Atkinson G and Nevill AM. Selected issues in the design and analysis of sport performance research. *J Sports Sci* 2001; 19: 811–827.
29. Hopkins WG. Measures of reliability in sports medicine and science. *Sport Med* 2000; 30: 1–15.
30. Fleiss J. Measuring nominal scale agreement among many raters. *Psychol Bull* 1971; 76: 378–382.
31. Landis JR KG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33: 159–174.
32. Sidor ML, Zuckerman JD, Lyon T, et al. The Neer classification system for proximal humeral fractures. An assessment of interobserver reliability and intraobserver reproducibility. *J Bone Joint Surg* 1993; 75: 1745–1750.
33. Resch H, Tauber M, Neviaser RJ, et al. Classification of proximal humeral fractures based on a pathomorphologic analysis. *J Shoulder Elb Surg* 2016; 25: 455–462.
34. Sukthankar AV, Leonello DT, Hertel RW, et al. A comprehensive classification of proximal humeral fractures: HGLS system. *J Shoulder Elb Surg* 2013; 22: e1–e6.
35. Krappinger D, Bizzotto N, Riedmann S, et al. Predicting failure after surgical fixation of proximal humerus fractures. *Injury* 2011; 42: 1283–1288.
36. Court-Brown CM, Garg A and McQueen MM. The translated two-part fracture of the proximal humerus. *J Bone Joint Surg* 2001; 83: 799–804.
37. Meleán P, Munjin A, Pérez A, et al. Coronal displacement in proximal humeral fractures: correlation between shoulder radiographic and computed tomography scan measurements. *J Shoulder Elb Surg* 2017; 26: 56–61.
38. Spross C, Meester J, Mazzucchelli RA, et al. Evidence-based algorithm to treat patients with proximal humerus fractures – a prospective study with early clinical and overall performance results. *J Shoulder Elb Surg* 2019; 28: 1022–1032.
39. Sumrein BO, Mattila VM, Lepola V, et al. Intraobserver and interobserver reliability of recategorized Neer classification in differentiating 2-part surgical neck fractures from multifragmented proximal humeral fractures in 116 patients. *J Shoulder Elb Surg* 2018; 27: 1756–1761.
40. Torrens C, Marí R, Cuenca M, et al. 3D reconstruction does not improve agreement and results in an increase in surgical indications in proximal humeral fractures. *J Orthop* 2018; 15: 967–970.