The interobserver reliability of classification systems for radial head fractures: the Hotchkiss modification of the Mason classification and the AO classification systems

David M. Sheps, MD*
Krystina R.L. Kiefer, BMSc*
Richard S. Boorman, MD†
John Donaghy, MB, Bch, BAO†
Aleem Lalani, MD*
Richard Walker, MD‡
Kevin A. Hildebrand, MD†

From the *Department of Surgery, University of Alberta, Edmonton, and the Departments of †Surgery and ‡Radiology, University of Calgary, Calgary, Alta.

Presented at the 61st Annual Meeting of the Canadian Orthopaedic Association, June 2–4, 2006, in Toronto, Ont.

Accepted for publication
Mar. 12, 2008

Correspondence to:
Dr. K.A. Hildebrand
Department of Surgery
Health Research Innovation Centre
Rm. 3A 16
3280 Hospital Dr. NW
Calgary AB T2N 4Z6
fax 403 270-3679
hildebrk@ucalgary.ca

Background: Radial head fractures are common injuries, and there is little information on the reliability of classification systems for such injuries. The purpose of our study was to report the interobserver reliability of 2 commonly used classification systems: the Hotchkiss modification of the Mason classification and the AO classification systems.

Methods: We compiled the radiographs from a cohort series of 43 patients with radial head fractures, and 5 observers classified the radiographs according to both classification systems. Additionally, we collapsed the systems, with types II and III combined for the Hotchkiss classification and the final digit dropped for the AO classification. We calculated percent agreement, the κ statistic and the associated 95% confidence intervals (CIs).

Results: The mean percent agreement was 72.3% (95% CI 65.8%–78.9%) for the Hotchkiss classification and 37.7% (95% CI 30.5%–44.9%) for the AO classification. The κ statistic was 0.585 (0.541–0.661) for the Hotchkiss classification and 0.261 (0.240–0.350) for the AO classification. The mean percent agreement was 89.3% (86.6%–92.0%) for the consolidated Hotchkiss classification and 67.4% (54.6%–80.3%) for the consolidated AO classification. The κ statistic was 0.760 (0.691–0.805) for the consolidated Hotchkiss classification and 0.455 (0.372–0.521) for the consolidated AO classification.

Conclusion: The interobserver reliability for the Hotchkiss modification of the Mason classification was moderate, and that for the AO classification was fair according to the criteria of Landis and Koch. Collapsing the Hotchkiss classification improved the reliability to substantial, and collapsing the AO system improved reliability to the lower end of moderate.

Contexte : Les fractures de la tête radiale sont des blessures fréquentes et on dispose de peu de données sur la fiabilité des systèmes de classification s'y rapportant. Notre étude avait pour but de faire rapport sur la fiabilité interobservateurs de 2 systèmes de classification d'usage courant, soit la modification Hotchkiss du système de classification Mason et le système de classification de l'Association pour l'étude de l'ostéosynthèse (AO).

Méthodes : Nous avons compilé les radiographies d'une cohorte en série de 43 patients victimes d'une fracture de la tête radiale. Cinq observateurs ont ensuite catégorisé les radiographies selon les 2 systèmes de classification. Nous avons en outre consolidé les systèmes en regroupant les types II et III de la classification de Hotchkiss et en laissant tomber le dernier chiffre de la classification numérique de l'AO. Nous avons calculé le pourcentage d'accord, le coefficient κ et les intervalles de confiance (IC) à 95 % associés.

Résultats : Nous avons obtenu un pourcentage d'accord moyen de 72,3 % (IC à 95 % 65,8 %–78,9 %) pour la classification de Hotchkiss et de 37,7 % (IC à 95 % 30,5 %–44,9 %) pour celle de l'AO. Nous avons enregistré un coefficient κ de 0,585 (0,541–0,661) pour la classification de Hotchkiss et de 0,261 (0,240–0,350) pour celle de l'AO. Nous avons obtenu un pourcentage d'accord moyen de 89,3 % (86,6 %–92,0 %) pour la classification de Hotchkiss consolidée et de 67,4 % (54,6 %–80,3 %) pour celle de l'AO consolidée. Nous avons noté un coefficient κ de 0,760 (0,691–0,805) pour la classification de Hotchkiss consolidée et de 0,455 (0,372–0,521) pour celle de l'AO consolidée.

Conclusion : À partir des critères de Landis et Koch, nous avons pu qualifier la fiabilité interobservateurs de modérée en ce qui concerne la modification Hotchkiss de la classification de Mason et de passable en ce qui concerne celle de l'AO. La consolidation a amélioré la fiabilité de la classification de Hotchkiss jusqu'à la rendre substantielle, tandis que la consolidation du système de l'AO en a amélioré la fiabilité jusqu'à la limite modérée inférieure.

Radial head fractures are a common orthopedic injury. The classification of this injury attempts to distinguish between fractures that may be treated nonoperatively, fractures that are amenable to open reduction and internal fixation, and fractures that cannot be reconstructed and require prosthetic replacement or excision. The interobserver and intraobserver reliability of the original Mason classification system was found to be unreliable.[1] The reliability of other existing classification systems has not been assessed.

The Hotchkiss[2] modification of the Mason classification and the AO classification[3] systems are used frequently in clinical practice and in the orthopedic literature to classify radial head fractures. The AO classification system is the accepted standard used by the Orthopaedic Trauma Association and the Association for the Study of Internal Fixation.

Hotchkiss[2] modified the original Mason classification of radial head fractures, dividing the fractures into 3 types. Type I fractures include nondisplaced or minimally displaced fractures of the head and neck that do not require operative treatment. Type II fractures include displaced fractures of the head and neck that are amenable to open reduction and internal fixation. Type III fractures include severely comminuted fractures of the head and neck that require prosthetic replacement or excision.

The AO classification system resulted from an attempt to provide a comprehensive classification system for long bone fractures. Fractures of the proximal radius and ulna are divided into 3 types. Type A includes extra-articular fractures of one or both bones of the forearm. Type B includes intra-articular fractures of one bone with or without an extra-articular fracture of the other bone. Type C includes intra-articular fractures of both bones. The fractures are then further subdivided into groups 1, 2 and 3 based on the involvement of the radius, the ulna or both the radius and ulna, and the location of the fracture line. They are then further subdivided into subgroups 0.1, 0.2 or 0.3 based on further specific qualifications such as the degree of fragmentation and the complexity of the articular and/or metaphyseal involvement.

The purpose of our study was to determine whether radial head fractures could be classified reliably among multiple examiners using these 2 classification systems. We used interobserver reliability to gauge whether either classification system was sufficiently reproducible among users for clinicians and researchers to use them to communicate concisely and reliably about radial head fractures.

Methods

We designed the study to assess the reliability of 2 classification systems for radial head fractures using 5 examiners with various levels of experience in the assessment and treatment of upper extremity injuries. Evaluation was undertaken with all observers blinded to patient identity and to the other examiners' ratings.

We identified patients who experienced radial head fractures between January 1999 and February 2004 from the practice of an upper extremity subspecialist orthopedic surgeon (K.A.H.), who did not participate in classifying the fractures, for inclusion in the study. Inclusion criteria were the presence of an acute radial head fracture, skeletal maturity and both an anterior-posterior and lateral radiograph of the elbow at the time of injury or before the start of active treatment. The Calgary Health Region Ethics Board approved our study.

The 5 observers selected to review the radiographs included 4 upper extremity orthopedic surgeons (D.M.S., R.S.B., J.D., A.L.) with differing levels of clinical experience, and a musculoskeletal radiologist (R.W.). The orthopedic surgeons included an upper extremity surgeon with 20 years of orthopedic trauma experience, 2 upper extremity surgeons in practice between 3 and 5 years and an upper extremity clinical orthopedic fellow. The observers had not previously been involved in the care of the patients included in the study and had not previously seen the radiographs.

We compiled the radiographs, and the treating surgeon (K.A.H.) selected the pretreatment anterior–posterior and lateral images. We hid all identifying data on the radiographs to protect patient confidentiality, and we numbered the images. In cases where the radiographs were electronic images, we appropriately magnified them to match the resolution of the original radiographic film images.

Each observer received diagrams illustrating and describing in detail the 2 classification systems being evaluated. We obtained the illustrations and descriptions from the initial descriptions by the original authors.[2,3]

We asked observers to classify the radiographs according to the Hotchkiss modification of the Mason classification and the AO classification systems. The observers classified the radiographs independent of each other. It took 45–60 minutes for the reviewers to go through the radiographs; they classified the radiographs individually, but all in one sitting. All observers reviewed the radiographs in the same order, and they used a standardized data entry sheet to record their classification for each set of radiographs. Prior to classifying the radiographs, we reviewed the classification systems with the observers. Reference sheets detailing the systems were available to the reviewers during the classification of the radiographs. We provided reviewers with rulers to measure displacement.

For the Hotchkiss modification of the Mason classification system, we asked observers to select either type I, II or III fractures. We asked them to ignore other fractures and dislocations apparent in the radiographs while classifying according to the Hotchkiss system.

For the AO classification system, we asked observers to select fracture type A, B or C, followed by group 1, 2 or 3 and finally subgroup 0.1, 0.2 or 0.3. We asked them to include fractures of the proximal ulna while classifying according to the AO system, and depression in this case meant any depression greater than 0 mm, following the description of the original authors.[3]

Statistical analysis

We entered data from each observer into Excel spreadsheets (Microsoft Corporation). We calculated percent agreement pairwise among observers, yielding 10 different results per classification system. We calculated the mean percent agreement and corresponding 95% confidence intervals (CIs) for each classification system using SPSS 11.0 for Windows (SPSS Inc.).

We also calculated the κ statistic for each classification system as a measure of chance-corrected agreement for nominal data. It compares an observed measure of agreement with the level of agreement expected by chance alone. The maximum value of 1.0 means that every observer agrees with every other observer on every case, whereas a value of 0 indicates no more agreement than what would be expected by chance alone.[4,5] We treated each category of the classification system as nominal data. Interpretation of the κ value was based on the guidelines proposed by Landis and Koch:[6] a κ statistic 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. We calculated κ statistic values using Stata 8 for Windows (StataCorp.).

We performed further analysis on consolidated versions of both classification systems. We collapsed the Hotchkiss modification of the Mason system into 2 types. The first consisted of type I fractures, whereas the second consisted of types II and III fractures. This, in effect, made the Hotchkiss modification of the Mason system a nonoperative versus operative classification system. We consolidated the AO classification system by removing the final digit. Thus, we reduced the original number of 26 subgroup categories to 9. Although this did not result in a nonoperative versus operative classification system, it was done to consolidate similar types of fractures and to reduce the complexity of the system. We recalculated percent agreement pairwise among observers, and we recalculated the mean percent agreement and corresponding 95% CIs for each consolidated classification system. Additionally, we calculated κ statistic values for the consolidated systems.

Results

We identified 58 patients as having radial head fractures between January 1999 and February 2004; however, owing to missing or destroyed radiographs, 43 patients provided radiographs for the study. Of these, 12 patients had concomitant proximal ulna fractures. Of the 43 sets of radiographs obtained, 37 sets were original radiographic film images and 6 sets were electronic images, which we magnified to match the resolution of the original radiographic film images.

Each of the 5 observers classified the 43 sets of radiographs. The mean percent agreement was 72.3% (95% CI 65.8%–78.9%) for the Hotchkiss modification of the Mason classification system and 37.7% (95% CI 30.5%–44.9%) for the AO classification system. The κ statistic was 0.585 (95% CI 0.54–0.661) for the Hotchkiss modification of the Mason classification system and 0.261 (95% CI 0.240–0.350) for the AO classification system (Table 1). We considered the κ statistic values to represent moderate and fair agreement for the Hotchkiss modification of the Mason classification and the AO classification systems, respectively, according to the interpretation of the κ statistic by Landis and Koch.[6]

Consolidation of the categories within the classification systems improved the percent agreement and κ statistic values for both classification systems. The mean percent agreement increased from 72.3% to 89.3% (95% CI 86.6%–92.0%) for the Hotchkiss modification of the Mason classification system and from 37.7% to 67.4% (95% CI 54.6%–80.3%) for the AO classification system. The κ statistic was 0.760 (95% CI 0.691–0.805) for the consolidated version of the Hotchkiss modification of the Mason classification system and 0.455 (95% CI 0.372–0.521) for the consolidated version of the AO classification system (Table 2). We considered these values to represent substantial and moderate (lower end) agreement for the Hotchkiss modification of the Mason classification and the AO classification systems, respectively, according to the interpretation by Landis and Koch.[6]

Table 1. Percent agreement and κ statistic of the classification systems

Classification system	Percent agreement (95% CI)	κ statistic (95% CI)
Hotchkiss	72.3 (65.8%–78.9%)	0.585 (0.541–0.661)
AO	37.7 (30.5%–44.9%)	0.261 (0.240–0.350)

CI = confidence interval.

Discussion

The objective of a classification system is to provide a reliable and reproducible means by which clinicians and researchers may communicate about various fractures and dislocations. The purpose of our study was to assess whether 2 common radial head fracture classification systems, the Hotchkiss modification of the Mason classification and the AO classification systems, were reliable among multiple examiners with varying levels of experience. To our knowledge, ours is the first study assessing the interobserver reliability of the Hotchkiss modification of the Mason classification and the AO classification systems for radial head fractures.

We found that the Hotchkiss modification of the Mason classification for radial head fractures had moderate reliability based on the interpretation of the κ statistic by Landis and Koch.[6] This reliability improved to substantial agreement when we consolidated the classification system. The Hotchkiss modification of the Mason system takes into account the suggested treatment in addition to the morphology of the fracture, and it builds upon the original classification system proposed by Mason.[2] Prior work on the reliability of the Mason classification by Morgan and colleagues[1] reported a median κ statistic of 0.54 and 0.64 at 2 separate readings of radiographs by the same observers. Although the authors concluded that the Mason system was unreliable based on their interpretation of their study results, using the interpretation of Landis and Koch, the median κ statistic values could be viewed as showing moderate reliability.

These findings suggest that the Hotchkiss modification of the Mason classification system is substantially successful at differentiating between fractures that require operative treatment and those that do not. Examining the classification of the individual fractures by the 5 observers in the present study revealed that no fracture classified as type I by one reviewer was classified as type III by another. Thus, all fractures that were labelled as type III by at least 1 reviewer would have received operative treatment based on the classification of all of the other reviewers.

Differentiating type I and II fractures requires that we determine whether the fractures are sufficiently displaced to require open reduction and internal fixation; however, a major concern is that the definition of sufficient displacement is not clear in the existing literature. Although, traditionally, fractures with more than 2 mm of displacement would be considered sufficiently displaced, anecdotally, many surgeons treat radial head fractures with more than 2 mm of displacement nonoperatively (Fig. 1).

Previous studies examining the outcomes of Mason I and II fractures of the radial head treated nonoperatively and operatively have produced mixed results. Herbertsson and colleagues[7] recently reviewed the long-term results of nonoperatively treated Mason I radial head fractures with more than 1 mm of displacement. They found a generally favourable outcome, with 91% of patients reporting no symptoms and no measured loss of motion among all patients included in the study. Studies by Arner and colleagues[8] and Poulson and Tophoj[9] found excellent results in 87% and 78% of patients, respectively, with marginal fractures of the radial head treated nonoperatively.

Table 2. Percent agreement and κ statistic of the consolidated classification systems

Classification system	Percent agreement (95% CI)	κ statistic (95% CI)
Consolidated Hotchkiss	89.3 (86.6%–92.0%)	0.760 (0.691–0.805)
Consolidated AO	67.4 (54.6%–80.3%)	0.455 (0.372–0.521)

CI = confidence interval.

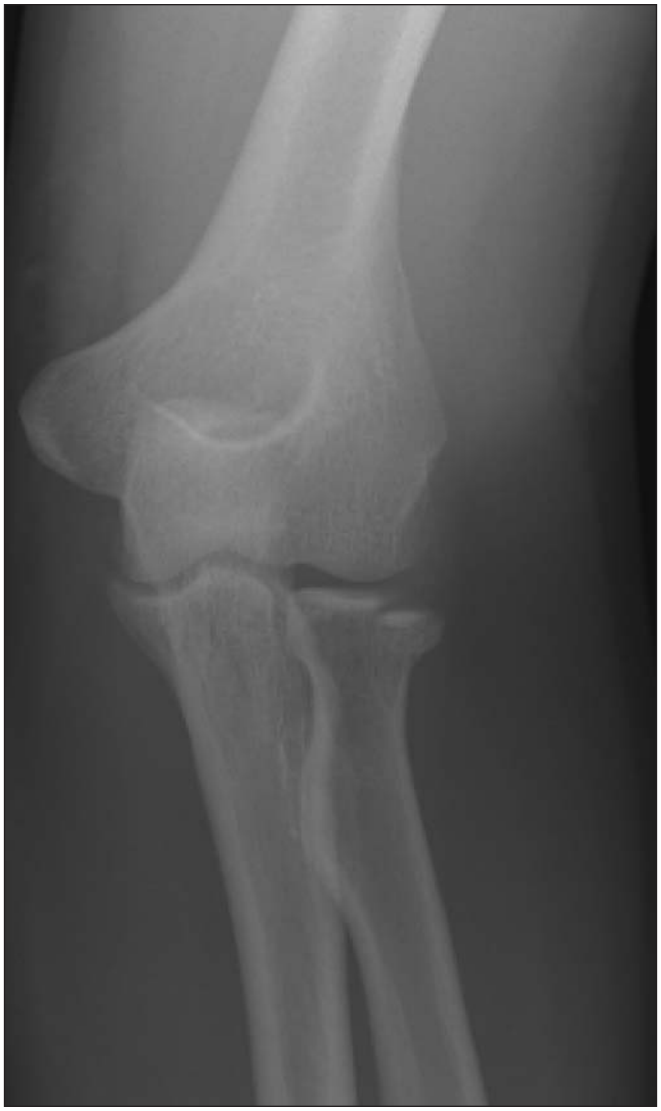

Fig. 1. The Hotchkiss modification of the Mason classification type II radial head fracture.

Conversely, Carstam[10] reported 46% poor or fair outcomes following nonoperative treatment of radial head fractures. Khalfayan and colleagues[11] found improved outcomes following open reduction and internal fixation for patients with Mason type II fractures of the radial head. Prospective studies on radial head fracture treatment and outcomes are required to better define sufficient displacement to require operative treatment.

We found that agreement in the AO classification by subgroup was fair, increasing to the lower end of moderate when we consolidated the system to group classification. Previous studies examining the interobserver reliability of the AO classification system for fractures of the proximal humerus, distal humerus and distal radius reported agreement that ranged from fair to moderate, with most studies finding fair subgroup agreement and moderate group agreement.[12–16]

As with previous studies examining the reliability of the AO classification system for various fractures, our findings suggest that the AO system is also generally unreliable for classifying radial head fractures. The AO system arose from efforts to produce a comprehensive method for classifying fractures of the long bones that could be used as a taxonomic system in research to make valid comparisons among groups.[3,17] However, most studies assessing the reliability of the AO system in various musculoskeletal injuries have found it to be cumbersome, not applicable to clinical decision-making and generally unreliable.[12–16]

A major criticism of the AO classification system specific to radial head fractures is its subgroup classification for such fractures. Isolated radial head fractures could be classified as either type B2.1, B2.2 or B2.3. The differences between these subgroups relate to the degree of comminution first and the degree of displacement second, with subgroup B2.1 including all displaced and nondisplaced simple fractures, subgroup B2.2 including nondisplaced multifragmentary fractures and subgroup B2.3 including displaced multifragmentary fractures. Treatment decisions based on this system may be limited, as nondisplaced simple and displaced simple fractures (both B2.1) may require different interventions. As treatment decisions are likely to be different, one could suppose that the outcomes may be different, which suggests this system may be limited in assisting clinical decision-making or assessing outcomes. A factor contributing to the lower agreement with the AO system and the limitation of its usefulness is the inclusion of ulnar fractures, when present, in the classification. Evaluators may agree on the radial head fracture, but differ on the ulnar fracture, lessening the agreement among observers.

In general, we found that the observers were likely more familiar with the Hotchkiss modification of the Mason classification system than the AO classification, which may have affected our observed reliability. However, it could be argued that this familiarity may be related to the greater ease of use of the Hotchkiss system than that of the AO system, which in turn suggests that the Hotchkiss system is more relevant as a treatment-based classification system. The percentage of displaced and comminuted radial head fractures may have been greater than what would be expected in the general population owing to the subspecialized nature of the practice from which we identified the patients. Thus, the reported interobserver agreement may be decreased as there was greater agreement in classifying nondisplaced fractures.

We did not randomize the order of presentation of the radiographs for each observer. However, the results do not appear to show increasing or decreasing agreement on radiographs reviewed earlier and those reviewed later, arguing against a learning bias or examiner fatigue. Additionally, we did not ask observers to repeat classification to assess the intraobserver reliability of the systems. A gold standard such as a computed tomography scan or operative intervention was not available for all patients, therefore we were not able to determine the accuracy of each reviewer in using the classification systems. The clinical experience of the reviewers did not appear to alter the percent agreement appreciably among the various reviewers. The number of observers and the number of patients is in keeping with similar previously completed studies.[12–16]

In conclusion, the Hotchkiss modification of the Mason classification system has moderate reliability, increasing to substantial reliability following consolidation of the type II and III fractures. The Hotchkiss system appears to be moderately successful in distinguishing between fractures requiring operative versus nonoperative intervention. Further work is required to determine the true applicability of this classification system: that is, whether fractures meeting the operative treatment criteria truly require such an intervention. As with previous studies on the interobserver reliability of the AO classification system for long bone fractures, our study confirms the unreliability of this system. Specifically, for fractures of the proximal radius and ulna, including the radial head, the AO subgroup classification system has only fair reliability. Consolidation of the classification system, which eliminates the ability to distinguish between different types of radial head fractures, increases the reliability to the lower end of moderate. Although this streamlines the system, the clinical relevance becomes limited by the consolidated version's decreased ability to describe injury severity and to identify injuries that require operative treatment.

Acknowledgements: This work was funded by a grant from the Calgary Health Region. Krystina Kiefer was supported by a summer student grant from the Alberta Heritage Foundation, David Sheps was supported by a Canada Graduate Student Masters Award from the Canadian Institutes of Health Research, and Kevin Hildebrand holds a Health Research Foundation – Canadian Institutes of Health Research Career Award.

Competing interests: None declared.

Contributors: Drs. Sheps, Kiefer, Boorman and Hildebrand designed the study. Drs. Sheps, Kiefer, Boorman Donaghy, Lalani and Walker acquired the data, which Drs. Sheps, Kiefer and Hildebrand analyzed. Drs. Sheps and Hildebrand wrote the article, which Drs. Kiefer, Boorman, Donaghy, Lalani and Walker reviewed. All authors approved the final version for publication.

References

1. Morgan SJ, Groshen SL, Itamura JM, et al. Reliability evaluation of classifying radial head fractures by the system of Mason. *Bull Hosp Jt Dis* 1997;56:95-8.
2. Hotchkiss RN. Displaced fractures of the radial head: internal fixation or excision? *J Am Acad Orthop Surg* 1997;5:1-10.
3. Müller ME. *The comprehensive classification of fractures in long bones*. Berlin (Germany): Springer-Verlag; 1990.
4. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960;20:37.
5. Posner KL, Sampson PD, Caplan RA, et al. Measuring interrater reliability among multiple raters: an example of methods for nominal data. *Stat Med* 1990;9:1103-15.
6. Landis JR, Koch G. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159-74.
7. Herbertsson P, Josefsson PO, Hasserius R, et al. Displaced Mason type I fractures of the radial head and neck in adults: a fifteen- to thirty-three-year follow-up study. *J Shoulder Elbow Surg* 2005;14:73-7.
8. Arner O, Ekengren K, Von Schreeb T. Fractures of the head and neck of the radius: a clinical and roentgenographic study of 310 cases. *Acta Chir Scand* 1957;112:115-34.
9. Poulsen JO, Tophoj K. Fracture of the head and neck of the radius. Follow-up on 61 patients. *Acta Orthop Scand* 1974;45:66-75.
10. Carstam N. Fractures of the head and neck of the radius. *Acta Orthop Scand* 1951;19:502-26.
11. Khalfayan EE, Culp RW, Alexander AH. Mason type II radial head fractures: operative versus nonoperative treatment. *J Orthop Trauma* 1992;6:283-9.
12. Andersen DJ, Blair WF, Steyers CM Jr, et al. Classification of distal radius fractures: an analysis of interobserver reliability and intraobserver reproducibility. *J Hand Surg [Am]* 1996;21:574-82.
13. Johnstone DJ, Radford WJ, Parnell EJ. Interobserver variation using the AO/ASIF classification of long bone fractures. *Injury* 1993; 24:163-5.
14. Kreder HJ, Hanel DP, McKee M, et al. Consistency of AO fracture classification for the distal radius. *J Bone Joint Surg Br* 1996;78:726-31.
15. Siebenrock KA, Gerber C. The reproducibility of classification of fractures of the proximal end of the humerus. *J Bone Joint Surg Am* 1993;75:1751-5.
16. Wainwright AM, Williams JR, Carr AJ. Interobserver and intraobserver variation in classification systems for fractures of the distal humerus. *J Bone Joint Surg Br* 2000;82:636-42.
17. Bernstein J, Monaghan BA, Silber JS, et al. Taxonomy and treatment: a classification of fracture classifications. *J Bone Joint Surg Br* 1997;79:706-7, discussion 8-9.