CHAPTER 5

Interobserver Reliability of Coronoid Fracture Classification: Two-Dimensional vs. Three-Dimensional Computed Tomography

Lindenhovius AL, Karanicolas PJ, Bhandari M, van Dijk CN, Ring D, and the COAST Collaborative*.

In press, J Hand Surg Am

Abstract

Purpose: This study tests the hypothesis that three-dimensional computed tomography reconstructions (3D-CT) improve interobserver agreement on classification and treatment of coronoid fractures when compared to two-dimensional computed tomography (2D-CT).

Methods: Twenty-nine orthopaedic surgeons evaluated ten coronoid fractures on two occasions (first with radiographs and 2DCT and then with radiographs and 3D-CT), separated by a minimum of two weeks. Surgeons classified fractures according to Regan & Morrey and O’Driscoll classifications, identified specific characteristics, recommended the most appropriate treatment approach, and made treatment recommendations. The kappa multirater measure ($\kappa$) was calculated to estimate agreement between observers.

Results: Regardless of the imaging modality used, there was fair to moderate agreement for the majority of observations. Three-dimensional CT improved interobserver agreement in Regan and Morrey ($\kappa_{3D} = 0.51$ vs. $\kappa_{2D} = 0.40$; $p < 0.001$) and O’Driscoll classifications ($\kappa_{3D} = 0.48$ vs. $\kappa_{2D} = 0.42$; $p = 0.009$). There were trends towards better reliability for 3D reconstruction in recognition of coronoid tip fractures ($\kappa_{3D} = 0.19$, $\kappa_{2D} = 0.03$; $p = 0.268$), comminution ($\kappa_{3D} = 0.41$ vs. $\kappa_{2D} = 0.29$; $p = 0.133$), impacted fragments ($\kappa_{3D} = 0.39$ vs. $\kappa_{2D} = 0.27$; $p = 0.0.094$), and in surgeons’ opinions on the need for something other than screws and/or plate for operative fixation ($\kappa_{3D} = 0.31$ vs. $\kappa_{2D} = 0.15$; $p = 0.138$). Interobserver agreement on treatment approach was better with 2D-CT ($\kappa_{3D} = 0.27$, $\kappa_{2D} = 0.32$; $p = 0.015$).

Conclusions: Three-dimensional CT reconstructions improve interobserver agreement with respect to fracture classification when compared to 2D-CT.

Level of Evidence: Level III, Diagnostic Study

Introduction

Recent studies have emphasized the importance of adequate coronoid fracture treatment in the setting of traumatic elbow instability.\textsuperscript{1-6} Recognition of size, location, and fracture morphology of the coronoid process may facilitate operative treatment. However, coronoid fractures are often part of complex elbow fracture-dislocations and can be difficult to identify and characterize on standard radiographs.\textsuperscript{1,7,8} This may result in inadequate treatment with less favorable results.\textsuperscript{6,8} Previous studies have established that two-dimensional computed tomography (2D-CT) scans and three-dimensional computed tomography (3D-CT) reconstructions are a useful adjunct in the evaluation of complex distal humeral fractures and fractures at other sites.\textsuperscript{9-12} It is felt that three-dimensional images are easier for a non-radiologist to interpret.\textsuperscript{9,13,14} Does 3D-CT improve interobserver reliability in the evaluation of coronoid fractures as compared to 2D-CT? This study tests the hypothesis that 3D-CT reconstructions improve interobserver agreement on classification, identification of specific
fracture characteristics, and treatment of coronoid fractures when compared to conventional 2D-CT scans. In a second analysis, we evaluated intra-observer reliability.

**Material and Methods**

**Study Design**

Orthopaedic surgeons from several countries were invited to evaluate ten cases of coronoid fractures in a two-round online survey: one round consisted of an evaluation based on radiographs and 2D-CT scans, and the other round consisted of an evaluation of the same cases based on radiographs and 3D-CT reconstructions. The study was performed under a protocol approved by the Institutional Research Board at our hospital.

This study was performed in cooperation with the international research group “COAST” (Collaboration for Outcome Assessment in Surgical Trials), founded in 2006 by a group of surgeons and clinical epidemiologists. The objectives of the collaborative are to foster international collaboration and to improve the process of outcomes assessment in surgical trials. The COAST group has created a web-based platform that facilitates large international outcome assessment and interobserver studies. With multiple surgeons from diverse countries and institutions participating in studies, this approach should result in improved reliability, validity and generalizability of results.

**Observers**

A total of 133 surgeons (including the existing members of the collaborative and new invitees) were invited by an e-mail that included a short study description. Other than an acknowledgement in the paper, no incentives were provided. Fifty-six surgeons that were interested in participation were e-mailed a username and password to log on to the website. They were randomized to a first round with either radiographs and 2D-CT scans or radiographs and 3D-CT reconstructions (Windows Excel by Microsoft, Redmond, WA). Two weeks after completion of the first round, observers received a reminder to complete the second evaluation of the same ten cases, now based on the other imaging modality. For both rounds, weekly reminders to complete the online survey were e-mailed with a maximum of six reminders. This study presents an analysis of the twenty-nine observers that completed both rounds.

**Fractures**

Radiographs and computed tomography scans of thirty coronoid fractures were blinded for identity by an independent research fellow for use in this study. Two of the authors (one subspecialty trained upper and extremity surgeon and one research fellow in upper extremity trauma) selected ten cases that had coronoid fractures of different size, morphology, and
location representing the known spectrum of traumatic elbow instability with coronoid fracture. Only computed tomography scans that were of adequate quality to create 3D-CT reconstructions (i.e. slice thickness of less than 1.25 mm) were selected. The 3D-CT reconstructions were created with use of Vitrea imaging software (Vital Images, Minnetonka, Minnesota). For each case, videos with 2D-CT and 3D-CT images along the sagittal and axial axes were created. The 3D-CT videos included a reconstruction of the entire elbow and a reconstruction with the distal humerus subtracted. Observers could scroll through the videos or play them automatically. Radiographs, 2D-CT scans, and 3D-CT reconstructions were uploaded to the research group’s website. (Figure 1)
Evaluation

Upon login to the website, surgeons were asked to provide demographic and professional information: (1) gender, (2) location of practice, (3) years in independent practice, (4) training of surgical trainees, (5) number of coronoid fracture cases treated per year, and (6) clinical specialty.

Subsequently, observers were asked to classify the fractures according to two classification systems: the one described by Regan and Morrey\(^\text{15}\) (figure 2) and the one suggested by O’Driscoll\(^\text{16}\) (figure 3). In addition, fractures were evaluated for the presence of five fracture characteristics: 1) fracture of the anteromedial facet, 2) fracture of the tip of the coronoid, 3) comminution of the fracture, 4) impacted articular fragments, and 5) subluxation or dislocation. Observers were provided with the original description and corresponding images of the classification systems. Then, observers were asked about management: 1) nonoperative, 2) Open Reduction and Internal Fixation (ORIF) through a lateral exposure, 3) ORIF through a medial exposure, and 4) ORIF through the olecranon fracture. Finally, observers were asked whether they agreed or disagreed with each of the two following statements: 1) some of the fracture fragments cannot be reliably repaired with screw and will require small wires, sutures, or another technique, and 2) the fracture fixation will be tenuous and should be protected with a hinged external fixator. All questions had to be completed in order to continue with the next case. Lastly, after completing all ten cases, surgeons were given the possibility to record their comments on the exercise.

Statistical Analysis

The kappa multirater measure (κ) was used to estimate agreement among surgeons with respect to fracture classification, fracture characteristics, treatment approach, and treatment statements. It is a commonly used statistic to describe chance-corrected agreement in a variety of intra-observer and interobserver studies.\(^\text{17-19}\) Intra-observer agreement was evaluated using Cohen’s kappa\(^\text{17}\), while agreement among observers was calculated with use of the multirater kappa measure described by Siegel and Castellan.\(^\text{20}\) We interpreted kappa values using the guidelines proposed by Landis and Koch\(^\text{21}\): values of 0.01 to 0.20 indicate 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.

Figure 1 (left). Images from a representative case.
A: Lateral radiographs of a complex coronoid base fracture as part of a posterior olecranon fracture-dislocation.
B: Two-dimensional CT image.
C: Three-dimensional CT image with distal humerus.
D: Three-dimensional CT image, distal humerus subtracted.
Zero indicates no agreement beyond that expected due to chance alone, – 1.00 means total disagreement, and + 1.00 represents perfect agreement.\textsuperscript{18, 21} The agreement based on 2D-CT scans was compared to the agreement based on 3D-CT reconstructions using Z-tests, which assume that the two samples are independent. Since the samples compared in this study were not independent (the same surgeons rated 2D and 3D images), this method produced conservative estimates of the p-values.

The study was powered at 80\% for detecting a kappa difference of 0.1 between results based on 2D-CT vs. 3D-CT using the O’Driscoll classification system. Statistical analysis was performed with SPSS software (version 16.0; SPSS, Chicago, Illinois).

Results

Surgeons

All surgeons that participated in this study were men. Fifteen had their practice located in the United States, six in Continental Europe, three in Canada, two in Australia, two in Central America, and one in Asia. Seven surgeons had been in independent practice for fewer than 5 years, six for 5 to 10 years, twelve for 11 to 20 years, and four for 21 to 30 years. All but one observer supervised surgical trainees in the operating room. Thirteen surgeons operated on zero to 5 coronoid fracture cases per year, eleven on 6 to 10 cases, and five on 11 to 20 cases. Three observers were general orthopaedic surgeons, thirteen were orthopaedic trauma surgeons, four were shoulder and elbow surgeons, and nine were hand and wrist surgeons.

Interobserver Reliability

Classification

When classifying fractures according to the Regan and Morrey classification, agreement among observers was fair with use of 2D-CT scans and moderate with use of 3D-CT agreement on O’Driscoll’s classification was moderate with both 2D-CT and 3D-CT ($k_{2D} = 0.42$, $SE = 0.013$, and $k_{3D} = 0.48$, SE 0.019; $p = 0.009$). (Table 1)
Figure 3. O’Driscoll Classification for Coronoid Fractures
Type 1: coronoid tip fracture
1. flake
2. fracture of tip
Type 2: anteromedial coronoid fractures
1. anteromedial fracture, between the tip of the coronoid and the sublime tubercle
2. comminution extending to the tip
3. comminution extending to the sublime tubercle
Type 3: basal coronoid fractures

Fracture Characteristics
With both imaging modalities, there was fair agreement among observers for diagnosis of fractures of the anteromedial facet ($\kappa_{2D} = 0.33$, SE 0.019, and $\kappa_{3D} = 0.30$, SE 0.030; $p = 0.398$). Interobserver agreement on presence of a coronoid tip fracture was poor with use of 2D-CT scans and slight with use of 3D-CT reconstructions ($\kappa_{2D} = 0.02$, SE 0.119, and $\kappa_{3D} = 0.19$, SE 0.082; $p = 0.268$). There was fair agreement among observers on fracture comminution with use of 2D-CT scans, and moderate agreement with use of 3D-CT reconstructions ($\kappa_{2D} = 0.29$, SE 0.045, and $\kappa_{3D} = 0.41$, SE 0.066; $p = 0.133$). Interobserver agreement on presence of impacted articular fragments was fair with use of both 2D-CT scans and 3D-CT reconstructions ($\kappa_{2D} = 0.27$, SE 0.047, and $\kappa_{3D} = 0.39$, SE 0.054; $p = 0.094$). Agreement on subluxation or dislocation was substantial with use of both 2D-CT scans and 3D-CT reconstructions ($\kappa_{2D} = 0.78$, SE 0.071, and $\kappa_{3D} = 0.62$, SE 0.070; $p = 0.109$).(Table 1)

Treatment
Interobserver agreement on operative exposure was fair with use of both 2D-CT scans and 3D-CT reconstructions ($\kappa_{2D} = 0.32$, SE 0.016, and $\kappa_{3D} = 0.27$, SE 0.013; $p = 0.015$). When observers were asked about the need for sutures, wires, or another technique, there was slight agreement with use of 2D-CT scans, and fair agreement with use of 3D-CT reconstructions ($\kappa_{2D} = 0.15$, SE 0.065, and $\kappa_{3D} = 0.31$, SE 0.086; $p = 0.138$). When observers decided on the
Table 1. Interobserver Reliability

<table>
<thead>
<tr>
<th>Classification</th>
<th>2D-CT</th>
<th></th>
<th>3D-CT</th>
<th></th>
<th>Significance comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Categorical</td>
<td>Kappa</td>
<td>SE</td>
<td>Categorical</td>
<td>Kappa</td>
</tr>
<tr>
<td>Regan &amp; Morrey</td>
<td>fair</td>
<td>0.40</td>
<td>0.014</td>
<td>moderate</td>
<td>0.51</td>
</tr>
<tr>
<td>O’Driscoll</td>
<td>moderate</td>
<td>0.42</td>
<td>0.013</td>
<td>moderate</td>
<td>0.48</td>
</tr>
<tr>
<td>Fracture Characteristics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anteromedial Facet Fracture</td>
<td>fair</td>
<td>0.33</td>
<td>0.019</td>
<td>fair</td>
<td>0.30</td>
</tr>
<tr>
<td>Tip Fracture</td>
<td>NS</td>
<td>0.03</td>
<td>0.119</td>
<td>slight</td>
<td>0.19</td>
</tr>
<tr>
<td>Comminution</td>
<td>fair</td>
<td>0.29</td>
<td>0.045</td>
<td>moderate</td>
<td>0.41</td>
</tr>
<tr>
<td>Impacted Articular Fragments</td>
<td>fair</td>
<td>0.27</td>
<td>0.047</td>
<td>fair</td>
<td>0.39</td>
</tr>
<tr>
<td>Subluxation / Dislocation</td>
<td>substantial</td>
<td>0.78</td>
<td>0.071</td>
<td>substantial</td>
<td>0.62</td>
</tr>
<tr>
<td>Treatment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approach</td>
<td>fair</td>
<td>0.32</td>
<td>0.016</td>
<td>fair</td>
<td>0.27</td>
</tr>
<tr>
<td>Need for Sutures / Wires / Other Technique</td>
<td>slight</td>
<td>0.15</td>
<td>0.065</td>
<td>fair</td>
<td>0.31</td>
</tr>
<tr>
<td>Need for Hinged External Fixator</td>
<td>slight</td>
<td>0.18</td>
<td>0.052</td>
<td>fair</td>
<td>0.22</td>
</tr>
</tbody>
</table>

SE = standard error; P = P-value
Intra-observer Reliability

We made an analysis of intra-observer reliability, but the reader should be mindful that observers looked at different images at the different time points, so we are measuring a combination of intra-observer reliability and changes with more sophisticated images.

Classification

Intra-observer agreement was moderate with respect to both the Regan and Morrey classification ($\kappa = 0.51$, SE 0.041), as well as the O’Driscoll classification ($\kappa = 0.44$, SE 0.042). (Table 2)

Fracture Characteristics

Intra-observer reliability was fair for recognition of coronoid tip fractures ($\kappa = 0.32$, SE 0.068) and comminution ($\kappa = 0.33$, SE 0.059), moderate for identification of fractures of the anteromedial facet ($\kappa = 0.44$, SE 0.053) and impaction ($\kappa = 0.41$, SE 0.057), and substantial for recognition of subluxation or dislocation ($\kappa = 0.73$, SE 0.047). (Table 2)

Treatment

Intra-observer reliability on operative exposure was moderate ($\kappa = 0.51$, SE 0.039), fair when asked about the need for sutures, wires, or another technique ($\kappa = 0.37$, SE 0.063), and moderate when asked about the need for an external fixator ($\kappa = 0.50$, SE 0.055). (Table 2)

Table 2. Intra-observer Reliability

<table>
<thead>
<tr>
<th></th>
<th>Categorical</th>
<th>Kappa</th>
<th>SE</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Classification</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regan &amp; Morrey</td>
<td>moderate</td>
<td>0.51</td>
<td>0.041</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>O’Driscoll</td>
<td>moderate</td>
<td>0.44</td>
<td>0.042</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td><strong>Fracture Characteristics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anteromedial Facet Fracture</td>
<td>moderate</td>
<td>0.44</td>
<td>0.053</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Tip Fracture</td>
<td>fair</td>
<td>0.32</td>
<td>0.068</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Communion</td>
<td>fair</td>
<td>0.33</td>
<td>0.059</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Impacted Articular Fragments</td>
<td>moderate</td>
<td>0.41</td>
<td>0.057</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Subluxation / Dislocation</td>
<td>substantial</td>
<td>0.73</td>
<td>0.047</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td><strong>Treatment</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approach</td>
<td>moderate</td>
<td>0.51</td>
<td>0.039</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Need for Sutures / Wires / Other Technique</td>
<td>fair</td>
<td>0.37</td>
<td>0.063</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Need for Hinged External Fixator</td>
<td>moderate</td>
<td>0.50</td>
<td>0.055</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

$SE =$ standard error; $P =$ P-value
Discussion

Regardless of the imaging modality used, there was fair to moderate agreement between observers for the majority of observations. Observations based upon 3D-CT reconstructions had significantly greater interobserver reliability than observations based on 2D-CT scans for fracture classification according both Regan and Morrey and O’Driscol, and there was a trend towards greater agreement for recognition of impacted articular fragments. Three-dimensional reconstructions, particularly those with the uninvolved bones subtracted, provide clear, intuitive images that our data would suggest are easier for surgeons to interpret. This is consistent with prior studies on the distal humerus and distal radius.\textsuperscript{9,12} However, agreement on operative exposure was significantly greater when observations were based on 2D-CT images.

Interobserver agreement was low for injury characteristics with disputable (e.g. what constitutes a “tip” fracture and optimal treatment technique) or less familiar characteristics (e.g. anteromedial facet fracture), and high for characteristics that are discrete and well defined (e.g. subluxation and dislocation). This supports the idea that interobserver reliability is influenced both by image quality as well as the individual observer’s understanding and experience. Factors with more variable definition and interpretation showed less interobserver reliability and less influence from 3D-CT reconstructions. In a similar vein, the limited interobserver agreement may reflect the relative infrequency of these fractures and concomitant paucity of reports in the literature, the complexity of these fractures in practice, as well as variations in training, practice, and experience. Interobserver reliability may measure what observers bring to an image as much as what they take away from that image.

Interobserver agreement is consistently much lower than intra-observer agreement across studies and seems to improve less with more sophisticated imaging techniques.\textsuperscript{9,12} Our study involving many fully trained and experienced surgeons practicing in different subspecialties in many parts of the world had lower intra-observer reliability than a prior study evaluating the same question for distal humeral fractures, but using trainees from one institution.\textsuperscript{9} The reasons for this are not entirely clear, but the greater number of experienced observers in the current study increases our confidence in the findings. Furthermore, in contrast to the study on distal humerus fractures\textsuperscript{9}, intra-observer reliability in the current study was based on evaluation of different images (2D-CT and 3D-CT) at the two time points. The bottom line is that even the most sophisticated imaging modalities are subject to interpretation.

A strength of this study compared to most reliability studies was the large number of observers and therefore of observations. In addition, participating surgeons came from multiple countries and continents, thereby increasing the generalizability of the results. Furthermore, whereas the typical reliability study relies on residents, fellows and other junior personnel that are available and willing to put in what often amounts to hours of effort, our
study included only fully trained and practicing surgeons. The web-based technology made this approach straightforward.

A shortcoming of our study includes the fact that the quality of the 2D-CT scans placed on the website was comparable to the actual 2D-CT scans available in the hospital, but the 3D-CT reconstructions were of somewhat less quality than those available in the hospital because observers could watch the elbows moving around the longitudinal and horizontal axis only, whereas intra-hospital 3D-CT images can be turned into any direction for a 360-degrees view of the joint in any axis. We believe that agreement might improve further with such an omni-angle view.

The addition of 3D-CT reconstructions to 2D-CT scans does not require repeat scanning or additional radiation. At our institution, using software to create 3D-CT reconstructions results in an approximate 20% increase in cost. In fact, some members of our research team quickly and easily learned to create 3D images on their own using the raw CT data, and we are confident that skill would be easily attainable for surgeons. Given that the size, location, and morphology of a coronoid fracture seems to correlate with overall injury pattern and guide treatment1-6,16,22-25, we believe the improved characterization of coronoid fracture fragments merit 3D reconstructions and has the potential to improve the results of treatment, although this should not be overstated given that even 3DCT interpretations varied substantially between observers. It appears that education and training may be more important determinants of observer variability than that the quality or sophistication of the imaging technique—a hypothesis that we plan to address in future studies.

References


PART III
Stiffness and Disability