Agreement between Initial Classification and Subsequent Reclassification of Fractures of the Distal Radius in a Prospective Cohort Study

Roderick H. van Leerdam · J. Sebastiaan Souer ·
Anneluuk L. C. Lindenhovius · David C. Ring

Received: 28 March 2009 / Accepted: 22 June 2009 / Published online: 9 July 2009
© American Association for Hand Surgery 2009

Abstract We tested the hypothesis that the original surgeon-investigator classification of a fracture of the distal radius in a prospective cohort study would have moderate agreement with the final classification by the team performing final analysis of the data. The initial post-injury radiographs of 621 patients with distal radius fractures from a multicenter international prospective cohort study were classified according to the Comprehensive Classification of Fractures, first by the treating surgeon-investigator and then by a research team analyzing the data. Correspondence between original and revised classification was evaluated using the Kappa statistic at the type, group and subgroup levels. The agreement between initial and revised classifications decreased from Type (moderate; $K_{\text{type}}=0.60$), to Group (moderate; $K_{\text{group}}=0.41$), to Subgroup (fair; $K_{\text{subgroup}}=0.33$) classifications (all $p<0.05$). There was only moderate agreement in the classification of fractures of the distal radius between surgeon-investigators and final evaluators in a prospective multicenter cohort study. Such variations might influence interpretation and comparability of the

data. The lack of a reference standard for classification complicates efforts to lessen variability and improve consensus.

Keywords Distal radius fractured · Prospective cohort · Classification · Agreement

Introduction

The intra- and interobserver reliability of fracture classifications are typically evaluated by having a few surgeons and surgeons-in-training evaluate radiographic studies and apply a classification. The overall kappa values for interobserver reliability of the classification of fracture of the distal radius according to the Comprehensive Classification of Fractures (CCF) range from slight to moderate agreement 0.37 to 0.60 [1, 6, 7]. Andersen [1] noted decreasing Kappa values from classification at the Type (0.64), Group (0.30), SubGroup (0.25) level. Kreder [7] also described kappa values decreasing with more detailed classification. It is recognized that the observer variability noted in these studies will have an effect primarily on the comparability of various scientific investigations rather than on the care of specific patients [1]. A large prospective cohort study of fractures of the distal radius provided us with an opportunity to compare the classifications of the surgeon-investigators providing the initial care for the patients (and based upon radiographic and intraoperative evaluation) with the consensus reclassifications of a group of researchers collating the data for final analysis. Our hypothesis was that there would be moderate correlation at best between the original surgeon-

R. H. van Leerdam · J. S. Souer · A. L. C. Lindenhovius
Orthopaedic Hand and Upper Extremity Service,
Massachusetts General Hospital,
Boston, MA, USA

D. C. Ring (✉)
Department of Orthopaedic Surgery,
Massachusetts General Hospital,
Yawkey Center for Outpatient Care,
55 Fruit Street, YAW-2-2C,
Boston, MA 02114, USA
e-mail: dring@partners.org

investigator classification and the revised classification by the data analysis team.

Materials and Methods

The immediate post-injury radiographs of 621 patients with distal radius fractures enrolled in a prospective cohort study approved by our Human Research Committee were classified according to the Comprehensive Classification of Fractures [10]. First, by the treating surgeons and study participants (original classification), and then by a research team analyzing the data (revised classification). The research team was initially blinded to the ratings of the original surgeon-investigators.

The Comprehensive Classification of Fractures

The Comprehensive Classification of Fractures, commonly known as the AO classification, was created in 1986 [10]. It is one of the most comprehensive and commonly used classification systems available. The classification is arranged in order of increasing severity according to the complexities of the fracture, difficulty of treatment and worsening prognosis. It consists of three main groups; types A, B, and C. Type A indicates extra-articular fractures; type B, partial articular fractures; and type C, complete articular fractures. Division in fracture type, group, and subgroup results in 27 fracture patterns. Both observers graded every distal radial fracture for three categories: group, type, and subgroup. Diagrams of the classification system were available during each grading.

Statistical Analysis

Correspondence between original and revised classification was evaluated using Cohen's kappa. The kappa coefficient is a commonly used statistic measure of inter-rater agreement for qualitative (categorical) items since its introduction by Cohen in 1960 [1]. The kappa value is a chance-corrected measure of agreement comparing the observed measure of agreement with the level of agreement expected by chance alone [11]. Kappa values have been arbitrarily assigned to subdivisions [8], with values of 0.01 to 0.20 indicating slight agreement; 0.21 to 0.40, fair agreement; 0.41 to 0.60, moderate agreement; 0.61 to 0.80, substantial agreement; and more than 0.81, almost perfect agreement. Significance of difference were considered when upper and lower boundaries of the respective 95% confidence interval did not overlap [5]. Zero indicates no agreement and 1.00 represents perfect agreement. Statistical analysis was performed using the SPSS software package (version 16.0).

Results

According to the original classification by the treating study surgeons there were 250 Type A, 45 Type B, and 290 Type C fractures. According to the revised classification, there were more Type A (261) and type B (49) fracture and fewer Type C (275) fractures (Fig. 1). Interobserver agreement for Type level was moderate ($K_{\text{type}}=0.60$), then decreased when the classification was expanded to Group and Subgroup grading. Interobserver agreement for Group level was moderate ($K_{\text{group}}=0.41$) and Subgroup was fair ($K_{\text{subgroup}}=0.33$). All kappa values were significant ($p<0.05$; Table 1).

Discussion

There are many classification systems for fractures of the distal radius, but few that are as detailed and comprehensive as the Comprehensive Classification of Fractures (CCF). The Comprehensive Classification of Fractures was developed for use in research and based on a large prospectively collected database [10]. As with most classification systems, the CCF was not validated prior to adoption [4, 9]. Unreliable classification systems may bias a study and limit comparability with other studies [2].

This study confirms the Comprehensive Classification of Fractures is most consistent among observers at the Type level, with more limited reliability at the Group and Subgroup levels. The experience in a prospective cohort study was comparable to those with a simulated research cohort [1, 3, 6]. The observation of only moderate agreement in the execution of a prospective multicenter cohort study demonstrates the importance and potential influence of the observer variability in fracture classification.

The strengths of this study include the large number of fractures, classification by multiple experienced surgeons in various parts of the world (greater external validity/reflective of actual practice), and prospective data collection with monitoring by study representatives. Most importantly, we have looked at variations in fracture classification where it may matter most; in the performance of prospective research. Shortcomings include the fact that, in essence, we are comparing only two reviewers or reviewing teams. It is also important to be mindful that we did not have a reference standard for assessing accuracy of classification, so we do not know that the consensus repeat classification was any better than the initial investigator classification.

The lack of a reference standard for classification complicates efforts to lessen variability and improve consensus. Nevertheless, the external validity and repro-

Figure 1 Bar chart showing the observer frequency distribution of group level classification.

ducibility of scientific investigations of fractures should be enhanced by using fracture characteristics or classifications with optimal reliability among observers. It is not clear if the moderate agreement observed with the AO classification in this and other studies is sufficient.

In our opinion, the agreement of different observers will always be limited due to differences in training, experience, and other factors. Rather than attempting to be comprehen-

sive, it may be preferable to adopt the practical and perhaps more realistic goal of developing simple, reliable, clinically meaningful classification systems for use in clinical research.

Acknowledgments The authors did not receive grants or outside funding in support of their research for or preparation of this manuscript. They did not receive payments or other benefits or a commitment or agreement to provide such benefits from a commercial entity

Table 1 Interobserver agreement.

Initial Investigator Classification		Research Group Classification	Difference between observers
A2.1	78	46	32
A2.2	48	55	7
A2.3	14	14	0
A3.1	36	1	35
A3.2	63	134	71
A3.3	11	11	0
B1.1	8	6	2
B2.3	8	5	3
B3.2	16	18	2
B3.3	13	20	7
C1.1	22	3	19
C1.2	23	9	14
C1.3	32	8	36
C2.1	32	31	1
C2.2	62	29	33
C2.3	19	17	2
C3.1	69	139	70
C3.2	26	31	6
C3.3	5	8	3
Total	585	585,0	343

References

1. Andersen DJ, Blair WF, Steyers CM Jr, Adams BD, el-Khoury GY, Brandser EA. Classification of distal radius fractures: an analysis of interobserver reliability and intraobserver reproducibility. *J Hand Surg Am*. 1996;21(4):574–82.
2. Audige L, Bhandari M, Hanson B, Kellam J. A concept for the validation of fracture classifications. *J Orthop Trauma*. 2005;19(6):401–6.
3. Belloti JC, Tamaoki MJ, da Silveira Franciozi CE, Gomes dos Santos JB, Balbachevsky D, Chap CE, et al. Are distal radius fracture classifications reproducible? Intra and interobserver agreement. *Sao Paulo Med J*. 2008;126(3):180–5.
4. Bernstein J. Fracture classification systems: do they work and are they useful? *J Bone Jnt Surg Am*. 1994;76(5):792–3.
5. Doornberg J, Lindenhovius A, Kloen P, van Dijk CN, Zurakowski D, Ring D. Two and three-dimensional computed tomography for the classification and management of distal humeral fractures. Evaluation of reliability and diagnostic accuracy. *J. Bone Jnt Surg Am*. 2006;88(8):1795–801.
6. Illarramendi A, Gonzalez D, V, Segal E, De CP, Maignon G, Gallucci G. Evaluation of simplified Frykman and AO classifications of fractures of the distal radius. Assessment of interobserver and intraobserver agreement. *Int Orthop*. 1998;22(2):111–115.
7. Kreder HJ, Hanel DP, McKee M, Jupiter J, McGillivray G, Swiontkowski MF. Consistency of AO fracture classification for the distal radius. *J. Bone Jnt Surg Br*. 1996;78(5):726–31.
8. Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*. 1977;33(2):363–74.
9. Martin JS, Marsh JL. Current classification of fractures. Rationale and utility. *Radiol Clin North Am*. 1997;35(3):491–506.
10. Muller ME, Nazarian S, Koch P, Schatzker J. The comprehensive classification of fractures of long bones. Berlin: Springer; 1990.
11. Posner KL, Sampson PD, Caplan RA, Ward RJ, Cheney FW. Measuring interrater reliability among multiple raters: an example of methods for nominal data. *Stat Med*. 1990;9(9):1103–15.