Evaluation of the Neer System of Classification of Proximal Humeral Fractures with Computerized Tomographic Scans and Plain Radiographs*

BY JOSEPH BERNSTEIN, M.D.†, LOUIS M. ADLER, M.D.†, JOHN E. BLANK, M.D.†, ROBERT M. DALSEY, M.D.†, GERALD R. WILLIAMS, M.D.†, AND JOSEPH P. IANNOTTI, M.D., PH.D.†, PHILADELPHIA, PENNSYLVANIA

Investigation performed at the University of Pennsylvania School of Medicine, Philadelphia

ABSTRACT: The intraobserver reliability and interobserver reproducibility of the Neer classification system were assessed on the basis of the plain radiographs and computerized tomographic scans of twenty fractures of the proximal part of the humerus. To determine if the observers had difficulty agreeing only about the degree of displacement or angulation (but could determine which segments were fractured), a modified system (in which fracture lines were considered but displacement was not) also was assessed. Finally, the observers were asked to recommend a treatment for the fracture, and the reliability and reproducibility of that decision were measured. The radiographs and computerized tomographic scans were viewed on two occasions by four observers, including two residents in their fifth year of postgraduate study and two fellowship-trained shoulder surgeons. Kappa coefficients then were calculated. The mean kappa coefficient for intraobserver reliability was 0.64 when the fractures were assessed with radiographs alone, 0.72 when they were assessed with radiographs and computerized tomographic scans, 0.68 when they were classified according to the modified system in which displacement and angulation were not considered, and 0.84 for treatment recommendations; the mean kappa coefficients for interobserver reproducibility were 0.52, 0.50, 0.56, and 0.65, respectively. The interobserver reproducibility of the responses of the attending surgeons regarding diagnosis and treatment did not change when the fractures were classified with use of computerized tomographic scans in addition to radiographs or with use of the modified system in which displacement and angulation were not considered; the mean kappa coefficient was 0.64 for all such comparisons. Over-all, the addition of computerized tomographic scans was associated with a slight increase in intraobserver reliability but no increase in interobserver reproducibility. The classification of fractures of the shoulder remains difficult because even experts cannot uniformly agree about which fragments are fractured. Because of this underlying difficulty, optimum patient care might require the development of new imaging modalities and not necessarily new classification systems.

The Neer system commonly is used to classify fractures of the proximal part of the humerus. Nevertheless, the authors of two recent studies questioned the reliability of this system when the classifications are assigned on the basis of plain radiographs[7,8]. Subsequent debate has centered over whether those studies[7,8] invalidate the Neer system or rather illustrate a need to modify its application[1,2,5,6]. Various authors have suggested that perhaps the classification should be assigned only at the time of operative exploration[6], that only experts should use the system[5], or that other imaging modalities should be used to assign the classification[1].

The purpose of the present study was to determine the reliability and reproducibility of the classification of fractures with the Neer system on the basis of computerized tomographic scans and plain radiographs. In addition, we hoped to determine whether the relatively low level of reproducibility that has been associated with the Neer system in previous studies[7,8] was due to difficulty in determining which segments of the proximal part of the humerus were fractured or rather to difficulty in determining whether a segment was sufficiently displaced to be considered a distinct part. To that end, we evaluated a modified system in which every fractured anatomical segment (as described by Neer) was considered to be a distinct fracture fragment even if the criteria for displacement or angulation described by Neer were not met — that is, even if the segment was displaced less than one centimeter or angulated less than 45 degrees. Finally, we attempted to determine whether the decision to explore a fracture operatively in order to obtain an intraoperative classification can be made reliably.

Materials and Methods

Our initial database comprised all proximal humeral fractures for which computerized tomographic scans

*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. No funds were received in support of this study.

†Department of Orthopedic Surgery, University of Pennsylvania School of Medicine, 3400 Spruce Street, Philadelphia, Pennsylvania 19104. E-mail address: orthodoc@mail.med.upenn.edu (Dr. Bernstein).

TABLE I
KAPPA COEFFICIENTS FOR INTRAOBSERVER RELIABILITY

	Over-All	Residents		Attendings	
		1	2	1	2
Radiographs alone	0.64	0.64	0.61	0.73	0.57
Computerized tomographic scan and radiographs					
Standard 16-type Neer classification	0.72	0.73	0.63	0.79	0.73
Modified 6-type Neer classification	0.72	0.70	0.70	0.76	0.70
Instances in which classification based on computerized tomographic scan agreed with that based on radiographs alone	0.81	0.76	0.87	0.84	0.75
Computerized tomographic scan (displacement and angulation ignored)	0.68	0.57	0.73	0.63	0.79
Decision to treat operatively	0.84	0.90	0.90	0.70	0.87

had been made at one of two trauma centers between 1992 and 1994. Additional consideration was given only to those fractures for which at least good axillary and anteroposterior radiographs also were available. The quality of the radiographs was assessed by the senior one us (J. P. I.), who did not serve as an observer in the study. From an initial group of forty-six fractures, twenty were found to have suitable radiographs and therefore were available for classification.

The radiographs and computerized tomographic scans were presented to four observers, on two separate occasions at least eight weeks apart, with use of a method similar to that described by Sidor et al. The observers included two chief residents in their fifth year of postgraduate study (J. B. and J. E. B.) and two fellowship-trained, board-certified shoulder specialists (R. M. D. and G. R. W.) who had extensive experience in the treatment of proximal humeral fractures. All of the observers were familiar with the Neer system. A goniometer, a ruler, and a chart showing the Neer classification were available at each viewing. The observers viewed the radiographs and computerized tomographic scans of each fracture one at a time. The plain radiographs were shown first, and the observer was asked several questions: (1) What segments of the proximal part of the humerus are fractured? (2) What segments of the proximal part of the humerus are fractured and displaced by at least one centimeter or angulated at least 45 degrees? (3) What is the Neer classification for this injury? (4) Would you treat such a fracture operatively? (5) Would you, in a clinical setting, order a computerized tomographic scan for this patient? (The observer was permitted to respond "not certain" to the first three questions.) With the radiographs still available, the observer then was shown the computerized tomographic scan and was asked to answer the first four questions again.

The answers were tabulated, and measures of intraobserver reliability and interobserver reproducibility were obtained for several parameters: the complete Neer classification based on plain radiographs alone; the complete Neer classification based on plain radiographs and computerized tomographic scans; a modified Neer classification that included six types of fractures (one, two, three, and four-part fractures; articular fractures; and fracture-dislocations); a modified complete Neer classification in which any displacement or angulation indicated that a fracture fragment was to be considered a part; and the recommended treatment. We also determined the consensus diagnosis for each fracture (by majority rule), first on the basis of radiographs alone and then on the basis of radiographs as well as computerized tomographic scans. When an observer was not certain of a classification on the basis of radiographs alone, the responses of that observer regarding that fracture were excluded from comparison with those of the other observers. The option to answer "not certain" was explicitly denied to the observers when they classified the fractures on the basis of the computerized tomographic scans. All possible pairs were compared, and means were calculated for each observer.

Reliability and reproducibility were assessed with use of kappa statistics, as described by Dunn. The kappa coefficient represents the percentage of instances of agreement while the likelihood of agreement based on chance alone is taken into account. A kappa coefficient of 1.00 indicates that there is perfect agreement, whereas a kappa coefficient of 0.00 implies no more agreement than would be expected by chance alone. The kappa coefficients were interpreted according to the guidelines described by Landis and Koch: values of less than 0.00 indicate poor agreement; 0.00 to 0.20, 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 0.80, excellent agreement.

Results

The consensus diagnoses that were based on the computerized tomographic scans included four nondisplaced, six two-part, six three-part, and four four-part fractures. Total unanimity — with all four observers assigning the same classification after both viewings — was achieved only three times.

Only four consensus diagnoses changed when the

TABLE II
KAPPA COEFFICIENTS FOR INTEROBSERVER REPRODUCIBILITY

	Over-All	Residents Only	Attendings Only	Best Over-All
Radiographs alone	0.52	0.57	0.64	Resident 1 (0.65)*
Computerized tomographic scan and radiographs				
Standard 16-type Neer classification	0.50	0.41	0.64	Attending 1 (0.59)
Modified 6-type Neer classification	0.54	0.58	0.64	Attending 1 (0.61)
Computerized tomographic scan (displacement and angulation ignored)	0.56	0.49	0.64	Attending 1 (0.64)
Decision to treat operatively	0.65	0.47	0.64	Attending 2 (0.87)

*Resident 1 was "not certain" of the classification of eight of the twenty fractures; this kappa coefficient represents the instances in which a specific diagnosis was chosen.

observers viewed the computerized tomographic scan after seeing the radiographs. The consensus regarding the treatment of three of those four fractures did not change.

After viewing the radiographs alone, the resident observers indicated that they were not certain of the diagnosis a total of nine times; the attending observers never did so. The resident observers said that they would order a computerized tomographic scan on both viewings nine times, and the attending observers did so eleven times.

Intraobserver Reliability (Table I)

Kappa statistical analysis showed substantial intraobserver reliability when the diagnosis was based on radiographs alone; the over-all mean kappa coefficient was 0.64. The mean kappa coefficient increased to 0.72 (also substantial reliability) when computerized tomographic scans were added to the analysis, and it remained at that level when the fractures were classified according to the modified Neer system that included only six types of fractures. The reliability was excellent ($\kappa = 0.81$) when the analysis included only the fractures for which the observer's classification that was based on both the computerized tomographic scan and the radiographs agreed with his observation based on radiographs alone, and the reliability was substantial ($\kappa = 0.68$) when the fractures were classified according to the modified complete Neer system that did not distinguish between displaced and non-displaced fractures. Finally, the decision to treat a given fracture with an operation was associated with excellent reliability ($\kappa = 0.84$).

Interobserver Reproducibility (Table II)

Including only the instances in which the observer had committed to a diagnosis, interobserver reproducibility was moderate ($\kappa = 0.52$) when the diagnosis was based on radiographs alone. Pairwise comparisons demonstrated moderate agreement between the responses of the resident observers ($\kappa = 0.57$) and substantial agreement between those of the attending observers ($\kappa = 0.64$).

Interobserver reproducibility remained moderate ($\kappa = 0.50$) when computerized tomographic scans were added to the analysis. Pairwise comparisons demonstrated only moderate agreement between the responses of the resident observers ($\kappa = 0.41$) but substantial agreement between the responses of the attending observers ($\kappa = 0.64$).

Use of the modified six-category Neer system in which all fractures of a given number of parts were considered the same also did not elevate interobserver reproducibility beyond the moderate range ($\kappa = 0.54$). However, there still was substantial agreement between the responses of the attending observers ($\kappa = 0.64$).

With use of the modified complete Neer system that did not distinguish between displaced and non-displaced fractures, we still noted only moderate reproducibility ($\kappa = 0.56$) between all pairs of observers.

Finally, the decision to treat a given fracture operatively was associated with substantial interobserver reproducibility ($\kappa = 0.65$). (This rate of agreement was, of course, affected not only by the observers' interpretation of the diagnostic images, but also by their understanding of what constitutes the best treatment of a given injury.)

Discussion

Substantial intraobserver reliability ($\kappa = 0.64$) and moderate interobserver reproducibility ($\kappa = 0.52$) were noted when the fractures in the present study were classified on the basis of radiographs alone; these findings are similar to those reported by Sidor et al., who reported kappa values of 0.66 and 0.50, respectively. When computerized tomographic scans were added to the analysis, we found that intraobserver reliability increased slightly but remained in the substantial range ($\kappa = 0.72$) and that interobserver reproducibility remained in the moderate range ($\kappa = 0.50$).

We noted substantial interobserver agreement between the responses of the attending observers when the fractures were classified on the basis of computerized tomographic scans ($\kappa = 0.64$). This kappa value was better than that associated with even the best single pair of viewings in the study by Sidor et al. Nevertheless, there also was substantial agreement among the responses of the attending observers when the classifica-

tion was based on radiographs alone. Thus, the use of computerized tomographic scans did not increase the reproducibility in this group. (As there was only one shoulder specialist in the study by Sidor et al., those authors were not able to evaluate the reproducibility between the responses of specialists.)

Use of the modified Neer system that included only six (as opposed to sixteen) types of fractures did not elevate interobserver reproducibility beyond the moderate range: the mean kappa coefficient increased only slightly, from 0.50 to 0.54. Thus, we infer that there is little to be gained by this modification: less information is conveyed, and there is no marked increase in reproducibility.

Perhaps the most surprising and important finding in the present study is that the reproducibility of the classifications that were based on the separation of segments (without regard to the degree of displacement or angulation) did not exceed the moderate range (κ = 0.56) — that is, a given pair of observers could not agree which segments were involved in nearly one-half of the fractures. Thus, the problem was not related to difficulty in agreeing whether the criteria for displacement were met. This finding implies that any classification system that is based on the anatomical segment schema of Neer cannot exceed moderate reproducibility. The difficulty in applying the Neer system probably is a manifestation of the difficulty that even experts have in interpreting imaging studies of the proximal part of the humerus and is not a conceptual flaw of the system itself. One might argue that surgeons could apply the system with ease if they could be certain which segments were fractured.

We believe that if one wished to devise a new system that could be applied with greater reproducibility, the classifications could not be conceptually based on anatomical segments; the findings of the present study indicate that reliable information regarding the status of anatomical segments is lacking even after experts consult a computerized tomographic scan and adequate trauma radiographs. Thus, the so-called bottleneck for systems in which the classification of proximal humeral fractures is based on anatomical segments is that even experts disagree more than one-third of the time about which segments are fractured.

The purpose of the present study was not to assess the underlying scientific validity of the Neer system. A system has such validity if the categories within it have their own unique natural history, prognosis, or treatment requirements. We did not assess that; we evaluated only the clinical application of radiographs and computerized tomographic scans as tools for employing the system. We found that the system could not be applied with excellent reproducibility, even by experts. This finding in no way vitiates the underlying validity of the system when it is applied correctly. Indeed, while each of the fracture types described by Neer may have its own unique natural history, prognosis, or treatment requirements, computerized tomographic scans and plain radiographs are insufficient to distinguish among them. As Neer stated in his classic article, "[It] is essential that the lesion under consideration be clearly defined."[5] Computerized tomographic scans and plain radiographs often are inadequate for that purpose.

We also did not address the question of whether computerized tomographic scans always are indicated. The consensus regarding the treatment of one of the twenty fractures changed from a non-operative plan to an operative plan after the computerized tomographic scan had been viewed. We infer from this finding that it is possible to underestimate the severity of the fracture on the basis of plain radiographs. The true rate at which computerized tomographic scanning changes the treatment plan can be measured only in a prospective series in which all proximal humeral fractures are evaluated with computerized tomography and the observers are asked to define a treatment plan before and after consulting the scan. (In the present study, we examined only humeral fractures that a clinician had decided to evaluate with a computerized tomographic scan.)

Computerized tomographic scans also may provide information that was not assessed in the present study, such as the presence of abnormalities of the glenoid or the type of fixation needed. Although the attending surgeons in the present study never indicated that they were "not certain" of the classification on the basis of plain radiographs alone, they said that, in a clinical situation, they would order a computerized tomographic scan to evaluate more than one-half of the fractures; this suggests that they used the scans to answer questions other than those regarding the classification of the fracture.

Over-all, the experts in our study classified proximal humeral fractures with substantial reproducibility on the basis of radiographs alone. The reliability of the non-experts was excellent when we considered only the instances in which the classification based on radiographs did not change after the computerized tomographic scan had been viewed. However, when we limited ourselves to the instances when such agreement took place, many of the fractures (more than one-half) were deemed impossible to classify.

We also question whether the guidelines described by Landis and Koch are appropriate for the assessment of fracture classifications. Those guidelines were devised to measure agreement, not accuracy. Ideally, we would like to be able to assess whether classifications made within a system are accurate — that is, whether they are in agreement with some standard. In the absence of such a standard, we must settle for an assessment of the agreement between observers in the hopes that errors will be distributed randomly and that agreement will be achieved only when both observers are correct.

The guidelines described by Landis and Koch may

be perfectly appropriate for assessing agreement in subjective settings in which there is no standard. Nonetheless, those guidelines may be too lenient as a means of assessing fracture classifications: what may be excellent in terms of agreement may be far from excellent in terms of accuracy.

The kappa statistic provides an indirect method of inferring the diagnostic accuracy of each observer. As the calculation of this statistic factors out random matches, a credited match occurs only when both observers make the correct diagnosis. Accordingly, the approximate upper bounds on the mean diagnostic accuracy for two observers is the square root of their rate of agreement (kappa). While a kappa of 0.81 indicates excellent agreement, this level of agreement implies an approximate diagnostic accuracy of only 90 per cent. If the observers agree on the diagnosis only 64 per cent of the time, the level of agreement is termed substantial according to the guidelines described by Landis and Koch but the mean diagnostic accuracy is only about 80 per cent. A kappa of 0.90, implying a diagnostic accuracy of 95 per cent, may be more appropriate for the purpose of assessing fracture classifications.

Classifications that are assigned on the basis of computerized tomographic scans and plain radiographs according to the Neer system are not very reliable or reproducible. This, we have shown, is due to difficulty in determining which segments are fractured. This is not a limitation of the Neer system *per se*, but of any system of classification of proximal humeral fractures that requires identification of the fractured fragments. If the treatment of proximal humeral fractures demands knowledge of which fragments are fractured, a point not addressed in the present study, then new imaging techniques or other diagnostic modalities (for example, a fluoroscopic examination with the patient under anesthesia) must be developed in order to diagnose the fracture correctly. The abandonment of the Neer classification may allow a more reproducible system to emerge. By definition, however, that system will convey less information regarding the anatomy of the fracture: it cannot both be reliable and assess which fragments are fractured, as such information routinely is lacking. Furthermore, to tout that new classification system as reliable may give the false impression that the difficulties that commonly are encountered in the interpretation of diagnostic images of the shoulder have been resolved.

References

1. **Bernstein, J.:** Correspondence. *J. Bone and Joint Surg.*, 76-A: 792-793, May 1994.
2. **Burstein, A. H.:** Fracture classification systems: do they work and are they useful? [editorial]. *J. Bone and Joint Surg.*, 75-A: 1743-1744, Dec. 1993.
3. **Dunn, G.:** *Design and Analysis of Reliability Studies. The Statistical Evaluation of Measurement Errors.* New York, Oxford, 1989.
4. **Landis, J. R.,** and **Koch, G. G.:** The measurement of observer agreement for categorical data. *Biometrics,* 33: 159-174, 1977.
5. **Neer, C. S.:** Displaced proximal humeral fractures: part I. Classification and evaluation. *Clin. Orthop.*, 223: 3-10, 1987.
6. **Rockwood, C. A., Jr.:** Correspondence. *J. Bone and Joint Surg.*, 76-A: 790, May 1994.
7. **Sidor, M. L.; Zuckerman, J. D.; Lyon, T.; Koval, K.; Cuomo, F.;** and **Schoenberg, N.:** The Neer classification system for proximal humeral fractures. An assessment of interobserver reliability and intraobserver reproducibility. *J. Bone and Joint Surg.*, 75-A: 1745-1750, Dec. 1993.
8. **Siebenrock, K. A.,** and **Gerber, C.:** The reproducibility of classification of fractures of the proximal end of the humerus. *J. Bone and Joint Surg.*, 75-A: 1751-1755, Dec. 1993.