Comparison of Hemiarthroplasty and Reverse Shoulder Arthroplasty for the Treatment of Proximal Humeral Fractures in Elderly Patients

Derek J. Cuff, MD, and Derek R. Pupello, MBA

Investigation performed at Suncoast Orthopaedic Surgery and Sports Medicine, Venice, and the Foundation for Orthopaedic Research and Education, Tampa, Florida

Background: Treatment of complex three and four-part proximal humeral fractures with hemiarthroplasty in elderly patients has yielded mixed clinical results. Reverse shoulder arthroplasty has emerged as a treatment option for comminuted proximal humeral fractures for these patients. The purpose of the study was to perform a prospective evaluation of patient outcomes comparing hemiarthroplasty and reverse shoulder arthroplasty for the treatment of comminuted proximal humeral fractures in elderly patients.

Methods: Fifty-three consecutive elderly patients (average age, 74.4 years) underwent an arthroplasty for a complex proximal humeral fracture. Indications for arthroplasty were four-part fractures, three-part fractures with severe comminution of the greater tuberosity, and fractures that involved an articular split of the humeral head. Twenty-six patients underwent hemiarthroplasty (the HA group), followed by twenty-seven patients who underwent reverse shoulder arthroplasty (the RSA group). A total of forty-seven patients (twenty-three in the HA group and twenty-four in the RSA group) were available for follow-up at a minimum of two years.

Results: Final average outcome scores were lower in the HA group than in the RSA group (American Shoulder and Elbow Surgeons [ASES] score of 62 versus 77 \( p = 0.0001 \) and Simple Shoulder Test [SST] of 5.8 versus 7.4 \( p = 0.0062 \)), and patient-reported satisfaction was lower in the HA group than in the RSA group (61% versus 91%; \( p = 0.038 \)). Radiographic healing of the tuberosities occurred in 61% of the patients in the HA group compared with 83% of the patients in the RSA group (\( p = 0.17 \)). Forward elevation of the arm was higher in the RSA group (139°) than in the HA group (100°) (\( p = 0.0002 \)), but no significant differences were observed for shoulder external rotation or internal rotation. Complication rates in both groups were similar. Three patients (13%) in the HA group elected revision to reverse shoulder arthroplasty because of failed tuberosity healing and resultant shoulder pseudoparesis.

Conclusions: In this series, reverse shoulder arthroplasty resulted in better clinical outcomes and a similar complication rate compared with hemiarthroplasty for the treatment of comminuted proximal humeral fractures in the elderly.

Level of Evidence: Therapeutic Level II. See Instructions for Authors for a complete description of levels of evidence.

Proximal humeral fractures are a common injury in the elderly and the third most frequently occurring fractures in patients over the age of sixty-five years\(^2\). Complex three-part and four-part fractures in elderly patients are often not amenable to repair because of poor bone quality, potential loss of fixation, risk of nonunion, or the risk of osteonecrosis, and traditionally are often treated with hemiarthroplasty\(^2\). Hemiarthroplasty for a comminuted proximal humeral fracture remains a technical challenge as the results are largely influenced by tuberosity position and healing, and this treatment has yielded mixed clinical results in elderly patients with respect to functional outcomes\(^4\).
Recently, reverse shoulder arthroplasty has emerged as an alternative option for the treatment of acute, comminuted proximal humeral fractures in elderly patients.2–12. Reverse shoulder arthroplasty is an attractive option in this scenario in that functional outcomes appear less dependent on tuberosity healing and rotator cuff integrity in comparison with hemiarthroplasty. Given that reverse shoulder arthroplasty primarily relies on the deltoid muscle to restore shoulder function, poor tuberosity healing or concomitant rotator cuff issues in the elderly may be circumvented by the use of this implant. The purpose of this study was to perform a prospective evaluation of patient outcomes comparing hemiarthroplasty and reverse shoulder arthroplasty for the treatment of comminuted proximal humeral fractures in elderly patients. Our hypothesis was that reverse shoulder arthroplasty would produce more consistent and superior functional outcomes in this patient population in comparison with hemiarthroplasty.

**Materials and Methods**

**Patient Inclusion Criteria and Demographics**

This study underwent institutional review board approval, and from September 2007 through March 2010, fifty-three consecutive elderly patients with a proximal humeral fracture who were an average of 74.4 years old were enrolled and treated with an arthroplasty performed by a single surgeon. The inclusion criteria were an age of seventy years or older and a four-part fracture, a three-part fracture with severe comminution of the greater tuberosity, or a fracture that involved an articular split of the humeral head. Patients were required to have a minimum of two years of follow-up to be included in the study. Patients had to demonstrate the mental ability to comply with a postoperative protocol, and patients with a diagnosis of dementia were excluded from the study.

The initial twenty-six patients in this series underwent a hemiarthroplasty (the HA group) as treatment for their injury. The subsequent twenty-seven patients in this series underwent a reverse shoulder arthroplasty (the RSA group). Six patients were lost to follow-up, leaving a total of forty-seven patients (the HA group) as treatment for their injury. Morselized bone graft from the humeral head was placed proximally around the humeral socket prior to tuberosity repair to aid in tuberosity healing.

**Outcome Measures and Range-of-Motion Analysis**

All patients were required to complete questionnaires to determine American Shoulder and Elbow Surgeons (ASES) scores and Simple Shoulder Test (SST) scores. Patient-reported satisfaction was recorded at one year postoperatively and at all subsequent follow-up visits, and patients were asked to rate their outcome as satisfactory or not satisfactory.

Patients had their shoulder range of motion digitally recorded at the one-year postoperative evaluation and at all subsequent yearly follow-up examinations. Active shoulder motion was evaluated in forward elevation, external rotation, and internal rotation in a manner that has been previously described. Three observers blinded to which procedure was performed analyzed the last available digital videos of the patients, using a digital goniometer (Screen Protractor; Iconico, New York, NY) in a method we previously described. The measurements of shoulder motion made by the three observers of each patient were averaged and recorded.

**Radiographic Analysis**

All radiographic analysis was performed by an independent musculoskeletal radiologist not involved with the surgical treatment of these patients. Standardized radiographs utilizing a true anteroposterior view, axillary lateral, internal and external rotation views, and a Y-view made at the one-year postoperative visit were used to judge final tuberosity position and healing. At this point, the tuberosities were graded as healed or resorbed. Those deemed to be healed were graded as healed in anatomic position (<5 mm of displacement) or malunited (≥5 mm of displacement). The final available radiographs of each patient made at the time of the latest follow-up were used to assess implant loosening on the humeral side for the HA group. For the RSA group, these radiographs were used to judge implant loosening on the humeral or glenoid side as well as scapular notching. Humeral component loosening was measured using the grading system described by Sperling et al. Glensphere and baseplate fixation was graded in a manner previously described as stable, at risk, or loose. Scapular notching was measured using the grading system of Sirveaux et al.

**Statistical Analysis**

Postoperative outcomes data for the HA and RSA groups were compared using an independent t test. Likewise, outcomes of patients who had a healed tuberosity and those who had a nonhealed tuberosity in both the HA and RSA groups were compared using an independent t test. Differences in proportions were compared using a chi-square test. An a priori power analysis was conducted to find the minimum sample needed to detect a 10-point difference in the ASES score. With alpha = 0.05, power set at 80%, and a standard deviation of 10, a minimum of sixteen patients in each group was needed.

**Results**

The outcome scores, range-of-motion data, and radiographic evaluation for patients included in this study are presented in Table I. Our results demonstrated that the RSA group in comparison with the HA group had significantly better mean ASES (77 versus 62; p = 0.0001) and SST scores (7.4 versus 5.8; p = 0.0062), higher mean forward elevation values (139° versus 100°; p = 0.0002), and a higher percentage...
of patient satisfaction (91% versus 61%; p = 0.038). There were no significant differences between the groups with respect to mean internal or external rotation values.

Tuberosity resorption was observed on the three, six, and twelve-month radiographs with no real changes in resorption occurring after the twelve-month examination. At the one-year postoperative evaluation, the tuberosity healing rate was 61% (fourteen) of twenty-three patients in the HA group with 39% (nine) having complete resorption of one or more tuberosities. There was no evidence of humeral stem loosening at the time of the latest follow-up for any patients in the HA group. In the RSA group, the tuberosity healing rate was 83% (twenty) of twenty-four patients, with 17% (four) having complete resorption of one or more tuberosities. At the time of the latest follow-up, there was no evidence of scapular notching, base-plate loosening, or humeral stem loosening in the RSA group.

Further stratification of the HA and RSA groups on the basis of their radiographic findings are presented in Tables II and III, respectively. These results demonstrate the importance of tuberosity healing in the HA group as those with healed tuberosities had significantly better mean ASES (71 versus 38; p < 0.0001) and SST scores (7.1 versus 2; p < 0.0001), as well as higher forward elevation (131° versus 52°; p < 0.0001) and external rotation values (28° versus 5°; p = 0.009). In contrast, the patients in the RSA group had superior clinical results irrespective of tuberosity healing, with no significant differences

---

**TABLE I Outcome Scores, Range-of-Motion Data, and Radiographic Evaluation for the Study Population**

<table>
<thead>
<tr>
<th></th>
<th>Hemiarthroplasty Group (N = 23)</th>
<th>Reverse Shoulder Arthroplasty Group (N = 24)</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total ASES score*</td>
<td>62 (28-84)</td>
<td>77 (67-82)</td>
<td>0.0001</td>
</tr>
<tr>
<td>ASES pain score*</td>
<td>35.3 (25-45)</td>
<td>40.2 (30-45)</td>
<td>0.0002</td>
</tr>
<tr>
<td>SST score*</td>
<td>5.8 (1-9)</td>
<td>7.4 (6-9)</td>
<td>0.0062</td>
</tr>
<tr>
<td>Satisfied patients†</td>
<td>14 (61)</td>
<td>22 (91)</td>
<td>0.038</td>
</tr>
<tr>
<td>Forward elevation*</td>
<td>100° (30°-170°)</td>
<td>139° (102°-172°)</td>
<td>0.0002</td>
</tr>
<tr>
<td>External rotation*</td>
<td>25° (0°-48°)</td>
<td>24° (8°-42°)</td>
<td>0.88</td>
</tr>
<tr>
<td>Full internal rotation (%)</td>
<td>30</td>
<td>46</td>
<td>0.27</td>
</tr>
<tr>
<td>Healed tuberosities†</td>
<td>14 (61)</td>
<td>20 (83)</td>
<td>0.17</td>
</tr>
<tr>
<td>Tuberosity resorption†</td>
<td>9 (39)</td>
<td>4 (17)</td>
<td>0.23</td>
</tr>
<tr>
<td>Anatomic tuberosity healing†</td>
<td>13 (57)</td>
<td>16 (67)</td>
<td>0.56</td>
</tr>
<tr>
<td>Malunited tuberosity healing†</td>
<td>1 (4)</td>
<td>4 (17)</td>
<td>0.34</td>
</tr>
<tr>
<td>Complications†</td>
<td>2 (9)</td>
<td>2 (8)</td>
<td>0.63</td>
</tr>
<tr>
<td>Revision surgeries†</td>
<td>3 (13)</td>
<td>0 (0)</td>
<td>0.22</td>
</tr>
</tbody>
</table>

*The values are given as the mean, with the range in parentheses. ASES = American Shoulder and Elbow Surgeons, and SST = Simple Shoulder Test. †The values are given as the number of patients, with the percentage in parentheses.

---

**TABLE II Stratification of Hemiarthroplasty Group on the Basis of Radiographic Findings**

<table>
<thead>
<tr>
<th></th>
<th>Healed</th>
<th>Resorption</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of patients</td>
<td>14</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Total ASES score*</td>
<td>71 (59-84)</td>
<td>38 (28-42)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>ASES pain score*</td>
<td>40.1 (30-45)</td>
<td>28.4 (25-35)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>SST score*</td>
<td>7.1 (5-9)</td>
<td>2 (1-3)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Forward elevation*</td>
<td>131° (104°-170°)</td>
<td>52° (30°-83°)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>External rotation*</td>
<td>28° (22°-48°)</td>
<td>5° (0°-12°)</td>
<td>0.009</td>
</tr>
<tr>
<td>Full internal rotation (%)</td>
<td>50</td>
<td>0</td>
<td>0.102</td>
</tr>
</tbody>
</table>

*The values are given as the mean, with the range in parentheses. ASES = American Shoulder and Elbow Surgeons, and SST = Simple Shoulder Test.
Discussion

The use of hemiarthroplasty for the treatment of comminuted proximal humeral fractures in elderly patients has produced mixed clinical results with respect to functional outcomes in multiple studies. Robinson et al. evaluated a large series of 138 patients undergoing hemiarthroplasty as the treatment for proximal humeral fractures and specifically examined the age of the patient in their regression analysis as a predictor of outcome. Of the multiple variables evaluated, they noted that an age of greater than seventy years at the time of surgery had the strongest correlation with a lower Constant score at the one-year follow-up evaluation. Issues with tuberosity bone quality and healing, along with underlying degenerative rotator cuff disease, in many elderly patients may be responsible for the poorer outcomes in these patients. Reverse shoulder arthroplasty becomes an attractive option in this setting in that it relies mainly on deltoid muscle function and may minimize the need for anatomic tuberosity healing. At the same time, given that the deltoid is the prime mover in reverse shoulder arthroplasty, underlying degenerative cuff tears in this elderly population may become less of a concern. These factors have led other authors to investigate the use of reverse shoulder arthroplasty for elderly patients with severely comminuted proximal humeral fractures that are not amenable to open reduction and internal fixation.

Gallinet et al. compared the outcomes of seventeen patients treated with hemiarthroplasty and sixteen patients who had reverse shoulder arthroplasty after a proximal humeral fracture. The authors identified a higher Constant score for the reverse shoulder arthroplasty group; however, the Disabilities of the Arm, Shoulder and Hand (DASH) scores for the two groups were identical.

Boyle et al. utilized the New Zealand Joint Registry to identify fifty-five patients who underwent reverse shoulder arthroplasty for a proximal humeral fracture. The authors identified a higher Constant score for the reverse shoulder arthroplasty group; however, the Disabilities of the Arm, Shoulder and Hand (DASH) scores for the two groups were identical.

In their series, the patients who had reverse shoulder arthroplasty fared better than the hemiarthroplasty subset with respect to outcomes, with a mean ASES score of 81 compared with 47. The patients who

| TABLE III Stratification of Reverse Shoulder Arthroplasty Group on the Basis of Radiographic Findings |
|--------------------------------------------------|-----------------|-----------------|-----------------|
| Healed                                           | Resorption      | P Value         |
| No. of patients                                 |                 |                 |
| 20                                              | 4               |                 |
| Total ASES score*                               | 78 (70-82)      | 75 (74-80)      | 0.344           |
| ASES pain score*                                | 40.4 (35-45)    | 39.2 (30-45)    | 0.588           |
| SST score*                                      | 7.7 (6-9)       | 7.1 (6-9)       | 0.385           |
| Forward elevation*                              | 147° (126°-172°)| 132° (102°-150°) | 0.213           |
| External rotation*                              | 28° (8°-40°)    | 12° (10°-12°)   | 0.02            |
| Full internal rotation (% of patients)           | 50              | 25              | 0.714           |

*The values are given as the mean, with the range in parentheses. ASES = American Shoulder and Elbow Surgeons, and SST = Simple Shoulder Test.
had reverse shoulder arthroplasty also had a substantially higher forward elevation of 122° compared with 90° in the patients who had hemiarthroplasty.

In our study, the RSA group had significantly better ASES and SST scores, as well as active shoulder elevation, than the HA group. Analysis of the outcomes relative to tuberosity healing yielded important findings. In the HA group, the patients in whom the tuberosities healed had relatively good outcome scores. However, 39% (nine) of the patients did not have healing of the tuberosities, resulting in tuberosity resorption. These patients had markedly worse outcomes and a lower satisfaction rating with regard to the surgical procedure. The patients in the HA group had a bimodal distribution with regard to elevation of the arm as previously described by Sirveaux et al.6. The overhead function and elevation in those patients appeared to be an “all or nothing” outcome and was associated with the healing of the tuberosities. In contrast, tuberosity healing was less important in the patients in the RSA group for whom the functional outcomes were equivalent regardless of tuberosity healing. In addition, all of the patients in the RSA group could elevate the arm above shoulder level. Tuberosity healing had an effect on external rotation after RSA. Overall, the patients who had reverse shoulder arthroplasty had a more consistent result with regard to overhead elevation compared with the patients managed with hemiarthroplasty.

Three patients from the HA group elected to undergo revision surgery to a reverse shoulder arthroplasty because of an unsatisfactory result from the primary surgery, while none of the patients in the RSA group required revision. All three patients had tuberosity resorption, with pseudoparalysis of the shoulder and anterior-superior escape of the hemiarthroplasty component. All three patients were satisfied with their outcome after revision surgery; however, the mean outcomes scores and shoulder motion values were inferior to those of the patients who had primary reverse shoulder arthroplasty. This is consistent with the report by Levy et al. and highlights the fact that the revision of a hemiarthroplasty after fracture is a difficult scenario9. The strength of the present investigation is its prospective design along with the fact that the study was appropriately powered on the basis of our power analysis. A weakness of this study was that the patients were not randomized to either treatment arm; rather, a consecutive series of hemiarthroplasties was performed followed by a series of reverse shoulder arthroplasties. This resulted in the HA group having a follow-up interval that was, on the average, ten months longer than that of the RSA group, which could generate some bias in underestimating complications or functional deterioration in the RSA group that could occur with longer follow-up. It must also be noted that preoperative ASES and SST scores were not obtained because the majority of the patients were admitted acutely to the hospital after the shoulder injury; thus, it was not possible to assess preinjury shoulder function.

In summary, the final clinical outcomes and range-of-motion values of these elderly patients who were treated with a hemiarthroplasty for an acute comminuted proximal humeral fracture exhibited a bimodal distribution of good outcomes if tuberosity healing occurred or poor outcomes if their tuberosities underwent resorption. In comparison, the patients who underwent reverse shoulder arthroplasty had more consistent and superior results irrespective of tuberosity healing. Overall, the elderly patients treated with reverse shoulder arthroplasty had better clinical outcomes, better forward elevation, higher tuberosity healing rates, and a lower rate of revision surgery compared with those who had hemiarthroplasty for the treatment of a comminuted proximal humeral fracture.

Appendix
eA A table showing data on demographics, fracture types, and comorbidities of the study population is available with the online version of this article as a data supplement at jbjs.org.

Derek J. Cuff, MD
Suncoast Orthopaedic Surgery and Sports Medicine, 836 Sunset Lake Boulevard, Venice, FL 34292.
E-mail address: Dcuff001@hotmail.com

Derek R. Pupello, MBA
Foundation for Orthopaedic Research and Education, 13020 North Telecom Parkway, Tampa, FL 33637.
E-mail address: dpupello@foreonline.org

References


The Impact of Three-Dimensional CT Imaging on Intraobserver and Interobserver Reliability of Proximal Humeral Fracture Classifications and Treatment Recommendations

Marschall B. Berkes, MD, Joshua S. Dines, MD, Milton T.M. Little, MD, Matthew R. Garner, MD, Grant Daniel Shifflett, MD, Lionel E. Lazaro, MD, David S. Wellman, MD, David M. Dines, MD, and Dean G. Lorich, MD

Investigation performed at the New York-Presbyterian Hospital and the Hospital for Special Surgery, New York, NY

Background: The classification systems for fractures of the proximal part of the humerus provide low interobserver and intraobserver reliability when radiographs or two-dimensional computed tomography scans are used. The purpose of this investigation was to determine whether the use of three-dimensional computed tomography scans could improve interobserver and intraobserver reliability of AO/OTA (Arbeitsgemeinschaft für Osteosynthesefragen/Orthopaedic Trauma Association) and Neer classifications and treatment recommendations.

Methods: Two trauma surgeons, one shoulder surgeon, two senior orthopaedic residents, and two junior orthopaedic residents reviewed the radiographs and two and three-dimensional computed tomography scans of forty fractures of the proximal part of the humerus. Each imaging modality was reviewed in isolation, and fractures were classified according to the Neer and AO/OTA classifications and treatment recommendations were provided. This process was repeated for intraobserver analysis. Interobserver agreement was calculated within and between levels of training for each classification and treatment recommendation with respect to radiographs and two and three-dimensional computed tomography scans.

Results: Among attending orthopaedic surgeons and senior residents, the use of three-dimensional computed tomography did not improve agreement compared with the use of two-dimensional computed tomography for the Neer classification based on planes, the AO/OTA classification, or the treatment recommendation, but it did improve agreement among junior residents. Comparing between levels of training, three-dimensional computed tomography increased agreement only between junior residents and more experienced reviewers for the Neer classification based on planes and for the AO/OTA classification but not for the treatment recommendation. Intraobserver agreement for each reviewer for classification and treatment ranged from slight to fair and was not improved through the use of three-dimensional computed tomography.

Conclusions: In this investigation, the use of three-dimensional computed tomography imaging did not offer improved interobserver and intraobserver agreement compared with the use of two-dimensional computed tomography imaging with regard to classification and treatment of fractures of the proximal part of the humerus, except among reviewers with limited clinical experience.

Level of Evidence: Diagnostic Level III. See Instructions for Authors for a complete description of levels of evidence.

Peer Review: This article was reviewed by the Editor-in-Chief and one Deputy Editor, and it underwent blinded review by two or more outside experts. It was also reviewed by an expert in methodology and statistics. The Deputy Editor reviewed each revision of the article, and it underwent a final review by the Editor-in-Chief prior to publication. Final corrections and clarifications occurred during one or more exchanges between the author(s) and copyeditors.

Disclosure: None of the authors received payments or services, either directly or indirectly (i.e., via his or her institution), from a third party in support of any aspect of this work. One or more of the authors, or his or her institution, has had a financial relationship, in the thirty-six months prior to submission of this work, with an entity in the biomedical arena that could be perceived to influence or have the potential to influence what is written in this work. No author has had any other relationships, or has engaged in any other activities, that could be perceived to influence or have the potential to influence what is written in this work. The complete Disclosures of Potential Conflicts of Interest submitted by authors are always provided with the online version of the article.

Effective treatment of fractures of the proximal part of the humerus begins with appropriate classification of the fracture pattern. For proximal humeral fractures, the Neer and AO/OTA (Arbeitsgemeinschaft für Osteosynthesefragen/Orthopaedic Trauma Association) classifications are the most widely used systems in both clinical and research settings. Despite their widespread use, there are substantial data to suggest that they provide low interobserver and intraobserver reliability when radiographs and two-dimensional (2-D) computed tomography (CT) scans are used. Advancements in CT imaging technology have resulted in the development of reformatted, three-dimensional (3-D) reconstructions. This technology is now frequently used by orthopaedic trauma and shoulder surgeons for the evaluation of fractures of the proximal part of the humerus. There are limits and conflicting data on whether the use of these 3-D CT scans can improve the reliability of the Neer and AO/OTA classifications. The purpose of this investigation was to re-examine the impact of modern, widely available 3-D CT imaging on the interobserver and intraobserver reliability of the Neer and AO/OTA classifications of proximal humeral fractures. Additionally, this study sought to determine whether the use of 3-D CT would impact treatment recommendations. The hypothesis was twofold: (1) 3-D CT will improve interobserver reliability of the Neer and AO/OTA classifications and treatment recommendations for fractures of the proximal part of the humerus among and between attending orthopaedic surgeons and orthopaedic residents with various levels of training, and (2) the use of 3-D CT imaging will result in improved intraobserver reliability of the Neer and AO/OTA classification and treatment recommendation compared with the use of radiographs or 2-D CT imaging.

Materials and Methods

The surgical log at a level-1 trauma center was used to develop a roster of forty cases of proximal humeral fracture. The cases for review were selected in reverse chronological order by an author who was not involved in the radiographic analysis (L.E.L.). Inclusion criteria included the presence of a confirmed fracture of the proximal part of the humerus; radiographs of the injury made in at least two different planes with at least one anteroposterior image; and a CT scan of the injury with 2-D reconstructions in the axial, sagittal, and coronal planes as well as 3-D reconstructions. All CT images were made with a GE scanner at the same hospital. Patients seen in the emergency room (n = 31) had images made with a sixteen-slice scanner, while those seen as outpatients (n = 9) had images made with a sixty-four slice scanner. With regard to the CT scan protocol, slice thickness was 0.625 mm, reformatted to 2.5 mm, with approximately seventy axial images obtained per shoulder; kV was 120, and mA was variable. Coronal and sagittal 2-D reconstructions were done on the scanner, and volume-rendering 3-D image reconstruction was done on either a separate GE Advantage workstation or a TeraRecon. Patients were excluded if the radiographic evaluation for them was incomplete. According to the radiographic review by the senior author (D.G.L.), this population consisted of two one-part fractures, twenty-one two-part fractures, thirteen three-part fractures, and four four-part fractures. The radiographs were reviewed by a diverse team of orthopaedic surgeons, with different numbers of years in training and different subspecialties. The group included two fellowship-trained orthopaedic traumatologists (D.G.L. and D.S.W.), one fellowship-trained shoulder surgeon (J.S.D.), two senior orthopaedic residents (M.B.B. and M.T.M.L.), and two junior orthopaedic residents (M.R.G. and G.D.S.). Each reviewer examined the radiographs using picture-archiving and communication systems in a blinded and randomized fashion. The participants were asked to determine whether fracture lines were present through the greater tuberosity, lesser tuberosity, surgical neck, or anatomic neck and to record the number of parts present in the fracture pattern, as defined by Neer. A modified Neer classification (one, two, three, and four-part fractures; articular fractures; and fracture dislocation) strictly based on the number of parts in each fracture, as well as a modified Neer classification based on the presence of a fracture line regardless of the magnitude of displacement, was assigned to each fracture, similar to the method used by Bernstein et al. The AO/OTA classification was recorded for each fracture. Finally, each reviewer was asked to provide a treatment recommendation for each fracture. Only the images and the age of the patient were provided for use in determining the proposed treatment, with the options being nonoperative treatment, open reduction and internal fixation, and arthroplasty.

The imaging review was repeated in the same manner, this time examining only 2-D CT images in a blinded fashion with a picture-archiving and communication system. Full-rotation 3-D images were then examined in isolation, and the data were recorded in an identical fashion. Reviewers recorded the classification and treatment information as described above for all 2-D and 3-D image sets. No set time between the evaluation of radiographs and the two sets of CT images was required. At a minimum of twenty-eight days after completion of the initial review process, all radiographs and CT scans were re-randomized, and data collection proceeded a second time in the same fashion to allow for intraobserver analysis.

The primary end points for analysis were interobserver reliability of the modified Neer classification according to the number of fracture parts; the modified Neer classification according to fracture lines; and the AO/OTA classification and treatment recommendation when only radiographs were used, when only 2-D CT imaging was used, and lastly when only 3-D CT imaging was used. The influence of the level of training on interobserver reliability with these three different imaging modalities was assessed by comparing interobserver reliability both within and between levels of clinical experience. Intraobserver reliability for each reviewer was calculated for each of the imaging modalities with regard to classification and treatment recommendation.

Statistical analysis was performed by a trained biomedical statistician. Interobserver and intraobserver agreement was determined by performing weighted kappa coefficient calculation. According to Landis and Koch, kappa coefficients <0 indicate no agreement; 0.0 to 0.2, slight agreement; 0.21 to 0.4, fair agreement; 0.41 to 0.6, moderate agreement; 0.61 to 0.8, substantial agreement; and 0.81 to 1.0, almost perfect agreement. P values <0.05 reflect the chance that the interobserver and intraobserver agreement is greater than zero (pure chance alone).

Source of Funding

No external funding was used for this investigation.

Results

Interobserver Agreement

Neer Classification Based on Parts

There was moderate to substantial agreement for the modified Neer classification based on parts among attending orthopaedic surgeons, with weighted kappa coefficients of 0.419 (95% confidence interval [CI], 0.275 to 0.563) for radiographs, 0.672 (95% CI, 0.549 to 0.795) for 2-D CT imaging, and 0.633 (95% CI, 0.499 to 0.767) for 3-D CT imaging. A similar pattern of interobserver agreement was seen between the reviewers who were senior residents. The interobserver reliability was the lowest among the reviewers who were junior residents. However, 3-D images had substantially higher levels of agreement within this level of training compared
Evaluating the AO/OTA classification was fair to substantial. Overall interobserver agreement within levels of training when using imaging modalities did not produce a substantial increase or decrease in agreement among senior-resident reviewers, while 3-D CT imaging generated the highest level of agreement among junior-resident reviewers (Table I). When comparing agreement between levels of training with regard to treatment recommendations, all analyses produced no better than fair agreement (Table II).

### Treatment

Among attending surgeons, radiographs produced the greatest level of agreement (0.658 [95% CI, 0.310 to 1.0]), while 3-D CT yielded the lowest level of agreement (0.423 [95% CI, 0.083 to 0.763]), for treatment recommendations. The different imaging modalities did not produce a substantial increase or decrease in agreement among senior-resident reviewers, while 3-D imaging generated the highest level of agreement among junior-resident reviewers (Table I). When comparing agreement between levels of training with regard to treatment recommendations, all analyses produced no better than fair agreement (Table II).

### Neer Classification Based on Fracture Planes

The modification of the Neer classification based on the presence of fracture planes resulted in a general modest increase in interobserver agreement. Two-dimensional CT yielded the highest agreement among attending-surgeon and senior-resident reviewers. Among the junior-resident reviewers, 3-D CT produced the highest level of agreement compared with the levels for radiographs or 2-D CT (Table I).

When evaluating interobserver agreement between levels of training, the highest levels of agreement were again seen between attending surgeons and senior residents with the use of 2-D CT (0.658 [95% CI, 0.568 to 0.747]) and 3-D CT (0.625 [95% CI, 0.536 to 0.714]). Three-dimensional CT provided higher levels of agreement between junior residents and reviewers with more clinical experience than did 2-D CT (Table II).

### AO/OTA Classification

Overall interobserver agreement within levels of training when evaluating the AO/OTA classification was fair to substantial (0.281 to 0.713). The highest levels of agreement among attending surgeons and senior residents was with 2-D CT imaging, while junior residents had the highest agreement when 3-D CT imaging was used (Table I).

Between levels of training, the highest levels of agreement were between attending surgeons and senior residents for all imaging modalities, with 2-D CT producing agreement superior to that with 3-D CT or radiographs. When comparing junior residents with reviewers who have more clinical experience, the highest levels of agreement were seen with 3-D CT (Table II).

### Neer Classification Based on Fracture Planes

The modification of the Neer classification based on the presence of fracture planes resulted in a general modest increase in interobserver agreement. Two-dimensional CT yielded the highest agreement among attending-surgeon and senior-resident reviewers. Among the junior-resident reviewers, 3-D CT produced the highest level of agreement compared with the levels for radiographs or 2-D CT (Table I).

When evaluating interobserver agreement between levels of training, the highest levels of agreement were again seen between attending surgeons and senior residents with the use of 2-D CT (0.658 [95% CI, 0.568 to 0.747]) and 3-D CT (0.625 [95% CI, 0.536 to 0.714]). Three-dimensional CT provided higher levels of agreement between junior residents and reviewers with more clinical experience than did 2-D CT (Table II).

### AO/OTA Classification

Overall interobserver agreement within levels of training when evaluating the AO/OTA classification was fair to substantial (0.281 to 0.713). The highest levels of agreement among attending surgeons and senior residents was with 2-D CT imaging, while junior residents had the highest agreement when 3-D CT imaging was used (Table I).

Between levels of training, the highest levels of agreement were between attending surgeons and senior residents for all imaging modalities, with 2-D CT producing agreement superior to that with 3-D CT or radiographs. When comparing junior residents with reviewers who have more clinical experience, the highest levels of agreement were seen with 3-D CT (Table II).

### Treatment

Among attending surgeons, radiographs produced the greatest level of agreement (0.658 [95% CI, 0.310 to 1.0]), while 3-D CT yielded the lowest level of agreement (0.423 [95% CI, 0.083 to 0.763]), for treatment recommendations. The different imaging modalities did not produce a substantial increase or decrease in agreement among senior-resident reviewers, while 3-D imaging generated the highest level of agreement among junior-resident reviewers (Table I). When comparing agreement between levels of training with regard to treatment recommendations, all analyses produced no better than fair agreement (Table II).

### Intraobserver Reliability

For all levels of training, the intraobserver agreement for each reviewer using the Neer classification based on parts or fractures planes was only slight to fair (–0.04 to 0.492). Intraobserver reliability was not improved through the use of 3-D CT imaging. The intraobserver agreement for each reviewer using the AO/OTA classification was consistently in the moderate range, again regardless of imaging modality or level of experience. Internal consistency regarding treatment recommendation was highly variable between individuals, with kappa coefficients ranging between 0.172 and 1.0 (see Appendix).
The main purpose of this investigation was to determine whether the addition of 3-D CT imaging improved agreement between orthopaedic surgeons with various levels of experience when classifying proximal humeral fractures according to the Neer and AO/OTA classification systems and making treatment recommendations. Overall, interobserver agreement was moderate to substantial. Three-dimensional CT did not improve agreement over 2-D CT among attending surgeons and senior residents, but it did improve agreement among junior residents. Additionally, the use of 3-D CT consistently increased agreement between junior-resident evaluators and more senior evaluators. While 3-D CT imaging was useful in increasing agreement on fracture classification, it did not have the same effect on treatment recommendation. Three-dimensional CT imaging did not improve intraobserver reproducibility.

Several studies have previously examined the reliability of systems for classifying fractures of the proximal part of the humerus without 3-D imaging. A low level of interobserver agreement was first reported by Kristiansen et al. Using fifty radiographs evaluated by a shoulder specialist, trauma surgeon, radiologist, junior resident, and senior resident, Sidor et al. reported interobserver kappa coefficients between 0.48 and 0.52 when analyzing the formal Neer classification. Reproducibility ranged from 0.83 to 0.50. Simplification of the Neer classification system did not seem to improve agreement or reproducibility. Siebenrock and Gerber reported interobserver reliability kappa coefficients of 0.40 for the Neer system and 0.53 for the basic AO classification, which only worsened with increasing levels of subclassification, as dictated by the AO. The addition of a third perpendicular radiograph did not improve agreement or reproducibility. Mahadeva et al. reported the highest interobserver reliability (kappa coefficient range, 0.61 to 0.80) for the Neer classification using a digital radiograph-viewing station. Other studies of interobserver reliability for the Neer classification (based on radiographs alone) have reported modest kappa coefficients of 0.27, 0.30, 0.33, and 0.36, which are more in line with the findings of Sidor et al. and Siebenrock and Gerber. Majed et al. also reported on interobserver reliability of the AO classification with radiographs, with a kappa coefficient of 0.11.

Given the modest levels of agreement between surgeons for classifications of the proximal part of the humerus, some authors have advocated for the use of advanced imaging to provide more detailed fracture analysis and more accurate classification. Bernstein et al. used both 2-D CT and radiographs to classify twenty fractures of the proximal part of the humerus and found no improved interobserver reliability compared with when only radiographs were used (0.50 with radiographs and CT compared with 0.52 with radiographs alone). Modifying the Neer system to account for only the presence of a fracture line produced a modest increase in reliability (0.56). Sjödén et al. also failed to show a dramatic increase in interobserver reliability through the use

<table>
<thead>
<tr>
<th>Imaging Modality</th>
<th>Neer Parts</th>
<th>Neer Fracture Planes</th>
<th>AO/OTA</th>
<th>Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radiographs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attending surgeon vs. senior resident</td>
<td>0.499 (0.385-0.612)</td>
<td>0.610 (0.504-0.716)</td>
<td>0.521 (0.433-0.608)</td>
<td>0.327 (0.125-0.529)</td>
</tr>
<tr>
<td>Attending surgeon vs. junior resident</td>
<td>0.468 (0.354-0.583)</td>
<td>0.534 (0.426-0.643)</td>
<td>0.498 (0.404-0.592)</td>
<td>0.151 (0.046-0.255)</td>
</tr>
<tr>
<td>Senior resident vs. junior resident</td>
<td>0.310 (0.189-0.430)</td>
<td>0.417 (0.298-0.536)</td>
<td>0.430 (0.334-0.526)</td>
<td>0.309 (0.188-0.431)</td>
</tr>
<tr>
<td>2-D CT scans</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attending surgeon vs. senior resident</td>
<td>0.605 (0.507-0.703)</td>
<td>0.658 (0.568-0.747)</td>
<td>0.638 (0.568-0.708)</td>
<td>0.328 (0.139-0.517)</td>
</tr>
<tr>
<td>Attending surgeon vs. junior resident</td>
<td>0.365 (0.246-0.485)</td>
<td>0.297 (0.178-0.416)</td>
<td>0.376 (0.275-0.476)</td>
<td>0.339 (0.166-0.513)</td>
</tr>
<tr>
<td>Senior resident vs. junior resident</td>
<td>0.280 (0.166-0.393)</td>
<td>0.323 (0.210-0.435)</td>
<td>0.383 (0.283-0.483)</td>
<td>0.267 (0.092-0.442)</td>
</tr>
<tr>
<td>3-D CT scans</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attending surgeon vs. senior resident</td>
<td>0.542 (0.437-0.648)</td>
<td>0.625 (0.536-0.714)</td>
<td>0.557 (0.470-0.644)</td>
<td>0.252 (0.045-0.460)</td>
</tr>
<tr>
<td>Attending surgeon vs. junior resident</td>
<td>0.538 (0.434-0.642)</td>
<td>0.420 (0.313-0.527)</td>
<td>0.479 (0.391-0.567)</td>
<td>0.372 (0.196-0.548)</td>
</tr>
<tr>
<td>Senior resident vs. junior resident</td>
<td>0.480 (0.380-0.580)</td>
<td>0.454 (0.350-0.558)</td>
<td>0.502 (0.418-0.585)</td>
<td>0.285 (0.117-0.453)</td>
</tr>
</tbody>
</table>
of 2-D CT compared with reliability through the use of radiographs.

Because of the failure of 2-D CT imaging to provide improved reliability of classification of proximal humeral fractures, interest has turned toward 3-D CT imaging as a means of evaluating such fractures. Three-dimensional imaging and its impact on classification has been investigated for other fracture types and pathology, including fractures of the calcaneus, tibial plateau, glenoid, distal part of the humerus, coronoid, and radial head. Its application to the proximal part of the humerus was first examined by Sjödén et al., in 1999. They found that 3-D CT provided no benefit when used with radiographs and 2-D CT, with kappa coefficients for interobserver reliability of 0.49 and 0.34 for the Neer and AO classifications, respectively. Similar findings were reported in a more recent study. Overall, neither 2-D nor 3-D CT imaging led to substantially higher interobserver reliability compared with radiographs alone (range, 0.03 to 0.57). However, they did report that advanced imaging generated increased agreement in more complex fracture patterns and among upper-extremity specialists. To our knowledge, increased interobserver agreement with use of 3-D CT has been reported in only one study, by Brunner et al. Special viewing stations were used to provide stereo-visualization of volume-rendered CT images of forty fractures of the proximal part of the humerus. Three-dimensional CT imaging led to increased interobserver reliability compared with radiographs and conventional 2-D CT for both the AO classification (0.64 compared with 0.39 and 0.48, respectively) and the Neer classification (0.80 compared with 0.48 and 0.58, respectively).

Our kappa coefficients for interobserver reliability for the modified Neer classification system based on parts (range, 0.23 to 0.0499), the modified Neer classification system based on fracture lines (range, 0.231 to 0.610), and the AO/OTA classification system (range, 0.43 to 0.521) with use of radiographs are consistent with those in the literature. This provides credibility for the remainder of our analysis, in that our reviewers performed similarly to those in other published studies. Unlike in most prior studies, in the present study 2-D CT did provide an increase in agreement for the modified Neer classifications based on parts and fracture lines among and between attending orthopaedic surgeons and senior orthopaedic residents. The same effect on agreement was not found among the junior residents and between junior residents and senior residents or attending surgeons. The relatively high agreement seen when CT scans are used can possibly be explained by the fact that a substantial number of the fractures evaluated had involvement of the tuberosity and were three or four-part fractures. Data suggest that CT can more accurately define these complex injuries and generate increasing agreement on fracture classification compared with radiographic assessment alone. In this study, the use of 3-D CT did not confer any additional benefit among our attending-surgeon and senior-resident reviewers compared with the agreement when they used 2-D CT. However, 3-D CT did outperform radiographic analysis. Junior residents did achieve substantially improved agreement among themselves and between those more experienced with 3-D CT. This may be because 3-D images provide better spatial understanding of the fracture pattern for individuals who have less experience interpreting 2-D radiographs and CT images. With that in mind, a 3-D image provides only visualization of the surface of the bone. Fracture planes deep to the cortex cannot be seen as well on 3-D CT as on 2-D CT. This can in part possibly explain why 3-D CT conferred no additional advantage among attending surgeons and senior residents.

In this study, we also examined the impact of imaging modality on interobserver treatment recommendation. Overall, neither 2-D CT nor 3-D CT conferred increased agreement in treatment recommendation. Other studies have similarly found that 2-D or 3-D CT does not appreciably change agreement between individuals compared with when only radiographs are used for treatment recommendations. Caution should be practiced when interpreting this result. This does not necessarily mean that CT imaging does not influence or aid in evaluation and treatment of fractures of the proximal part of the humerus. Rather, this result should be interpreted as meaning that the addition of CT does not increase the agreement between different physicians in their treatment recommendation of choice. This finding creates an opportunity for potential health-care cost savings for the patient and society if 3-D CT is used only sparingly or is avoided all together.

Based on these data, advanced imaging is more likely to generate more uniform answers regarding classification than regarding treatment recommendation. This less than excellent agreement between orthopaedic surgeons with regard to the cornerstone classifications of fractures of the proximal part of the humerus is troubling. First, it casts doubt on prior and future research that categorizes and stratifies patient outcomes based on these classifications, with a lack of homogeneity in the study populations and potential mischaracterization of results. The fact that two surgeons cannot predictably agree on how to classify a common fracture suggests the need for modification of the existing schemes or development of an entirely new classification scheme. Modifications of the Neer classification made in an effort to streamline this system have not substantially increased interobserver agreement. Other classification schemes have been proposed and described, but they have not been widely adopted. Thus, until a new classification scheme is accepted into the orthopaedic mainstream or new imaging modalities become available that increase agreement between surgeons using existing classification schemes, the past, current, and future literature should be interpreted with appropriate skepticism.

There are limitations to this investigation. First, fractures were chosen for analysis based on a surgical log, with a heavy bias toward complex fractures with involvement of the tuberosity and neck. The inclusion of nondisplaced fractures within the cohort could have potentially increased agreement between and among observers; conversely, the inclusion of predominantly complex fracture patterns may have negatively influenced our results. Second, only two or three reviewers were included within
each level of training. Agreement within and between these tiers of training is based on only a few individuals and is thus prone to sampling bias. Lastly, all imaging modalities were used in isolation to characterize fracture patterns rather than using radiography in combination with 2-D and 3-D CT imaging in aggregate. Perhaps the agreement would have increased more if all images had been viewed together, thus producing an additive effect. However, the analysis performed in this study involved the evaluation of radiographs, 2-D CT scans, and 3-D CT scans as singular entities. This is not realistic in clinical practice, where the surgeon would use all images available in combination to evaluate these injuries.

Conclusions

The use of 3-D CT imaging did not prove superior to the use of 2-D CT with regard to the classification of proximal humeral fractures according to the Neer and AO/OTA systems, except among reviewers with limited clinical experience. Neither 2-D CT nor 3-D CT use resulted in increased agreement on treatment recommendation compared with the use of radiographs alone. Despite the advantages that 3-D CT has for the evaluation of fracture patterns and preoperative planning, this modality does not eliminate the lack of agreement in classifications between surgeons that is seen with the Neer and AO/OTA classifications of proximal humeral fractures.

Appendix

Tables showing the values for intraobserver reliability of all reviewers for all classifications and treatment recommendations, by modality, are available with the online version of this article as a data supplement at jbjs.org.

References