Advancements in classification, treatment and outcome of radial head fractures
Guitton, T.G.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

Download date: 11 Dec 2018
CHAPTER 5
Interobserver Reliability of Radial Head Fracture Classification: Two-Dimensional vs. Three-Dimensional Computed Tomography

Thierry G. Guitton, MSc
David Ring, MD PhD
Science of Variation Group*

*Science of Variation Group:
Abstract

**Background:** The Broberg and Morrey modification of the Mason classification of radial head fractures has substantial interobserver variation. This study used a large web-based collaborative of experienced orthopaedic surgeons to test the hypothesis that three-dimensional reconstructions of computed tomography scans (3D-CT) improve the interobserver reliability of the classification of radial head fractures according to the Broberg and Morrey modification of the Mason classification.

**Methods:** Eighty-five orthopaedic surgeons evaluated twelve radial head fractures and were randomly assigned to review either radiographs and two-dimensional computed tomography scans (2D-CT) or radiographs, and 3D-CT images to determine the fracture classification, fracture characteristics and treatment plans. The kappa multirater measure ($\kappa$) was calculated to estimate agreement between observers.

**Results:** 3D-CT had moderate and 2D-CT had fair agreement among observers for the Mason classification ($\kappa_{3D} = 0.49$ vs. $\kappa_{2D} = 0.37; p < 0.001$). Among seven fracture characteristics (fracture line, comminution, articular surface involvement, gap/step of 2mm or greater, central impaction, recognition of three articular fragments and articular fragments to small to repair) there was a significant difference in kappa value between 3D-CT and 2D-CT for three variables (articular fragments too small to repair [$\kappa_{3D} = 0.61$ vs. $\kappa_{2D} = 0.47; p = < 0.001$], recognition of three articular fragments [$\kappa_{3D} = 0.60$ vs. $\kappa_{2D} = 0.38; p = < 0.001$] and central impaction [$\kappa_{3D} = 0.15$ vs. $\kappa_{2D} = 0.22; p = 0.006$]). Among treatment recommendations there was fair agreement for both 3D-CT and 2D-CT ($\kappa_{3D} = 0.40$ vs. $\kappa_{2D} = 0.26; p = < 0.001$).

**Conclusion:** Although 3D-CT led to some small but significant decreases in interobserver variation, there is still a notable degree of disagreement regarding classification and characterization of radial head fractures. Improvements in imaging may not be sufficient to optimize interobserver agreement.

*Level of Evidence: Diagnostic Level III*

---

**Introduction**

The classification of radial head fractures according to Broberg and Morrey’s modification of the Mason classification has substantial interobserver variation. As with classification and characterization of most fractures the interobserver variation is greater than the intraobserver variation. Evidence suggests that more sophisticated images such as 3D-CT improve intraobserver reliability more than interobserver reliability. A major limitation of most studies of observer variation is the use of only a few observers, most of them typically relatively junior surgeons.

A new collaboration motivated to better understand interobserver variation consists of observers who have completed all training and are independently treating patients. This provides an opportunity to further investigate interobserver variability and how to reduce it.

Treatment decisions for radial head fractures are often based on radiological criteria and measurements according to Broberg and Morrey’s modification of the Mason classification. This investigation tested the hypothesis that 3D-CT images improve the interobserver reliability of the classification and characterization of radial head fractures over 2D-CT and radiographs.

**Materials and Methods**

**Study Design**

Independent observers (all orthopaedic surgeons) from several countries were invited to evaluate twelve cases from a convenience sample of radial head fractures (selected to represent a full spectrum of radial head fracture morphologies and overall injury patterns) in an online survey: they were randomly assigned to review either radiographs and 2D-CT or radiographs and 3D-CT and then to determine the fracture classification, fracture characteristics and treatment plans. The randomization sequence was determined by a computer random number generator (Windows Excel; Microsoft, Redmond, WA). The study was performed under a protocol approved by the Institutional Research Board at the principal investigators hospital.

This was the inaugural study from a nascent collaborative called the Science of Variation Group (SOVG). The objectives of the collaborative are to study variation in the definition, interpretation, and classification of injury and disease. The Science of Variation Group has created a web-based platform that facilitates large international interobserver studies. With multiple fully trained surgeons from diverse countries and institutions participating in studies, this approach should provide a powerful forum for studying, understanding, and ultimately reducing interobserver variation in aspects of patient care.
Observers
A total of 206 surgeons were invited via e-mail to join the Science of Variation Group. We used lists of various professional organizations as well as friends and acquaintances (along with their friends and acquaintances) to identify surgeons to invite for participation. We welcome any interested surgeon to join. Other than an acknowledgement as part of the author collaborative in the paper, no incentives were provided. One-hundred surgeons were interested in participation and logged on to the website. Forty-eight surgeons were randomized to 2D-CT scans and radiographs and fifty-two to 3D-CT scans and radiographs. Four weekly reminders to complete the online survey were e-mailed. Eighty-eight surgeons completed the study, from which 3 observers were excluded because of inability to view the online study due to hospital restriction. This study presents an analysis of the eighty-five observers that completed the study; 39 in the 2D-CT group and 46 in the 3D-CT group.

Fractures
Radiographs and computed tomography scans of radial head fractures were identified from a list of all cases treated by the senior investigator between 2000 and 2006 at one level-1 trauma center. The scanning technique was evaluated to determine suitability for 3D reconstructions (slice thickness between 0.62 and 1.25 mm, no metallic implants). Inclusion criteria were: 1) Radial head fracture; 2) CT scan appropriate for 3D reconstruction; 3) Age 18 or older. Inadequate quality of the CT scan prompted exclusion from the study. Radiographs and CT scans of radial head fractures from 30 patients were blinded by an independent research fellow for use in this study. Two of the authors (one subspecialty trained upper extremity surgeon and one research fellow in upper extremity trauma) selected twelve cases that had radial head fractures of different size, morphology, and location; representing most of the different patterns of traumatic elbow instability with radial head fracture. Radiographs, 2D-CT scans, and 3D-CT reconstructions were uploaded to the research group’s website. 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, coronal and axial cuts 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.

Evaluation
Observers logged in independently on the website. Upon login to the website, they were asked to provide demographic and professional information: 1) location of practice; 2) years in independent practice; 3) training of surgical trainees; 4) number of radial head fractures treated per year, and; 5) clinical specialty. Subsequently, observers were asked to classify the fractures according to Broberg and Morrey’s modification of the Mason classification. Type 4 fractures were excluded, since we were interested in the radial head fracture independent of associated injuries. Observers were provided with the original description and corresponding images of the classification system.

The observers were also asked 7 questions regarding fracture characteristics:

1) Does the fracture line separate the entire articular surface from the radial neck?
2) Is there any comminution of the radial neck?
3) Does the fracture involve the articular surface?
4) Is there an articular step or gap of greater than 2 millimeters?
5) Is there any central impaction of the articular surface?
6) Are there more than 3 articular fragments?
7) Are any of the fragments too small to repair?

They were also asked their preferred management: 1) Non-operative management; 2) Open reduction and internal fixation with screws, wires or pins; 3) Open reduction and internal fixation with plate and screws; 4) Radial head excision; 5) Radial head replacement (arthroplasty). Observers were blinded to clinical information. Observers could comment on each case and all questions had to be completed in order to continue with the next case. The observers completed the study at their own time and pace.

Statistical Analysis
The kappa multirater measure (κ) was used to estimate agreement among surgeons with respect to fracture classification, fracture characteristics and treatment approach. It is a commonly used statistical method to describe chance-corrected agreement in a variety of intra-observer and interobserver studies. Agreement among observers was calculated with use of the multirater kappa measure described by Siegel and Castellan. Agreement among observers was calculated with use of the multirater kappa measure described by Siegel and Castellan. Kappa values were interpreted using the guidelines proposed by Landis and Koch: 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. Zero indicates no agreement beyond that expected due to chance alone. – 1.00 means total disagreement, and + 1.00 represents perfect agreement. Two-sample independent Z-tests were performed for each variable to compare the kappa for 2 dimensional CT with that of 3 dimensional CT. Since the samples compared in this study were not independent (the same set of fractures were rated by the 2D-CT and 3D-CT group), this method produced conservative estimates of the p-values. A post-hoc power analysis was performed using nQuery Advisor (version 7.0, nQuery Advisor, Statistical Solutions, Saugus, MA) to identify the power of each comparison and the sample size necessary to achieve 80% given both effect size and rater ratio remain constant at each iteration (Table V).
**Sources of Funding**
No funding was received in direct support of this study.

**Results**

**Observer Demographics**
A total of 85 observers participated in this investigation. The observer demographics are summarized in Table I. Among the others surgeons there were 3 hand surgeons (no wrist), 2 trauma surgeons and 3 upper extremity surgeons (hand, wrist, elbow and shoulder).

<table>
<thead>
<tr>
<th>Variable</th>
<th>2-Dimensional CT (N = 39)</th>
<th>3-Dimensional CT (N = 46)</th>
<th>Total (N = 85)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Practice</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asia</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Australia</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Canada</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Europe</td>
<td>6</td>
<td>15</td>
<td>21</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>United States</td>
<td>28</td>
<td>72</td>
<td>100</td>
</tr>
<tr>
<td>Other</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Years in practice</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-5</td>
<td>11</td>
<td>28</td>
<td>18</td>
</tr>
<tr>
<td>6-10</td>
<td>6</td>
<td>15</td>
<td>21</td>
</tr>
<tr>
<td>11-20</td>
<td>18</td>
<td>46</td>
<td>35</td>
</tr>
<tr>
<td>21-30</td>
<td>4</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>Supervise Residents</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>35</td>
<td>90</td>
<td>125</td>
</tr>
<tr>
<td>No</td>
<td>4</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>Fractures per year</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-5</td>
<td>10</td>
<td>26</td>
<td>13</td>
</tr>
<tr>
<td>6-10</td>
<td>10</td>
<td>26</td>
<td>33</td>
</tr>
<tr>
<td>11-20</td>
<td>13</td>
<td>33</td>
<td>26</td>
</tr>
<tr>
<td>21-30</td>
<td>6</td>
<td>15</td>
<td>11</td>
</tr>
<tr>
<td>Specialization</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General orthopaedics</td>
<td>2</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Orthopaedic traumatology</td>
<td>12</td>
<td>31</td>
<td>33</td>
</tr>
<tr>
<td>Shoulder and elbow</td>
<td>7</td>
<td>18</td>
<td>15</td>
</tr>
<tr>
<td>Hand and wrist</td>
<td>14</td>
<td>36</td>
<td>38</td>
</tr>
<tr>
<td>Other</td>
<td>4</td>
<td>10</td>
<td>8</td>
</tr>
</tbody>
</table>

**Classification**
The interobserver variation for classification of the fractures according to Broberg and Morrey’s modification of the Mason classification was fair with use of 2D-CT scans and moderate with use of 3D-CT reconstructions ($\kappa_{2D} = 0.37, SE 0.010$ and $\kappa_{3D} = 0.49, SE 0.033; p < 0.001$) (Table IIA-B).

**Fracture Characteristics**
Agreement on central impaction of the articular surface was fair with use of 2D-CT scans and slight with use of 3D-CT reconstructions ($\kappa_{2D} = 0.22, SE 0.027$ and $\kappa_{3D} = 0.15, SE 0.026; p < 0.006$). Interobserver agreement on presence of more than 3 art-
ticular fragments was fair with use of 2D-CT scans and substantial with use of 3D-CT reconstructions ($\kappa = 0.38$, SE $0.01$, and $\kappa = 0.61$, SE $0.010$; $p < 0.001$). Agreement on presence of fragments too small to repair was moderate with use of 2D-CT scans and substantial with use of 3D-CT reconstructions ($\kappa = 0.47$, SE $0.013$, and $\kappa = 0.61$, SE $0.010$; $p < 0.001$) (Table IIIA-B).

**Treatment**

Interobserver agreement on treatment was fair with both 2D-CT scans and 3D-CT reconstructions ($\kappa = 0.26$, SE $0.012$, and $\kappa = 0.40$, SE $0.013$; $p < 0.001$) (Table II).

**Observer demographics and Mason classification**

When classifying fractures according to the Mason classification agreement among United States observers was fair with use of 2D-CT scans and moderate with use of 3D-CT reconstructions ($\kappa = 0.32$, SE $0.01$, and $\kappa = 0.52$, SE $0.03$; $p < 0.001$) (Table IIIA-B).

Agreement among observers who were in practice 5 or fewer years was moderate with use of 2D-CT scans and substantial with use of 3D-CT reconstructions ($\kappa = 0.44$, SE $0.01$, and $\kappa = 0.62$, SE $0.18$; $p = 0.039$). Agreement among observers who were in practice from 11 to 20 years was fair with use of 2D-CT scans and moderate with use of 3D-CT reconstructions ($\kappa = 0.35$, SE $0.02$, and $\kappa = 0.45$, SE $0.04$; $p = 0.011$).

Agreement among observers who treated 5 or fewer radial head fractures per year was fair with both 2D-CT scans and 3D-CT reconstructions ($\kappa = 0.27$, SE $0.03$, and $\kappa = 0.32$, SE $0.14$; $p = 0.76$). Agreement among observers who treated 6 to 10 radial head fractures per year was fair with use of 2D-CT scans and moderate with use of 3D-CT reconstructions ($\kappa = 0.39$, SE $0.04$, and $\kappa = 0.48$, SE $0.04$; $p = 0.069$). Agreement among observers who treated 6 to 10 radial head fractures per year was fair with both 2D-CT scans and 3D-CT reconstructions ($\kappa = 0.44$, SE $0.03$, and $\kappa = 0.46$, SE $0.05$; $p = 0.66$). Agreement among observers who treated more than 20 radial head fractures per year was moderate with both 2D-CT scans and 3D-CT reconstructions ($\kappa = 0.47$, SE $0.05$; $p = 0.44$).

Agreement among orthopaedic traumatology specialist observers was fair with use of 2D-CT scans and moderate with use of 3D-CT reconstructions ($\kappa = 0.37$, SE $0.03$, and $\kappa = 0.47$, SE $0.04$; $p = 0.05$). Agreement among hand and wrist spe-
reliability of radial head fracture classification   | chapter 5

...classification among imaging modality

<table>
<thead>
<tr>
<th>Variable</th>
<th>2-Dimensional CT</th>
<th>3-Dimensional CT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Practice</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Europe vs. United States</td>
<td>0.006*</td>
<td>0.06</td>
</tr>
<tr>
<td>Years in practice</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-5 vs. 21-30</td>
<td>0.715</td>
<td>0.037*</td>
</tr>
<tr>
<td>Fractures per year</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-5 vs. &gt;20</td>
<td>0.006*</td>
<td>0.175</td>
</tr>
<tr>
<td>Specialization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ortho Trauma vs. Shoulder &amp; elbow</td>
<td>0.017*</td>
<td>0.175</td>
</tr>
</tbody>
</table>

* = Significant (p < 0.05)

Discussion

The collaborative, internet based approach has facilitated large, international studies of inter-rater variation. Additionally, only fully trained surgeons, many with substantial clinical experience participated. Inclusion of surgeons from multiple countries and continents should increase the generalizability of the results. Using high-speed internet connections and improved compression techniques, we were able to provide high quality reproduction images and movies via the internet.

There are many weaknesses in this study. First, the quality of the radiographs was limited to what had been obtained at the time of injury, which reflects usual practice, but not what might be achieved with specific protocols. In addition, we provided limited information about the patient and the injury. There was also a spectrum bias by selecting cases to represent the known variety of injuries, with the result that less common complex fractures were over represented compared to the more common minimally or slightly displaced fractures. Our study reflects what would be expected with relatively complex fractures of the radial head—the reliability would be expected to be higher if we included more of the non-displaced or minimally displaced fractures that makeup the majority of radial head fractures.

Another shortcoming is the fact that a small number of observers either uncommonly or never treat radial head fractures, but we did not plan for exclusions on this basis and therefore did not do so after the fact to avoid introducing bias. The power is based on the total number of observations allowing us to use a smaller number of cases and thereby decrease burden and increase participation of observers. Given the small kappa differences between 2D and 3D for certain questions (e.g. articular surface involvement) and the large variabilities, power was low, and huge sample sizes would be required for 80% power for detecting differences between 2D and 3D given the levels of agreement observed. For other questions (e.g. three articular fragments), the observed power was very high with the numbers of surgeons participating and even fewer would have attained the traditional 80% power. Finally, this is an artificial research situation given that in clinical practice patients would have both the two and three-dimensional reconstructions available to them.

We speculate that the very poor agreement regarding articular surface involvement might reflects misunderstanding of the question-based on comments received as part of the survey some observers probably thought we were referring to involvement of the part of the radial head that articulates with the lesser sigmoid notch of the ulna. The poor agreement regarding central impaction likely reflects the lack of a precise or consistent definition of this term. The findings of this study are otherwise consistent with prior studies on the distal humerus, distal radius and the coronoid. Three-dimensional reconstructions are made from CT-scans and there-
fore do not require additional scanning or expose the patient to additional radiation. It has been calculated at the investigators institution, the cost for additional 3D reconstructions are an additional 20% of the costs of a CT-scan. Free software such as OsiriX is available which makes it possible for every orthopedic surgeon to quickly and easily create 3D reconstructions themselves with minimal training.

Three-dimensional CT images led to small but significant decreases in variation between observers for fracture classification and some fracture characteristics compared to 2D-CT, but a notable amount of variation remains even with more sophisticated imaging. Our belief that 3D-CT images are easier for surgeons to interpret is supported by the observation that 3D-CT produced a higher agreement for Broberg and Morrey’s modification of the Mason classification than previously reported in the literature and 3D-CT was associated with less disagreement in classification than 2D-CT across various cultures, training, subspecialty and levels of experience. Nonetheless agreement was only fair or moderate at best even with 3D-CT. Furthermore, some might interpret this data as showing much less influence one interobserver variation than one might guess.

Other potential sources of interobserver variation include unfamiliar or unclear definitions, and differences in culture, training, and exposure. In our opinion, the fact that well-trained, experienced observers disagree indicates that there are variations in these factors that lead different experts to see different things in sophisticated images. In other words, reducing interobserver variation seems to depend on something more than better imaging. Additional research to identify and reduce sources of observer variation in the interpretation of diagnostic images is merited.

**References**