The impact of stereo-visualisation of three-dimensional CT datasets on the inter- and intraobserver reliability of the AO/OTA and Neer classifications in the assessment of fractures of the proximal humerus

A. Brunner,
P. Honigmann,
T. Treumann,
R. Babst

*From The Cantonal
Hospital, Lucerne,
Switzerland*

We evaluated the impact of stereo-visualisation of three-dimensional volume-rendering CT datasets on the inter- and intraobserver reliability assessed by kappa values on the AO/OTA and Neer classifications in the assessment of proximal humeral fractures. Four independent observers classified 40 fractures according to the AO/OTA and Neer classifications using plain radiographs, two-dimensional CT scans and with stereo-visualised three-dimensional volume-rendering reconstructions. Both classification systems showed moderate interobserver reliability with plain radiographs and two-dimensional CT scans. Three-dimensional volume-rendered CT scans improved the interobserver reliability of both systems to good. Intraobserver reliability was moderate for both classifications when assessed by plain radiographs. Stereo visualisation of three-dimensional volume rendering improved intraobserver reliability to good for the AO/OTA method and to excellent for the Neer classification.

These data support our opinion that stereo visualisation of three-dimensional volume-rendering datasets is of value when analysing and classifying complex fractures of the proximal humerus.

A number of systems have been proposed to classify fractures of the proximal humerus.¹⁻⁴ The two most frequently used are the AO/OTA system of Müller et al³ and the Neer⁴ classification. Several authors⁵⁻⁸ have evaluated the inter- and intraobserver reliability of these two classifications on the basis of plain radiographs, but have not found satisfactory reproducibility. Additionally, the use of two- (2D) and three-dimensional (3D) CT scans has failed to increase their reliability.^{5,8,9} The need for further research on modifications to imaging techniques to improve reproducibility in the classification of these injuries has been discussed.^{5,7} Three-dimensional images of CT datasets may be generated using a variety of computer algorithms,¹⁰ of which the most frequently used in skeletal radiology to date has been shaded surface display. This technique creates an external 3D view of the surface of the bone on the basis of approximately 10% of the CT data.¹⁰

With more powerful computer hardware, volume rendering^{10,11} has been proposed as an alternative imaging technique to evaluate the musculoskeletal system. The volume rendering algorithm uses a higher percentage of the CT data, depending on the transfer table,

to create the 3D image.^{10,11} The technique makes structures transparent thereby allowing the observer to look into the interior of the bone.¹¹ During the last decade, the 3D volume rendering technique has been increasingly used in orthopaedic and trauma practice.¹²⁻¹⁵

A major limitation of this and other algorithms is that they create a 3D image which is usually displayed on a 2D screen. The spatial information has then to be generated in the mind of the observer.

In order to enable real stereo-visualisation of 3D CT data we used a special workstation to create volume rendering datasets with separate images for the right and left eyes. These images were displayed to the corresponding eye resulting in a realistic 3D image.

However, little is known regarding the impact of stereo-visualisation of 3D volume rendering CT datasets on the reproducibility of classification systems for proximal humeral fractures. Our aim therefore was to evaluate the impact of stereo-visualisation of 3D volume rendering CT datasets on the inter- and intraobserver reliability of the AO/OTA and Neer classifications in the assessment of fractures of the proximal humerus.

■ A. Brunner, MD, Trauma
Surgery Resident
■ P. Honigmann, MD, Trauma
Surgery Resident
■ R. Babst, MD, Professor
Department of Trauma Surgery
■ T. Treumann, MD, Radiologist
Department of Radiology
Cantonal Hospital,
Spitalstrasse 42, 6000 Lucerne
16, Switzerland.

Correspondence should be sent
to Dr A. Brunner; e-mail:
a-r.brunner@web.de

©2009 British Editorial Society
of Bone and Joint Surgery
doi:10.1302/0301-620X.91B6.
22109 \$2.00

J Bone Joint Surg [Br]
2009;91-B:766-71.

Received 25 November 2008;
Accepted 9 March 2009

Table I. Interobserver reliability (kappa values) for the AO/OTA classification

	Plain radiographs	2D CT	2D CT + 3D volume rendering
Observer			
I/II			
Type	0.56	0.69	0.87
Group	0.47	0.59	0.78
Subgroup	0.51	0.54	0.68
I/III			
Type	0.64	0.78	0.87
Group	0.50	0.55	0.78
Subgroup	0.42	0.47	0.64
I/IV			
Type	0.65	0.60	0.87
Group	0.56	0.53	0.75
Subgroup	0.38	0.39	0.62
II/III			
Type	0.63	0.73	0.83
Group	0.46	0.59	0.69
Subgroup	0.34	0.50	0.65
II/IV			
Type	0.55	0.64	0.91
Group	0.41	0.49	0.78
Subgroup	0.28	0.42	0.68
III/IV			
Type	0.64	0.63	0.83
Group	0.50	0.61	0.72
Subgroup	0.42	0.53	0.56
Mean interobserver reliability			
Type	0.61	0.68 (p = 0.091)	0.86* (p = 0.002)
Group	0.48	0.56 (p = 0.026)	0.75* (p = 0.001)
Subgroup	0.39	0.48 (p = 0.021)	0.64* (p = 0.004)

* mean kappa values were significantly higher (*t*-test, $p < 0.01$) for type, group and subgroup when classified on the basis of 3D volume rendering reconstructions compared with classification by 2D CT scans and plain radiographs

Patients and Methods

Plain anteroposterior (AP) radiographs and CT scans of fractures of the proximal humerus were acquired from 40 consecutive adult patients over a period of three years (2001 to 2003). All the CT scans were performed on a Somatom scanner (Siemens, Erlangen, Germany) with a slice thickness of 1.5 mm. The primary image plane was axial with a restricted field of view to the shoulder of 180 mm. Oblique axial, coronal and sagittal images adapted to the plane of the shoulder were reformatted from the bone algorithm with a slice thickness of 3 mm for the reformats.

The datasets were transferred to a Virtuoso workstation (Siemens) to create stereovisible 3D volume rendering images. The workstation used a SUN Sparc II Computer (Silicon Graphics, Delaware, Canada) running the Unix operating system (The Open Group, San Francisco, California) with the Virtuoso software installed. After uploading the original CT datasets the workstation generated different 3D volume rendering images for the right and the

left eyes. These images were displayed on a computer screen alternating at a frequency of 25 Hz. The observer wore spectacles with polarised lenses. Each lens could be switched from lucent to opacified by the application of an electric field to the glass. This alternately impaired and allowed vision separately in each eye, with the field of vision being generated by an in-built battery in the spectacles. A transmitter placed on the computer monitor, with the spectacles having a receiver between the eyes, triggered the restriction of the field of vision in one eye and then the other. With this tool, real 3D imaging of the object could be achieved.

For evaluation the datasets were blinded by a third party. In order to consider a possible effect of the observers' experience on the classification of proximal humeral fractures, we chose four observers (RB, TT, VR, UC) with different levels of clinical training and experience as follows: one trauma surgeon sub-specialising in problems of the shoulder (observer I, RB), one general trauma surgeon (observer II, VR), a radiologist specialising in imaging of the

Table II. Intraobserver reliability (kappa value) for the AO/OTA classification

	Plain radiographs	2D CT	2D CT + 3D volume rendering
Observer			
I			
Type	0.64	0.78	0.96
Group	0.46	0.62	0.81
Subgroup	0.38	0.49	0.67
II			
Type	0.63	0.74	0.91
Group	0.50	0.53	0.72
Subgroup	0.34	0.42	0.54
III			
Type	0.62	0.77	0.87
Group	0.52	0.58	0.81
Subgroup	0.32	0.43	0.67
IV			
Type	0.63	0.68	0.91
Group	0.52	0.52	0.72
Subgroup	0.31	0.41	0.62
Mean intraobserver reliability			
Type	0.63	0.74 (p = 0.015)	0.91* (p = 0.008)
Group	0.50	0.56 (p = 0.170)	0.77* (p < 0.001)
Subgroup	0.34	0.44* (p = 0.001)	0.63* (p = 0.005)

* the mean kappa value for subgroup was significantly higher (*t*-test, $p < 0.01$) when classified with 2D CT scans compared with classification using plain radiographs. After the addition of 3D volume rendering kappa values were significantly higher (*t*-test, $p < 0.01$) for type, group, and subgroup compared with classification with 2D scans and plain radiographs

musculoskeletal system (observer III, TT) and a radiology resident (observer IV, UC). Each observer was provided with diagrams of the AO/OTA and Neer classification systems as well as with the original publications.

According to the AO/OTA system, each fracture is classified into one of three types as follows: type A indicates extra-articular unifocal fractures; type B, extra-articular bifocal fractures; and type C, articular fractures. Each of these categories is subdivided into three groups (1, 2 and 3) on the basis of the pattern of the fracture. Finally, each group is then divided into three subgroups (1, 2 and 3) according to the degree of fragmentation resulting in 27 different potential patterns of injury. On the Neer system, each fracture is classified as follows: group I, non-displaced; group II, two-part; group III, three-part; and group IV, four-part. Each group is further divided according to the pattern of the fracture defining a total of 16 different potential types of fracture.

The images were evaluated in a randomised order. Three evaluations were performed. First, the classification of the fracture on the sole basis of plain radiographs was evaluated. Four weeks later, a combination of plain radiographs and 2D CT scans was evaluated. Finally, after another four weeks, the fractures were classified on the basis of plain radiographs, 2D CT scans and stereo-visualised 3D volume rendering reconstructions. The use of 3D volume rendering imaging was not analysed alone because it is most commonly used in conjunction with 2D scans.¹⁶

After an interval of eight weeks, these rounds were repeated in a newly randomised order to evaluate intraobserver reliability. The observers were not provided with any feedback and the radiographs were not available to any of them between the readings.

Statistical analysis. Inter- and intraobserver reliability was assessed by calculating the kappa (κ) correlation coefficient as described by Cohen¹⁷ and the interpretation of kappa coefficients was performed using the criteria of Landis and Koch.¹⁸ They defined a kappa of more than 0.8 as excellent, of between 0.6 and 0.8 as good, of between 0.4 and 0.6 as moderate, and of less than 0.4 as poor correlation. The paired Student *t*-test was used to calculate statistical differences between the mean kappa values. In order to consider statistical differences between single kappa values, we calculated the 99% confidence intervals (CI). According to the recommendations of Doornberg et al¹⁹ differences between single kappa values were considered to be significant when the upper and lower boundaries of the 99% CI did not overlap. A p -value ≤ 0.01 was considered to be statistically significant.

Results

AO/OTA classification. The mean interobserver reliability for the AO/OTA classification is given in Table I. According to type (A, B, or C) it was good ($\kappa = 0.61$) when classified solely on the basis of plain radiographs and slightly

Table III. Interobserver reliability (kappa values) of the Neer classification

	Plain radiographs	2D CT	2D CT + 3D volume rendering
Observer			
I/II			
Group	0.53	0.64	0.81
Fracture type	0.42	0.53	0.77
I/III			
Group	0.48	0.59	0.76
Fracture type	0.39	0.54	0.72
I/IV			
Group	0.48	0.63	0.82
Fracture type	0.42	0.55	0.82
II/III			
Group	0.46	0.54	0.75
Fracture type	0.44	0.58	0.76
II/IV			
Group	0.48	0.57	0.81
Fracture type	0.39	0.59	0.70
III/IV			
Group	0.45	0.53	0.85
Fracture type	0.45	0.54	0.81
Mean interobserver reliability			
Group	0.48	0.58* (p < 0.001)	0.80* (p < 0.001)
Fracture type	0.42	0.56* (p < 0.001)	0.76* (p < 0.001)

* mean kappa values were significantly higher (*t*-test, $p < 0.01$) for group and fracture type when classified with 2D CT scans compared with classification using plain radiographs. After the addition of 3D volume rendering kappa values were significantly higher (*t*-test, $p < 0.01$) for group and fracture type compared with classification by classified 2D CT scans and plain radiographs

improved ($\kappa = 0.68$) after the addition of 2D CT scans. When classified on the basis of 2D CT scans and 3D volume rendering reconstructions interobserver reliability significantly improved to excellent ($\kappa = 0.86$). The classification of proximal humeral fractures according to group (nine categories) showed moderate mean interobserver reliability with plain radiographs ($\kappa = 0.48$) and 2D CT scans ($\kappa = 0.56$). The mean interobserver reliability significantly improved to good when fractures were classified on the basis of 3D volume rendering reconstruction scans ($\kappa = 0.75$). The mean interobserver reliability according to subgroup (27 categories) was poor with plain radiographs ($\kappa = 0.39$) and only moderate with CT scans ($\kappa = 0.48$), but significantly improved to good after the addition of 3D volume rendering reconstructions ($\kappa = 0.64$).

The mean intraobserver reliability for the AO/OTA classification is given in Table II. For type (A, B, or C) it was good with plain radiographs ($\kappa = 0.63$) and 2D CT scans ($\kappa = 0.74$) but again significantly improved to excellent after the addition of 3D volume rendering reconstructions ($\kappa = 0.91$). The mean intraobserver reliability for group (nine categories) was moderate with plain radiographs ($\kappa = 0.50$) and 2D CT scans ($\kappa = 0.56$) and significantly improved to good after adding 3D volume rendering recon-

structions ($\kappa = 0.77$). For classification according to subgroup (27 categories) the mean intraobserver reliability was poor ($\kappa = 0.34$) when the observers used plain radiographs and improved to moderate with 2D CT scans ($\kappa = 0.44$). After the addition of 3D volume rendering reconstructions the mean intraobserver reliability significantly improved to good ($\kappa = 0.63$). Finally, statistical analysis showed no significant differences regarding inter- and intraobserver reliability between observers with different levels of experience.

Neer classification. The mean interobserver reliability for the Neer classification is given in Table III. According to group (I, II, III, IV) it showed moderate reliability with plain radiographs ($\kappa = 0.48$) and 2D CT scans ($\kappa = 0.58$), but improved to excellent after adding 3D volume rendering reconstructions ($\kappa = 0.80$). The mean interobserver reliability according to type of fracture (16 categories) was moderate with plain radiographs ($\kappa = 0.42$) and 2D CT scans ($\kappa = 0.56$), but significantly improved to good with 3D volume rendering reconstructions ($\kappa = 0.76$).

The mean intraobserver reliability is given in Table IV. For group (I, II, III, IV) it was moderate with plain radiographs ($\kappa = 0.58$), good with CT scans ($\kappa = 0.69$), and excellent with 3D volume rendering reconstructions ($\kappa = 0.88$).

Table IV. Intraobserver reliability (kappa values) for the Neer classification

	Plain radiographs	2D CT	2D CT + 3D volume rendering
Observer			
I			
Group	0.61	0.72	0.95
Fracture type	0.51	0.61	0.88
II			
Group	0.58	0.71	0.85
Fracture type	0.47	0.62	0.81
III			
Group	0.55	0.70	0.90
Fracture type	0.45	0.65	0.87
IV			
Group	0.58	0.64	0.81
Fracture type	0.49	0.62	0.81
Mean intraobserver reliability			
Group	0.58	0.69* (p = 0.009)	0.88* (p = 0.002)
Fracture type	0.48	0.63* (p = 0.006)	0.84* (p = 0.001)

* mean kappa values were significantly higher (*t*-test, $p < 0.01$) for group and type of fracture when classified with 2D CT scans compared to that when classified with plain radiographs. After the addition of 3D volume rendering kappa values were significantly higher (*t*-test, $p < 0.01$) for group and fracture type compared with classification with 2D CT scans and plain radiographs

Intraobserver reliability for type of fracture (16 categories) was moderate with plain radiographs ($\kappa = 0.48$), good with 2D CT scans ($\kappa = 0.63$), and excellent with 3D volume rendering reconstructions ($\kappa = 0.84$). Once again, statistical analysis showed no significant differences regarding inter- and intra-observer reliability between observers with different levels of experience.

Discussion

In our study the AO/OTA and Neer classifications both showed moderate interobserver reliability when classified solely on the basis of plain radiographs and when classified with the use of both plain radiographs and 2D CT scans. Stereo-visualisation of 3D volume-rendered CT datasets significantly improved the interobserver reliability of both systems to good. Likewise, intraobserver reliability was moderate for both classifications with plain radiographs. After the addition of 2D CT scans it remained moderate for the AO/OTA system and improved to good for the Neer classification. Stereo-visualisation of 3D volume rendering reconstructions significantly improved intraobserver reliability to good for the AO/OTA and to excellent for the Neer classifications.

Several other studies^{5,8,9} have evaluated the reliability of the AO/OTA and Neer classifications on the basis of plain radiographs. Their results were comparable with our data. Bernstein et al⁵ evaluated the impact of 2D CT scans on the reliability of the Neer classification. The adding of CT scans to plain radiographs resulted in a slight increase in intraobserver reliability, but no increase in interobserver reliability. Likewise, Sjöden et al^{8,9} evaluated the impact of

2D and 3D CT scans on the reproducibility of both classifications. Both imaging techniques could not significantly improve inter- and intraobserver reliability compared with that of plain radiographs alone. However, in the studies of Sjöden et al^{8,9} CT scanning had several technical limitations. Slice thickness was 3 mm, resulting in reformats of reduced quality, and the images were provided on plain films which did not allow interactive assessment of the fracture.

In our series, the analysis was performed interactively on a computer workstation with a slice thickness of 1.5 mm which enhanced the quality of coronal and sagittal reformatting, as well as 3D reconstruction. These technical improvements in conjunction with additional stereo-visualisation may have led to an improved reliability of the classification of proximal humeral fractures than was previously suggested by Sjöden et al.^{8,9} More advanced imaging techniques with high-definition monitors instead of special polarising spectacles may further enhance the results towards higher inter- and intraobserver reliability.

Some authors^{6,20} have reported that repeated training of observers on a classification system improves its reliability. Since the AO/OTA system is used as the standard classification in our department to classify fractures, this may have led to a bias in the results. However, in our study both classification systems showed comparable reproducibility.

Another potential flaw of our study was its retrospective nature. Patients who did not have CT as part of the initial management of the fracture were excluded from the study which might also have biased our patients collectively towards more complex patterns of fracture. However, it has been found that the reproducibility of the AO/OTA and

Neer classifications was better in less complex types of fractures⁷⁻⁹ and therefore we do not believe that this compromised our results.

Our results suggest that stereo-visualisation of 3D volume rendering datasets is of value when analysing and classifying complex fractures of the proximal humerus.

The authors would like to thank Dr. V. Roth and Dr. U. Caspar for their contribution to this work.

Dr A. Brunner and Dr P. Honigmann contributed equally to this work.

No benefits in any form have been received or will be received from a commercial party related directly or indirectly to the subject of this article.

References

1. Bahrs C, Schmal H, Lingenfelter E, et al. Inter- and intraobserver reliability of the MTM-classification for proximal humeral fractures: a prospective study. *BMC Musculoskelet Disord* 2008;9:21.
2. Edelson G, Kelly I, Vigder F, Reis ND. A three-dimensional classification for fractures of the proximal humerus. *J Bone Joint Surg [Br]* 2004;86-B:413-25.
3. Müller ME, Nazarian S, Koch P, Schatzker J. *The comprehensive classification of fractures of long bones*. Berlin: Springer Verlag, 1990:120-1.
4. Neer CS. *Shoulder Reconstruction*. Philadelphia: W. B. Saunders, 1990:363-98.
5. Bernstein J, Adler LM, Blank JE, et al. Evaluation of the Neer system of classification of proximal humeral fractures with computerized tomographic scans and plain radiographs. *J Bone Joint Surg [Am]* 1996;78-A:1371-5.
6. Sidor ML, Zuckerman JD, Lyon T, et al. The Neer classification system for proximal humeral fractures: an assessment of interobserver reliability and intraobserver reproducibility. *J Bone Joint Surg [Am]* 1993;75-A:1745-50.
7. Siebenrock KA, Gerber C. The reproducibility of classification of fractures of the proximal end of the humerus. *J Bone Joint Surg [Am]* 1993;75-A:1751-5.
8. Sjöden GO, Movin T, Guntner P, et al. Poor reproducibility of classification of proximal humeral fractures: additional CT of minor value. *Acta Orthop Scand* 1997;68:239-42.
9. Sjöden GO, Movin T, Aspelin P, Guntner P, Shalabi A. 3D-radiographic analysis does not improve the Neer and AO classifications of proximal humeral fractures. *Acta Orthop Scand* 1999;70:325-8.
10. Calhoun PS, Kuszyk BS, Heath DG, Carley JC, Fishman EK. Three-dimensional volume rendering of spiral CT data: theory and method. *Radiographics* 1999;745-64.
11. Pretorius ES, Fishman EK. Volume-rendered three-dimensional spiral CT: musculoskeletal applications. *Radiographics* 1999;19:1143-60.
12. Coan P, Mollenhauer J, Wagner A, Muehleman C, Bravin A. Analyzer-based imaging technique in tomography of cartilage and metal implants: a study at the ESRF. *Eur J Radiol* 2008;68:541-8.
13. Lawler LP, Corl FM, Fishman EK. Multi- and single detector CT with 3D volume rendering in tibial plateau fracture imaging and management. *Crit Rev Comput Tomogr* 2002;43:251-82.
14. Purnell ML, Larson AI, Clancy W. Anterior cruciate ligament insertions on the tibia and femur and their relationships to critical bony landmarks using high-resolution volume-rendering computed tomography. *Am J Sports Med* 2008;36:2083-90.
15. Taser F, Shafiq Q, Ebraheim NA. Three-dimensional volume rendering of tibiofibular joint space and quantitative analysis of change in volume due to tibiofibular syndesmosis diastases. *Skeletal Radiol* 2006;35:935-41.
16. Bahrs C, Rolaufts B, Sudkamp NP, et al. Indications for computed tomography (CT-) diagnostics in proximal humeral fractures: a comparative study of plain radiography and computed tomography. *BMC Musculoskelet Disord* 2009;10:33.
17. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960;20:37-46.
18. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159-74.
19. Doornberg J, Lindenhovius A, Kloen P, et al. Two and three-dimensional computed tomography for the classification and management of distal humeral fractures: evaluation of reliability and diagnostic accuracy. *J Bone Joint Surg [Am]* 2006;88-A:1795-801.
20. Rasmussen S, Madsen PV, Bennicke K. Observer variation in the Lauge-Hansen classification of ankle fractures: precision improved by instruction. *Acta Orthop Scand* 1993;64:693-4.