ANZJSurg.com

Interobserver agreement of Neer and AO classifications for proximal humeral fractures

Maritsa K. Papakonstantinou,* Melissa J. Hart,†‡ Richard Farrugia,§ Belinda J. Gabbe,¶ Afshin Kamali Moaveni,‖ Dirk van Bavel,*†† Richard S. Page‡‡§§ and Martin D. Richardson††¶¶

*Department of Orthopaedics, St Vincent's Hospital, Melbourne, Victoria, Australia
†Victorian Orthopaedic Trauma Outcomes Registry (VOTOR), Melbourne, Victoria, Australia
‡Department of Epidemiology and Preventive Medicine, School of Public Health and Preventive Medicine, Monash University, Melbourne, Victoria, Australia
§Department of Orthopaedics, The Royal Melbourne Hospital, Melbourne, Victoria, Australia
¶Department of Epidemiology and Preventative Medicine, Monash University, Melbourne, Victoria, Australia
‖Department of Orthopaedics, The Alfred Hospital, Melbourne, Victoria, Australia
††Department of Orthopaedics, The Epworth Hospital, Melbourne, Victoria, Australia
‡‡Department of Orthopaedics, University Hospital, Geelong, Victoria, Australia
§§School of Medicine, Deakin University, Geelong, Victoria, Australia
¶¶Department of Surgery, The University of Melbourne, Melbourne, Victoria, Australia

Key words
classification, humerus, observer variation, shoulder fractures, X-rays.

Correspondence
Dr Maritsa K. Papakonstantinou, Department of Orthopaedics, St Vincent's Hospital, 59 Victoria Parade, Fitzroy, Vic. 3065, Australia.
Email: m.papakonstantinou@ymail.com

M. K. Papakonstantinou BMed, BMedSci; **M. J. Hart** RN, BEd; **R. Farrugia** BN, BAppSci (Physics); **B. J. Gabbe** PhD; **A. Kamali Moaveni** FRACS, FAOrthA; **D. van Bavel** MBBS, FRACS (Orth); **R. S. Page** BMedSci, FRACS; **M. D. Richardson** BS, MS, FRACS.

Accepted for publication 7 December 2015.

doi: 10.1111/ans.13451

Abstract

Background The classification of proximal humeral fractures remains challenging. The two main classification systems used, the Neer and the AO classification, have both been shown to have less than ideal interobserver agreement. Agreement in classification is required, however, to guide fracture management.

Method Data from the Victorian Orthopaedic Trauma Outcomes Registry were collected and the X-rays of 104 proximal humeral fractures were reviewed by three orthopaedic consultants. They classified the fractures according to the Neer and AO classifications, as well as their simplified versions. Interobserver agreement was then assessed using kappa statistics.

Results Interobserver agreement was better overall in the Neer classification, which was moderate (kappa = 0.40–0.58), than the AO classification, which was fair to moderate (kappa = 0.31–0.54). When simplified, the Neer and AO classification interobserver agreement remained similar.

Conclusion The classification of proximal humeral fractures with both the Neer and the AO systems remains difficult with minimal improvements seen when reducing the number of categories in each classification system. From these results, the Neer classification system would appear slightly more useful in clinical practice to guide treatment.

Introduction

Classification systems for fractures are important because they should help guide their management.[1] They should be easy to use and understand and allow the observer to come to the same classification regardless of experience. Both the Neer classification, devised by Charles S. Neer II in 1970[2] and AO classification developed in the 1980s,[3] are used to guide proximal humeral fracture management.

The Neer classification has as its basis, a system devised by Codman in 1934, in which the proximal humerus was described as fracturing into four main fragments (humeral head, greater tuberosity, lesser tuberosity and shaft).[4] Neer described the effect of displacement forces exerted on the fracture fragments by their musculotendinous attachments enabling 16 fracture categories to be identified.[2]

The AO classification is based on the severity of the fracture and the likely disruption to the vascularity of the proximal humerus. Three broad types of fracture exist. Type A fractures are extra-articular and unifocal, type B fractures are extra-articular and bifocal and type C fractures are articular. These three fracture types are divided into three groups and three subgroups based on the degree

of displacement, impaction and dislocation of fracture fragments. This adds up to 27 subgroups in the Arbeitsgemeinschaft für Osteosynthesefragen/Association for the Study of Internal Fixation classification system.[3]

A number of studies have evaluated the interobserver agreement that exists in the Neer classification. Most studies show only fair agreement, as shown in a study by Brorson *et al*. who achieved a kappa value of 0.33,[5] or moderate agreement as shown by Brien *et al.* (k = 0.45).[6] There have been considerably fewer studies analysing interobserver agreement in the AO system, in particular, the full 27 category version. Majed *et al.* demonstrated only slight interobserver agreement in the full AO system with a kappa coefficient of 0.11.[7] In most studies, one of the disadvantages has been the relatively low number of X-rays reviewed[6,8–12] and that cases with an incomplete series of images were excluded.

Many authors have tried to use more specialized imaging modalities with the aim of improving interobserver variation. Their results have not so far supported this hypothesis. Bernstein *et al*. compared X-rays and computed tomography (CT) using the Neer classification and achieved kappa coefficients of 0.52 when X-rays alone were used and 0.50 when X-rays and CT were used.[8] Similar results were obtained by Sjoden *et al*.[9] Even when three-dimensional CT was used, results did not improve.[10] Foroohar *et al.* compared X-rays with CT and three-dimensional reconstructions and found no significant improvement in kappa values for the Neer classification (k = 0.14 for X-rays, k = 0.09 for three-dimensional CT and k = 0.07 for two-dimensional CT).[12]

Given that there is a paucity of work comparing interobserver agreement between the Neer and AO classification systems and that advanced imaging modalities have not been shown to improve interobserver agreement, our aim was to compare the full and simplified versions of these two systems using only X-rays.

Method

Ethics approval was sought and granted from the ethics committee at The Alfred Hospital (Protocol Number: 216/09), which participates in the Victorian Orthopaedic Trauma Outcomes Registry (VOTOR). VOTOR is a clinical quality registry that forms a comprehensive database of orthopaedic injuries, treatments, complications and outcomes based on admissions to four Victorian hospitals. It collects information on all patients over 16 years of age admitted with a new orthopaedic injury with a length of hospital stay over 24 h. Pathologic orthopaedic injuries are excluded. Information is prospectively collected from patients who are able to opt-out from the registry at any time. Through the registry, 108 cases of proximal humeral fractures from 107 participants were identified between April 2007 and July 2008. From these, one case was excluded as X-rays were not available, two were excluded because the fractures were of the proximal diaphysis and one excluded as no fracture was identified. This left 104 proximal humeral fractures for classification.

Plain X-rays alone were used to classify fractures. These were taken at the time of injury. X-ray series usually included an AP and trans-scapular lateral view. Some cases had additional views (e.g. axillary). In a few trauma cases, only an AP X-ray was performed. X-ray quality varied substantially. All X-rays were digital.

Three orthopaedic surgeons with an interest in upper limb pathology and trauma classified the fractures. The surgeons had between 2 and 15 years of experience in the upper limb practice. All three surgeons reviewed and classified the fractures according to the full Neer (17 categories including the four-part valgus impacted category described by Jacob *et al.*[13]) and AO (27 categories) classifications, as well as their simplified versions. Documents with the full

Table 1 Distribution of rater classifications using the Neer classification

Category		Rater 1	Rater 2	Rater 3
		n (%)	*n* (%)	*n* (%)
Full	1-Part minimally displaced	0 (0.0)	2 (1.9)	6 (5.9)
	2-Part anatomical neck	1 (1.0)	1 (1.0)	2 (2.0)
	2-Part surgical neck	48 (47.5)	48 (46.6)	37 (36.3)
	2-Part greater tuberosity	18 (17.8)	10 (9.7)	13 (12.7)
	2-Part lesser tuberosity	0 (0.0)	0 (0.0)	0 (0.0)
	2-Part fracture dislocation (anterior)	9 (8.9)	13 (12.6)	6 (5.9)
	2-Part fracture dislocation (posterior)	2 (2.0)	1 (1.0)	3 (2.9)
	3-Part greater tuberosity	18 (17.8)	22 (21.4)	23 (22.5)
	3-Part lesser tuberosity	1 (1.0)	0 (0.0)	3 (2.9)
	3-Part fracture dislocation (anterior)	0 (0.0)	1 (1.0)	4 (3.9)
	3-Part fracture dislocation (posterior)	0 (0.0)	0 (0.0)	1 (1.0)
	4-Part valgus impacted	1 (1.0)	2 (1.9)	2 (2.0)
	4-Part 'classic'	1 (1.0)	2 (1.9)	0 (0.0)
	4-Part fracture dislocation (anterior)	1 (1.0)	1 (1.0)	0 (0.0)
	4-Part fracture dislocation (posterior)	0 (0.0)	0 (0.0)	0 (0.0)
	Articular surface-head splitting	1 (1.0)	0 (0.0)	2 (2.0)
	Articular surface-head crushing	0 (0.0)	0 (0.0)	0 (0.0)
Total (%)		101 (100.0)	103 (100.0)	102 (100.0)
4-Part	1-Part	0 (0.0)	2 (1.9)	6 (5.9)
	2-Part	78 (77.2)	73 (70.9)	61 (59.8)
	3-Part	19 (18.8)	23 (22.3)	31 (30.4)
	4-Part	4 (4.0)	5 (4.9)	4 (3.9)
Total (%)		101 (100.0)	103 (100.0)	102 (100.0)

Table 2 Level of agreement between raters – Neer classification

Neer	Raters	% Agreement	Kappa (95% CI)	Weighted kappa (95% CI)
Complete classification	1 versus 2	71.0	0.59 (0.47, 0.72)	0.58 (0.43, 0.71)
	1 versus 3	55.0	0.41 (0.29, 0.52)	0.40 (0.28, 0.54)
	2 versus 3	57.4	0.44 (0.34, 0.56)	0.46 (0.33, 0.59)
4-Category classification	1 versus 2	80.0	0.52 (0.33, 0.69)	0.59 (0.41, 0.74)
	1 versus 3	70.0	0.37 (0.21, 0.53)	0.43 (0.27, 0.58)
	2 versus 3	70.3	0.41 (0.26, 0.57)	0.45 (0.28, 0.60)

CI, confidence interval.

classifications including pictures and descriptions were provided to all observers.

The simplified Neer classification included the following four categories: one-part fractures (minimally displaced), two-part fractures, three-part fractures and four-part fractures. Two versions of the simplified AO classification were used. A nine-category system where the three types of proximal humeral fracture (Type A, B and C) was each divided into three groups related to fracture pattern and a three-category system where only the type of fracture was classified.

Interobserver agreement was assessed with kappa statistics that quantifies the absolute agreement of observers, accounting for the agreement that would occur by chance alone. Kappa, and weighted kappa, coefficients were calculated for each rater pair for the full Neer classification, the four-category Neer classification, the full AO classification, the nine-category AO classification and the three-category AO classification. The 95% confidence intervals for kappa and weighted kappa coefficients were calculated using the 95th percentile interval from 1000 bootstrap replications as described by Landis and Kock. Kappa values <0.00 were regarded as poor agreement, values between 0.00 and 0.20 as slight agreement, values of 0.21–0.40 showed fair agreement, 0.41–0.60 demonstrated moderate agreement, 0.61–0.80 showed substantial agreement and 0.81–1.00 indicated almost perfect agreement.[14]

Table 3 Distribution of rater classifications using the AO classification

Category		Rater 1	Rater 2	Rater 3
		n (%)	*n* (%)	*n* (%)
Full	A1.1	8 (7.8)	3 (2.9)	5 (5.0)
	A1.2	9 (8.8)	7 (6.8)	7 (7.0)
	A1.3	7 (6.9)	12 (11.6)	2 (2.0)
	A2.1	12 (11.8)	5 (4.8)	6 (6.0)
	A2.2	6 (5.9)	3 (2.9)	6 (6.0)
	A2.3	2 (2.0)	3 (2.9)	1 (1.0)
	A3.1	2 (2.0)	20 (19.4)	0 (0.0)
	A3.2	24 (23.5)	14 (13.6)	8 (8.0)
	A3.3	3 (2.9)	6 (5.8)	9 (9.0)
	AB1.1	7 (6.9)	9 (8.7)	5 (5.0)
	B1.2	0 (0.0)	0 (0.0)	1 (1.0)
	B1.3	10 (9.8)	1 (1.0)	5 (5.0)
	B2.1	0 (0.0)	3 (2.9)	2 (2.0)
	B2.2	1 (1.0)	5 (4.8)	10 (10.0)
	B2.3	1 (1.0)	3 (2.9)	3 (3.0)
	B3.1	0 (0.0)	0 (0.0)	0 (0.0)
	B3.2	0 (0.0)	1 (1.0)	10 (10.0)
	B3.3	1 (1.0)	0 (0.0)	1 (1.0)
	C1.1	2 (2.0)	0 (0.0)	3 (3.0)
	C1.2	1 (1.0)	0 (0.0)	10 (10.0)
	C1.3	0 (0.0)	0 (0.0)	0 (0.0)
	C2.1	1 (1.0)	5 (4.8)	4 (4.0)
	C2.2	1 (1.0)	0 (0.0)	0 (0.0)
	C2.3	1 (1.0)	0 (0.0)	2 (2.0)
	C3.1	0 (0.0)	2 (1.9)	0 (0.0)
	C3.2	1 (1.0)	0 (0.0)	0 (0.0)
	C3.3	2 (2.0)	1 (1.0)	0 (0.0)
Total (%)		102 (100.0)	103 (100.0)	100 (100.0)
9-Category	A1.1–A1.3	24 (23.5)	22 (21.4)	14 (14.0)
	A2.1–A2.3	20 (19.6)	11 (10.7)	13 (13.0)
	A3.1–A3.3	29 (28.4)	40 (38.8)	17 (17.0)
	B1.1–B1.3	17 (16.7)	10 (9.7)	11 (11.0)
	B2.1–B2.3	1 (1.0)	11 (10.7)	15 (15.0)
	B3.1–B3.3	1 (1.0)	1 (1.0)	11 (11.0)
	C1.1–C1.3	3 (2.9)	0 (0.0)	13 (13.0)
	C2.1–C2.3	3 (2.9)	5 (4.8)	6 (6.0)
	C3.1–C3.3	4 (3.9)	3 (2.9)	0 (0.0)
Total		102 (100.0)	103 (100.0)	100 (100.0)
3-Category	A	73 (71.6)	73 (70.9)	44 (44.0)
	B	19 (18.6)	22 (21.4)	37 (37.0)
	C	10 (9.8)	8 (7.8)	19 (19.0)
Total (%)		102 (100.0)	103 (100.0)	100 (100.0)

Results

The results of the study indicate that overall interobserver agreement was fair to moderate in both classification systems. There was overall moderate agreement in the full Neer classification (k = 0.40–0.58) and fair to moderate overall agreement in the full AO classification (k = 0.31–0.54).

Regarding the simplified versions of the Neer and AO classifications, results did not improve substantially. Agreement in the four-category Neer classification improved only marginally (k = 0.43–0.59). Agreement remained almost the same for the nine-category (k = 0.29–0.54) and three-category (k = 0.31–0.54) AO systems compared with the full AO classification (Tables 1–4).

Discussion

The present study aimed to evaluate interobserver agreement of the Neer and AO classification systems using the X-ray images available at the time of injury. As the study was retrospective, it was not possible to order additional imaging or to repeat X-rays that were not of high quality. This was one of the limitations of the study. Despite lack of standardization in X-ray quality and type, the results of the current study are similar to previous studies.

Regarding interobserver agreement in the Neer system, an overall kappa value of 0.40–0.58 was achieved. This equates to 'moderate' agreement and is similar to results obtained from other researchers such as Bernstein *et al.* (k = 0.52), Siebenrock and

Table 4 Level of agreement between raters – AO classification

Tile/AO	Raters	% Agreement	Kappa (95% CI)	Weighted kappa (95% CI)
Complete classification	1 versus 2	40.6	0.36 (0.27, 0.46)	0.54 (0.40, 0.65)
	1 versus 3	29.3	0.25 (0.17, 0.35)	0.31 (0.18, 0.43)
	2 versus 3	28.3	0.25 (0.16, 0.34)	0.32 (0.19, 0.44)
9-Category classification	1 versus 2	57.4	0.47 (0.37, 0.59)	0.54 (0.39, 0.67)
	1 versus 3	36.4	0.27 (0.17, 0.37)	0.29 (0.16, 0.42)
	2 versus 3	40.4	0.31 (0.20, 0.41)	0.32 (0.19, 0.46)
3-Category classification	1 versus 2	77.2	0.49 (0.33, 0.66)	0.54 (0.36, 0.69)
	1 versus 3	60.6	0.35 (0.21, 0.49)	0.33 (0.19, 0.48)
	2 versus 3	56.6	0.27 (0.15, 0.41)	0.31 (0.18, 0.46)

CI, confidence interval.

Gerber (k = 0.40) and Sidor *et al*. (k = 0.48).[8,15,16] Likewise, there was a slightly lower rate of interobserver agreement in the AO system both in our study (kappa value of 0.31–0.54, equal to 'fair to moderate' agreement) and in the literature. Sjoden *et al*. obtained a kappa value of 0.31 while Majed *et al*. had even lower agreement with a kappa of 0.11.[7,9] Interestingly, Siebenrock and Gerber had higher interobserver agreement in the AO system in comparison with the Neer system (k = 0.53).[515]

One of the reasons posed as the cause for lower agreement is the high number of categories in the respective classification systems. Sidor *et al*. achieved an overall kappa value of 0.48 between all observers for the full Neer classification and in fact achieved a lower kappa value of 0.42 when a six-category version was used.[16] Bernstein *et al*. used a six-category Neer classification and achieved kappa values of 0.54 compared with the full Neer classification that had values of 0.50.[8] The six-category version of the Neer classification was also used by Mahadeva *et al*. who surprisingly achieved 'substantial' agreement (kappa range 0.617–0.730).[17] Siebenrock and Gerber used a four-category Neer classification and had interobserver agreement rate of 0.40.[15] Brorson *et al*. reduced the number of categories in the Neer classification to two (non-displaced and displaced) achieving 'moderate' agreement with a kappa coefficient of 0.41 (up from 0.27 when the six-category system was used).[18] Overall, while in some studies there was an improvement in agreement, this has not been as large as hoped. In the present study, reducing the Neer categories to four did not improve interobserver variation.

Simplified AO classifications have been used by a small number of researchers. Majed *et al*. simplified the AO classification to three categories and achieved an interobserver kappa value of 0.30 compared with 0.11 for the full 27-category system.[7] Siebenrock and Gerber also demonstrated an improvement in agreement with the three-category system (k = 0.53) compared with the nine-category AO system (k = 0.42).[15] No substantial improvement was shown in our study when simplifying the full classification system to the nine- and three-category systems.

Given the somewhat limited agreement seen with all classification systems, it is not surprising that the management of proximal humeral fractures can be quite varied and challenging. One factor that could positively influence the agreement in classification of proximal humeral fractures is experience in the field. Although Sidor *et al*. reported kappa values of 0.83 when an orthopaedic shoulder surgeon used the Neer classification, compared with orthopaedic residents (k = 0.48), the hypothesis that more experience equates the better interobserver agreement has not been supported by other authors.[5,16]

Conclusion

The classification of proximal humeral fractures with both the Neer and AO systems remains difficult. Very minimal improvements have so far been seen when reducing the number of categories in each classification system. In addition, more advanced imaging modalities have failed to improve interobserver agreement significantly. The present study suggests that slightly better results for interobserver agreement are found in the full Neer classification system. This would make it the more useful classification to use in clinical practice.

Acknowledgements

The Victorian Orthopaedic Trauma Outcomes Registry (VOTOR) is funded by the TAC via the Institute for Safety Compensation and Recovery Research (ISCRR). Study sponsors had no involvement in the study design, collection, analysis or interpretation of data. We would like to thank the VOTOR telephone interviewers, VOTOR Steering Committee and the participating hospitals of the VOTOR.

References

1. Burstein AH. Fracture classification systems: do they work and are they useful? *J. Bone Joint Surg.* 1993; **75-A**: 1743–4.
2. Neer CS II. Displaced proximal humeral fractures: part I. Classification and evaluation. *J. Bone Joint Surg.* 1970; **52A**: 1077–89.
3. Muller M, Nazarian S, Koch P, Schatzker J. *The Comprehensive Classification of Fractures in Long Bones*. Berlin: Springer-Verlag, 1990.
4. Codman EA. Fractures in relation to the subacromial bursa. In: Codman EA (ed.). *The Shoulder*. Boston: Thomas Todd, 1934; chapter X.
5. Brorson S, Olsen BS, Frich LH *et al*. Surgeons agree more on treatment recommendations than on classification of proximal humeral fractures. *Musculoskelet. Disord.* 2012; **13**: 114.

6. Brien H, Notfall F, MacMaster S, Cummings T, Landells C, Rockwood P. Neer's classification system: a critical appraisal. *J. Trauma* 1995; **38**: 257–60.
7. Majed A, Macleod I, Bull AMJ *et al*. Proximal humeral fracture classification systems revisited. *J. Shoulder Elbow Surg.* 2011; **20**: 1125–32.
8. Bernstein J, Adler LM, Blank JE, Dalsey RM, Williams GR, Iannotti JP. Evaluation of the Neer system of classification of proximal humeral fractures with computerized tomographic scans and plain radiographs. *J. Bone Joint Surg. Am.* 1996; **78-A**: 1371–5.
9. Sjoden GOJ, Movin T, Guntner P *et al*. Poor reproducibility of classification of proximal humeral fractures: additional CT of minor value. *Acta Orthop. Scand.* 1997; **68**: 239–42.
10. Sjoden GOJ, Movin T, Aspelin P, Shalabi A. 3D-radiographic analysis does not improve the Neer and AO classifications of proximal humeral fractures. *Acta Orthop. Scand.* 1999; **70**: 325–8.
11. Sallay PI, Pedowitz RA, Mallon WJ, Vandemark RM, Dalton JD, Speer KP. Reliability and reproducibility of radiographic interpretation of proximal humeral fracture pathoanatomy. *J. Shoulder Elbow Surg.* 1997; **6**: 60–9.
12. Foroohar A, Tosti R, Richmond JM, Gaughan JP, Ilyas AM. Classification and treatment of proximal humeral fractures: inter-observer reliability and agreement across imaging modalities and experience. *J. Orthop. Surg.* 2011; **6**: 38.
13. Jakob RP, Miniaci A, Anson PS, Jaberg H, Osterwalder A, Ganz R. Four-part valgus impacted fractures of the proximal humerus. *J. Bone Joint Surg.* 1991; **73-B**: 295–8.
14. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33**: 159–74.
15. Siebenrock KA, Gerber C. The reproducibility of classification of fractures of the proximal end of the humerus. *J. Bone Joint Surg.* 1993; **75-A**: 1751–5.
16. Sidor ML, Zuckerman JD, Lyon T, Koval K, Cuomo F, Schoenberg N. The Neer classification system for proximal humeral fractures. *J. Bone Joint Surg.* 1993; **75-A**: 1745–50.
17. Mahadeva D, Dias RG, Deshpande SV, Datta A, Dhillon SS, Simons AW. The reliability and reproducibility of the Neer classification system – digital radiography (PACS) improves agreement. *Injury* 2011; **42**: 339–42.
18. Brorson S, Bagger J, Sylvest A, Hrobjartsson A. Low agreement among 24 doctors using the Neer-Classification; only moderate agreement on displacement, even between specialists. *Int. Orthop.* 2002; **26**: 271–3.