International Orthopaedics (SICOT) (2002) 26:271–273
DOI 10.1007/s00264-002-0369-x

ORIGINAL PAPER

S. Brorson · J. Bagger · A. Sylvest · A. Hróbjartsson

Low agreement among 24 doctors using the Neer-classification; only moderate agreement on displacement, even between specialists

Accepted: 23 April 2002 / Published online: 8 June 2002
© Springer-Verlag 2002

Abstract Twenty-four orthopaedic surgeons classified 42 pairs of radiographs according to the Neer system for proximal humeral fractures. Mean kappa value for inter-observer agreement was 0.27 (95% CI 0.26–0.28) with no clinically significant difference between orthopaedic residents (*n*=9), fellows (*n*=6) and specialists (*n*=9). Mean kappa for agreement of displacement versus non-displacement was 0.41 (95% CI 0.39–0.43) overall, and 0.50 (95% CI 0.45–0.56) within the specialist group. The agreement found in our study is unsatisfactory from a clinical perspective.

Résumé 24 chirurgiens orthopédistes ont classé 42 paires de radiographies selon la classification de Neer pour les fractures humérales proximales. La valeur moyenne du kappa pour l'accord inter-observateur était 0.27 (95% CI 0.26–0.28) sans différence significative entre résidents (*n*=9), chirurgiens juniors (*n*=6) et chirurgiens spécialistes (*n*=9). La valeur moyenne du kappa pour l'accord sur le critère déplacement ou non-déplacement était globalement 0.41 (95% CI 0.39–0.43), et, dans le groupe spécialiste 0.50 (95% CI 0.45–0.56). Dans une perspective clinique l'accord trouvé dans notre étude est peu satisfaisant.

Introduction

The Neer classification [11] (Fig. 1) of proximal humeral fractures is commonly used. Several previous studies have shown poor agreement between observers using the system [1, 2, 3, 9, 12, 13, 14, 15, 16]. Low agreement might explain conflicting study outcomes of the treatment of three- and four-part fractures [4]. All previous observer studies have used a fairly limited number of observers or cases. Thus, the general agreement among doctors using the Neer system, and especially sub-analyses of the agreement on certain Neer categories, and on the agreement within different levels of clinical experience, have been somewhat imprecisely estimated.

We conducted a large study based on a consecutive series of cases, and included as observers the full medi-

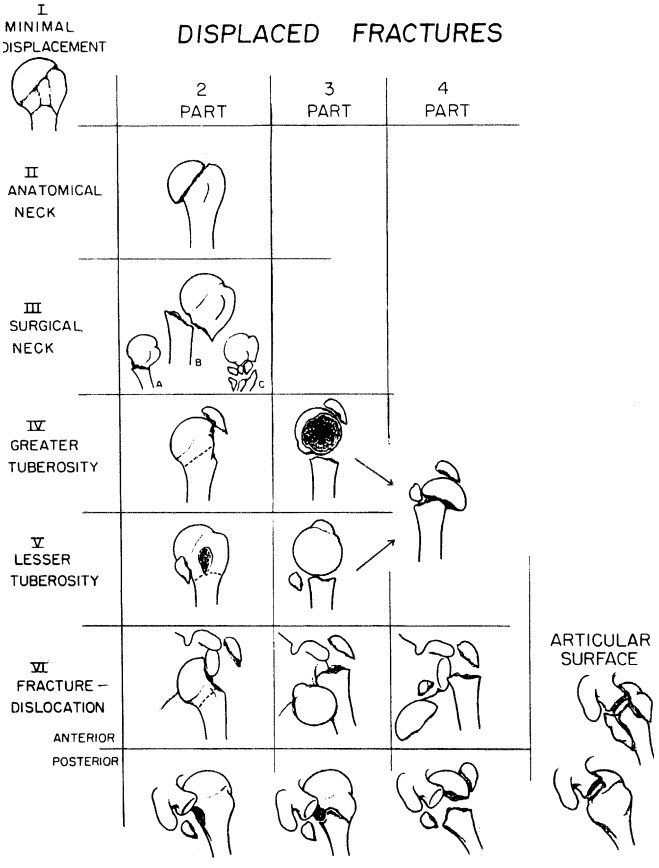

Fig. 1 The Neer system for classification of proximal humeral fractures. Reprinted with permission

S. Brorson (✉) · A. Hróbjartsson
Department of Medical Philosophy and Clinical Theory,
University of Copenhagen, Panum Institute,
Blegdamsvej 3, 2200 Copenhagen N, Denmark
e-mail: e-mail: s.brorson@medphil.ku.dk
Tel.: +45-35327289, Fax: +45-35327938

J. Bagger · A. Sylvest
Department of Orthopaedic Surgery,
Bispebjerg University Hospital, 2400 Copenhagen NV, Denmark

cal staff at the department of orthopaedic surgery in a large university hospital. Our primary aim was to investigate the inter-observer variation within a large group of orthopaedic surgeons using the Neer system. Our secondary aims were to study whether the inter-observer variation differed between doctors at three levels of clinical experience and how much the inter-observer variation changed if the Neer system was reduced to a binary classification (i.e. displaced versus non-displaced).

Material and methods

A consecutive series of 42 proximal humeral fractures in patients discharged from our department within an arbitrarily chosen 10-week period was identified. Plain anterior-posterior and lateral radiographs were available for all cases. No selection according to the quality of radiographs was performed, and none of the cases underwent additional CT or MRI scans subsequently.

The entire medical staff on duty in our department on a certain day was included as observers. After a brief introduction the observers were provided with a ruler and goniometer, and a diagram of the Neer system with a written definition of displacement. They were all asked to classify all cases independently. None of the observers were informed about the study prior to the classification and were not permitted to communicate during the classification. The authors did not participate as observers.

Statistics

Agreement between observers was calculated using kappa statistics. Kappa statistics adjusts simple observed agreement for chance agreement. The kappa coefficients range from 1 (perfect agreement) through 0 (chance agreement) to less than 0 (systematic disagreement). According to Landis and Koch [10] values less than 0 indicate poor reliability, 0.00–0.20 slight reliability, 0.21–0.40 fair reliability, 0.41–0.60 moderate reliability, 0.61–0.80 substantial agreement, and 0.81–1.00 excellent or almost perfect agreement.

Prior to calculations the original 16 Neer groups were reduced to six groups (Table 1). Mean kappa values for all pairs of observers were calculated, and a pre-defined subgroup analysis of agreement within three levels of clinical experience was performed. In addition, a calculation of mean kappa values for displaced versus non-displaced fractures was performed, and 95% confidence intervals were calculated according to Svanholm et al. [17].

Results

Twenty-four doctors (nine orthopaedic residents, six fellows and nine specialists) participated in the study and classified all cases. Mean kappa for agreement between all pairs of observers was 0.27 (95% CI 0.26–0.28). No clinically important difference between the mean kappa values of residents, fellows and specialists was detected (Table 2). Inexperienced doctors tended to classify more fractures as displaced, but otherwise there were no differences between the three levels of experience. The highest agreement was found for the categories "non-displaced" and "fracture dislocation" (Table 1). If the Neer system was reduced to a binary classification (Neer group 1 versus all other groups), mean kappa was 0.41 (95% CI 0.39–0.43) overall and for specialists 0.50 (95% CI 0.45–0.56).

Table 1 Proportional distribution and inter-observer (n=24) agreement of the six main categories

	Proportion	Mean kappa
Non-displaced	0.21	0.41
Two-part fractures	0.42	0.18
Three-part fractures	0.21	0.22
Four-part fractures	0.07	0.16
Fracture-dislocation	0.08	0.44
Articular surface	0.01	0.12

Table 2 Mean kappa values (95% CI) for pairs of observers in a six-group classification

	Number	Mean kappa
All observers	24	0.27 (0.26–0.28)
Residents	9	0.26 (0.23–0.29)
Fellows	6	0.21 (0.16–0.25)
Specialists	9	0.33 (0.30–0.35)

Discussion

Our study confirmed the poor inter-observer agreement found in previous studies of the Neer system based on plain radiographs. Kristiansen et al. [9] reported paired kappa values between 0.07 and 0.48. Sidor et al. [13] reported a mean kappa value of 0.50. Siebenrock and Gerber [14] reported a mean kappa value of 0.40. Brien et al. [3] reported a mean kappa-value of 0.45. Bernstein et al. [2] reported a mean kappa value of 0.52. No significant improvement of mean kappa values has been demonstrated despite exclusive use of high-quality radiographs [2, 13] or by adding CT scans and three-dimensional reconstructions [2, 12, 15, 16].

We used a larger number of observers than in any previous study and imitated the clinical situation by exclusively using unselected and consecutive radiographs. We found even lower mean kappa values than the previous studies. However, comparisons of kappa values from different studies may be problematic, as kappa changes with the prevalence of the diagnosis [6]. In our sample we found a prevalence of displaced fractures of 79%, which is considerably higher than the prevalences reported by Neer [11] (20%), Horak and Nilsson [7] (39%), Kiær et al. [8] (57%) and Court-Brown et al. [5] (51%). Our finding may be due to differences in classifying among our observers or due to an unrepresentative sample of radiographs. However, from a clinical perspective, the reported levels of observer agreement are far from satisfying. A different distribution of the categories may be expected if the observers were asked initially to classify the radiographs into two or six categories instead of 16 categories. However, Sidor et al. [13] and Bernstein et al. [2] found no increase in kappa when they reduced the number of units in the Neer system prior to classification.

Several studies have addressed the impact of clinical experience when using the Neer system, but the results have been inconsistent. A study including four observers [9]

found the lowest level of agreement in pairs involving the less experienced observer. Sidor et al. [13] reported a slightly higher agreement between attending physicians. Siebenrock and Gerber [14] included shoulder surgeons exclusively and ruled out lack of experience as a main factor contributing to low agreement. Sallay et al. [12] concluded that inter-observer agreement was suboptimal regardless of the level of experience. We included substantially more observers than in any previous study and found that specialists performed slightly better than residents and fellows. Even so, all kappa values were unsatisfactorily low.

In making clinical decisions, not all categories in the Neer system are of equal significance. The initial decision of whether to treat the patient conservatively with a sling and early exercises or to consider other procedures depends on the ability to distinguish displaced fractures from non-displaced. We found the highest agreement on these two main types of fractures among orthopaedic specialists. However, an increase in kappa is expected due to a reduced number of categories, and a kappa-value of 0.50 is still dissatisfying from a clinical perspective.

To safely recommend the Neer system for high quality clinical care and research, ways of improving inter-observer variation must be identified. We recommend further studies into means of improving observer agreement, e.g. educational interventions.

Acknowledgments We thank the medical staff at the Department of Orthopaedic Surgery, Bispebjerg University Hospital, for their participation. Stig Brorson was supported by a grant from the Danish Medical Research Council. 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.

References

1. Ackermann C, Lam Q, Linder P, Kull C, Regazzoni P (1986) Zur Problematik der Frakturklassifikation am proximalen Humerus. Z Unfallchir Vers med Berufskr 79:209–215
2. Bernstein J, Adler LM, Blank JE, Dalsey RM, Williams GR, Iannotti JP (1996) Evaluation of the Neer system of classification of proximal humeral fractures with computerized tomographic scans and plain radiographs. J Bone Joint Surg [Am] 78:1371–1375
3. Brien H, Notfall F, MacMaster S, Cummings T, Landells C, Rockwood P (1995) Neer's classification system: a critical appraisal. J Trauma 38:257–260
4. Burstein AH (1993) Fracture classification systems: Do they work and are they useful? J Bone Joint Surg [Am] 75:1743–1744
5. Court-Brown CM, Garg A, McQueen MM (2001) The epidemiology of proximal humeral fractures. Acta Orthop Scand 72:365–371
6. Gjørup T (1988) The kappa coefficient and the prevalence of a diagnosis. Meth Inform Med 27:184–186
7. Horak J, Nilsson BE (1975) Epidemiology of fracture of the upper end of the humerus. Clin Orthop 112:250–253
8. Kiær T, Larsen CF, Blicher J (1986) [Proksimale humerusfrakturer: En epidemiologisk og frakturbeskrivende undersøgelse]. An epidemiological and descriptive investigation of proximal fractures of the humerus. Ugeskr Laeger 148:1984–1987 [in Danish]
9. Kristiansen B, Andersen ULS, Olsen CA, Varmarken JE (1988) The Neer classification of fractures of the proximal humerus: an assessment of interobserver variation. Skeletal Radiol 17:420–422
10. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174
11. Neer CS (1970) Displaced proximal humeral fractures. Part I: classification and evaluation. J Bone Joint Surg [Am] 52:1077–1089
12. Sallay PI, Pedowitz RA, Mallon WJ, Vandemark RM, Dalton JD, Speer KP (1997) Reliability and reproducibility of radiographic interpretation of proximal humeral fracture pathoanatomy. J Shoulder Elbow Surg 6:60–69
13. Sidor ML, Zuckerman JD, Lyon T, Koval K, Cuomo F, Schoenberg H (1993) The Neer classification system for proximal humeral fractures: an assessment of interobserver reliability and intraobserver reproducibility. J Bone Joint Surg [Am] 75:1745–1750
14. Siebenrock KA, Gerber C (1993) The reproducibility of classification of fractures of the proximal end of the humerus. J Bone Joint Surg [Am] 75:1751–1755
15. Sjödén GOJ, Movin T, Güntner P, Aspelin P, Ahrengart L, Ersmark H (1997) Poor reproducibility of classification of proximal humeral fractures: additional CT of minor value. Acta Orthop Scand 68:239–242
16. Sjödén GOJ, Movin T, Aspelin P, Güntner P, Shalabi A (1999) 3D-radiographic analysis does not improve the Neer and AO classifications of proximal humeral fractures. Acta Orthop Scand 70:325–328
17. Svanholm H, Starklint H, Gundersen HJG, Fabricius J, Barlebo, H, Olsen S (1989) Reproducibility of histomorphologic diagnosis with special reference to the kappa statistic. APMIS 97:689–698