



Interobserver reliability of classification and characterization of proximal humeral fractures: a comparison of two and three-dimensional CT

AUTHOR(S)

W Bruinsma, T Guitton, J Warner, D Ring, Richard Page

PUBLICATION DATE

04-09-2013

HANDLE

[10536/DRO/DU:30058954](https://hdl.handle.net/10536/DRO/DU:30058954)

Downloaded from Deakin University's Figshare repository

Deakin University CRICOS Provider Code: 00113B

Deakin Research Online

This is the published version:

Bruinsma, Wendy E., Guitton, Thierry G., Warner, Jon J. P., Ring, David and Page, Richard
2013, Interobserver reliability of classification and characterization of proximal humeral
fractures: a comparison of two and three-dimensional CT, *Journal of bone & joint surgery*,
vol. 95, no. 17, pp. 1600-1604.

Available from Deakin Research Online:

<http://hdl.handle.net/10536/DRO/DU:30058954>

Reproduced with the kind permission of the copyright owner.

Copyright : 2013, Journal of bone & joint surgery

Interobserver Reliability of Classification and Characterization of Proximal Humeral Fractures

A Comparison of Two and Three-Dimensional CT

Wendy E. Bruinsma, MD, Thierry G. Guitton, MD, PhD, Jon J.P. Warner, MD, David Ring, MD, PhD,
and the Science of Variation Group*

Investigation performed at the Department of Orthopaedic Surgery, Massachusetts General Hospital, Boston, Massachusetts

Background: Interobserver reliability for the classification of proximal humeral fractures is limited. The aim of this study was to test the null hypothesis that interobserver reliability of the AO classification of proximal humeral fractures, the preferred treatment, and fracture characteristics is the same for two-dimensional (2-D) and three-dimensional (3-D) computed tomography (CT).

Methods: Members of the Science of Variation Group—fully trained practicing orthopaedic and trauma surgeons from around the world—were randomized to evaluate radiographs and either 2-D CT or 3-D CT images of fifteen proximal humeral fractures via a web-based survey and respond to the following four questions: (1) Is the greater tuberosity displaced? (2) Is the humeral head split? (3) Is the arterial supply compromised? (4) Is the glenohumeral joint dislocated? They also classified the fracture according to the AO system and indicated their preferred treatment of the fracture (operative or nonoperative). Agreement among observers was assessed with use of the multirater kappa (κ) measure.

Results: Interobserver reliability of the AO classification, fracture characteristics, and preferred treatment generally ranged from “slight” to “fair.” A few small but statistically significant differences were found. Observers randomized to the 2-D CT group had slightly but significantly better agreement on displacement of the greater tuberosity ($\kappa = 0.35$ compared with 0.30, $p < 0.001$) and on the AO classification ($\kappa = 0.18$ compared with 0.17, $p = 0.018$). A subgroup analysis of the AO classification results revealed that shoulder and elbow surgeons, orthopaedic trauma surgeons, and surgeons in the United States had slightly greater reliability on 2-D CT, whereas surgeons in practice for ten years or less and surgeons from other subspecialties had slightly greater reliability on 3-D CT.

Conclusions: Proximal humeral fracture classifications may be helpful conceptually, but they have poor interobserver reliability even when 3-D rather than 2-D CT is utilized. This may contribute to the similarly poor interobserver reliability that

continued

*The Science of Variation Group: Parag Melvanki, Rudolf W. Poolman, Brett D. Crist, Lars C. Borris, Vishwanath M. Iyer, Reto H. Babst, Robert D. Zura, Huub van der Heide, Frede Frihagen, Iain McGraw, Eckart Schwab, George Thomas, Axel Jubel, James Kellam, Andrew Schmidt, Philipp Lenzlinger, Fred Baumgaertel, Elena Grosso, Matt Mormino, Marc Swiontkowski, Kyle Jeray, Daphne Beingessner, Neeraj Bijlani, Michael Prayson, Ladislav Mica, David Sonnabend, Darren Drosdowech, Francisco Lopez-Gonzalez, W. Jaap Willems, Frank Walter, Charalampos Zalavras, Richard S. Page, Thomas Wright, Scott Duncan, Taco Gosens, George S.M. Dyer, Grant Garrigues, German Ricardo Hernandez, Jeffrey A. Greenberg, Phani Dantuluri, Jose A. Ortiz Jr., Charles Cassidy, Alberto Pérez Castillo, Rick Papandrea, Sanjeev Kakar, Steve Kronlage, Leon Benson, Julie Adams, Lawrence Weiss, Gustavo Mantovani Ruggiero, Jay Pomerance, Ramon de Bedout, Eric Hofmeister, Marc J. Richard, Fabio Suarez, Theresa Wyrick, Michael Baskies, Thomas Hughes, Neil Wilson, J.C. Goslings, Thakkar Navin, Kevin Eng, Qiugen Wang, Pradeep Choudhari, Henry Broekhuysse, Richard Jenkinson, K.J. Ponsen, Arie B. van Vugt, Leon Elmans, Steven J. Rhemrev, Peter Kloen, Andreas Platz, Peter R.G. Brink, Rajat Varma, Alan Kawaguchi, James V. Nepola, Thomas DeCoster, Raymond Malcolm Smith, Kenneth Egol, Joseph M. Conflitti, Antonio Barquet, Rodrigo Pesantez, Jin-Young Park, Chunyan Jiang, Martin Richardson, Jeremy Hall, George Kontakis, Denise Eygendaal, Augustus D. Mazocco, Xavier A. Duralde, Gerald Williams Jr., Donald Endrizzi, Steven J. Hattrup, Steve Petersen, Sergio L. Checchia, Sander Spruijt, Jason C. Fanuele, Taizoon Baxamusa, Chris Wilson, Francisco Javier Aguilar Sierra, Jorge Boretto, Karel Chivers, Jorge Rubio, Peter Schandelmaier, John L. Marsh, and Giuseppe Porcellini.

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.

was observed for selection of the treatment for proximal humeral fractures. The lack of a reliable classification confounds efforts to compare the outcomes of treatment methods among different clinical trials and reports.

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

The Neer and AO (Arbeitsgemeinschaft für Osteosynthesefragen) classifications of proximal humeral fractures have limited intraobserver and interobserver reliability¹⁻⁴. Neer indicated that his classification was meant to be applied after operative exposure and believed that radiographs alone would be unreliable⁵; however, treatment protocols and scientific experiments rely on accurate and reliable fracture characterization prior to surgery. Some investigators have reported that the addition of two-dimensional (2-D) computed tomography (CT) scans did not improve the interobserver reproducibility of either of these classification systems⁶.

Recently, Foroohar et al. studied interobserver agreement among sixteen observers (including four upper extremity specialists, four general orthopaedic surgeons, four senior orthopaedic residents, and four junior orthopaedic residents) who utilized radiographs, 2-D CT scans, and three-dimensional (3-D) CT scans to classify such fractures⁷. There was slight agreement on the Neer classification, as indicated by the kappa measure ($\kappa = 0.069$ to 0.14), and fair agreement on treatment ($\kappa = 0.28$ to 0.33) across all three modalities⁷, but neither characteristic exhibited strong performance for clinical or research use.

The aim of the present study was to test the null hypothesis that interobserver reliability of the AO and Neer classifications, preferred treatment, and fracture characteristics is the same for 2-D and 3-D CT. A large cohort of practicing surgeons underwent randomization to review either 2-D or 3-D CT scans along with radiographs.

Materials and Methods

The study protocol was approved by the institutional review board at the principal investigator's hospital. We invited the members of the Science of Variation Group (fully trained practicing orthopaedic and trauma surgeons from around the world) to participate in an online evaluation of the characteristics, classification, and preferred treatment of proximal humeral fractures. The only incentive for observers to participate was the group authorship of the present manuscript.

We constructed a list of consecutive patients with a proximal humeral fracture, treated at one institution from 2000 to 2010, for which both high-quality CT scans (slice thickness, ≤ 1.25 mm) and anteroposterior and axillary or scapular-Y radiographs were available. Fifteen fractures were then selected by general consensus of the authors to represent a full spectrum of proximal humeral fracture patterns. The approach of the Science of Variation Group is to gain statistical power through increasing the number of observers rather than the number of observations per observer; this limits the burden on individual observers and thereby hopefully increases participation, allows for more comparisons among observers, and improves external validity.

The 2-D CT slices were converted into a movie (in AVI format) that the observers could scroll through in three dimensions (in the transverse, sagittal, and coronal planes). Two sets of 3-D reconstructions were also made; one set was in the form of a movie showing a vertically and horizontally rotating humerus with a proximal fracture, and the other set showed both the humerus and the surrounding articulating structures rotating around a horizontal as well as a vertical axis.

The members of the Science of Variation Group were randomized in an ideally 1:1 ratio (with use of a computerized random-number generator at the time of invitation) to review radiographs and either 2-D or 3-D CT scans. A total of 371 invitations were sent, of which 193 (52%) went to the 2-D group and 178 (48%) went to the 3-D group. Responses were received from 135 surgeons (36%); sixty-two respondents were from the 2-D group and seventy-three were from the 3-D group. Of the respondents, 107 (79%) completed the online survey; 93% of these were male and 7% were female. Forty-six (43%) of these 107 surgeons had been randomized to the 2-D group and sixty-one

TABLE I Observer Demographics

	2-D (N = 46)		3-D (N = 61)		Total (N = 107)
	N	%	N	%	N
Sex					
Male	44	96	55	90	99
Female	2	4	6	10	8
Area					
United States	16	35	33	54	49
Europe	12	26	13	21	25
Asia	5	11	3	5	8
Canada	3	7	1	2	4
U.K.	1	2	2	3	3
Australia	2	4	2	3	4
Other	7	15	7	11	14
Years in independent practice					
0-5	8	17	15	25	23
6-10	8	17	14	23	22
11-20	17	37	19	31	36
21-30	13	28	13	21	26
Specialization					
General orthopaedic surgery	2	4	1	2	3
Orthopaedic traumatology	22	48	25	36	44
Shoulder and elbow	14	30	12	20	26
Hand and wrist	7	15	20	33	27
Other	1	2	3	5	4
Supervises trainees in the operating room					
Yes	41	89	51	84	92
No	5	11	10	16	15

TABLE II Overall Interobserver Agreement

Question	2-D			3-D			P Value
	κ	Agreement	95% CI*	κ	Agreement	95% CI*	
Is the humeral head split?	0.26	Fair	0.12, 0.40	0.30	Fair	0.20, 0.41	0.594
Is the greater tuberosity displaced?	0.35	Fair	0.33, 0.37	0.30	Fair	0.28, 0.32	<0.001*
Is the arterial supply compromised?	0.19	Slight	0.14, 0.25	0.22	Fair	0.20, 0.23	0.438
Is the glenohumeral joint dislocated?	0.12	Slight	-0.12, 0.36	0.09	Slight	-0.09, 0.27	0.846
What is the AO classification?	0.18	Slight	0.17, 0.19	0.17	Slight	0.16, 0.18	0.018*
What is the preferred treatment?	0.25	Fair	0.18, 0.32	0.24	Fair	0.18, 0.30	0.828

*Significant.

(57%), to the 3-D group. After they had logged on, observers were asked where they practiced, how long they had been in independent practice, whether or not they supervised trainees in the operating room, and what their orthopaedic subspecialty was.

Each observer answered the following four yes-or-no questions for each of the fifteen proximal humeral fractures: (1) Is the greater tuberosity displaced? (2) Is the humeral head split? (3) Is the arterial supply compromised? (4) Is the glenohumeral joint dislocated? He or she also classified the fracture according to the AO system⁸ and indicated the preferred treatment (operative, percutaneous pinning, open reduction and internal fixation, or hemiarthroplasty) by selecting it from a pull-down menu. No training or criteria were provided.

In addition, each observer was asked to classify the fracture according to the Neer system by choosing the correct classification from a subset of pull-down menus. The observer first had to choose the number of "parts," between one and four, and then use the corresponding pull-down menu to choose the structures involved. However, this arrangement proved unsuccessful as only seventeen observers used it correctly.

Statistical Analysis

Interobserver agreement was determined with use of the multirater kappa measure described by Siegel and Castellan⁹, which is a frequently used measure

of chance-corrected agreement between ratings made by multiple observers (interobserver reliability). The generated kappa values were interpreted according to the guidelines of Landis and Koch: a value of 0.01 to 0.20 indicates slight agreement; 0.21 to 0.40, fair agreement; 0.41 to 0.60, moderate agreement; 0.61 to 0.80, substantial agreement; and 0.81 to 0.99, almost perfect agreement. Zero indicates no agreement beyond that expected because of chance alone; -1.00, total disagreement; and +1.00, perfect agreement¹⁰.

A post hoc power analysis was performed with use of nQuery Advisor software (version 7.0; Statistical Solutions, Saugus, Massachusetts). It was calculated that the forty-six observers in the 2-D group and sixty-one observers in the 3-D group would have yielded 80% power to detect a difference of 0.015 in the kappa value for the AO fracture classification.

Source of Funding

No external funding was received for this study.

Results

The forty-six observers who were randomized to review radiographs and 2-D CT scans and the sixty-one observers who were randomized to review radiographs and 3-D CT scans

TABLE III Interobserver Agreement on AO Classification According to Observer Demographics and Imaging Modality

	2-D CT				Radiography and 3-D CT				P Value
	N	Agreement	κ	SE*	N	Agreement	κ	SE*	
Area									
Europe and U.K.	13	Slight	0.14	0.012	15	Slight	0.14	0.01	0.65
United States and Canada	19	Fair	0.23	0.011	34	Slight	0.18	0	<0.001†
Other	14	Slight	0.16	0.012	12	Slight	0.15	0.01	0.40
Years in independent practice									
0-10	16	Slight	0.14	0.01	29	Slight	0.17	0.01	0.014†
11-20	17	Slight	0.20	0.01	19	Slight	0.19	0.01	0.22
>20	13	Slight	0.18	0.02	13	Slight	0.15	0.01	0.12
Specialization									
Orthopaedic traumatology	22	Slight	0.20	0.008	25	Slight	0.16	0.01	0.002†
Shoulder and elbow	14	Fair	0.24	0.016	12	Slight	0.18	0.02	0.008†
Other	10	Slight	0.09	0.015	24	Slight	0.17	0.01	<0.001†

*SE = standard error. †Significant.

TABLE IV Kappa Compared with Overall Rate of Agreement

Question	2-D			3-D		
	κ	Agreement	Overall Rate of Agreement	κ	Agreement	Overall Rate of Agreement
Is the humeral head split?	0.26	Fair	0.95	0.30	Fair	0.97
Is the greater tuberosity displaced?	0.35	Fair	0.96	0.30	Fair	0.97
Is the arterial supply compromised?	0.19	Slight	0.92	0.22	Fair	0.95
Is the glenohumeral joint dislocated?	0.12	Slight	0.87	0.09	Slight	0.87
What is the AO classification?	0.18	Slight	0.96	0.17	Slight	0.97
What is the preferred treatment?	0.25	Fair	0.96	0.24	Fair	0.97

were comparable except for the distribution of surgical subspecialties. The 3-D CT group had a greater percentage of hand surgeons and surgeons in the United States as well as a lower percentage of shoulder and elbow surgeons and orthopaedic traumatologists (Table I).

The observers randomized to the 2-D CT group had minimally, but significantly, better agreement on displacement of the greater tuberosity ($\kappa = 0.35$ compared with 0.30 for the observers randomized to the 3-D group) and on the AO classification ($\kappa = 0.18$ compared with 0.17). However, the agreement was no better than fair for any question (Table II).

Shoulder and elbow surgeons, orthopaedic trauma surgeons, and surgeons in the United States had greater reliability for the AO classification on 2-D CT, whereas surgeons in practice for ten years or less and surgeons from other subspecialties had greater reliability on 3-D CT. However, the differences were again small and the reliability was no better than fair (Table III).

European surgeons achieved moderate agreement on the splitting of the humeral head ($\kappa = 0.42$), and American surgeons achieved moderate agreement on displacement of the greater tuberosity ($\kappa = 0.42$), on 2-D CT. The corresponding agreement using 3-D CT was fair in both cases, and all other ratings by the three geographic subgroups (North America, Europe, and other) had slight or fair agreement. Surgeons with more than twenty years in independent practice achieved moderate agreement on displacement of the greater tuberosity ($\kappa = 0.41$) on 2-D CT but fair agreement using 3-D CT; agreement on all other ratings by the subgroups defined on the basis of years in practice was slight or fair. Shoulder and elbow surgeons achieved moderate agreement on the displacement of the greater tuberosity on both 2-D and 3-D CT, but agreement on all other ratings in the subspecialty groups was slight or fair. Agreement on preferred treatment was fair or worse in all subgroups (see Appendix).

Discussion

Consistent with Foroohar et al.⁷, we documented slight to fair agreement overall on proximal humeral fracture classification, characterization, and treatment on both 2-D and 3-D CT scans combined with radiographs. In fact, interobserver agreement was slightly but significantly lower overall on displacement of the greater tuberosity and on the

AO classification among surgeons viewing 3-D reconstructions. In particular, 3-D CT was beneficial only for surgeons with a subspecialty other than shoulder and elbow or orthopaedic traumatology and for surgeons with less than ten years in independent practice. One could argue that surgeons who are not used to looking at radiographs and 2-D CT scans benefit from the more intuitive 3-D reconstructions, but the differences were very small and should not be overinterpreted, particularly in the context of the limited reliability overall. Also, cost is a lesser issue because the 3-D reconstructions add only \$100 at our institution and can be done for free with use of files in DICOM format and software such as OsiriX (Pixmeo, Bernex, Switzerland).

The strengths of this study include the large number of surgeons, which allowed randomization and subgroup analysis, and the fact that all observers were experienced, practicing surgeons. The study should be interpreted in light of several limitations. First, the data may be subject to the so-called “kappa paradox” because the kappa measure was considerably lower than the overall percentage of agreement (Table IV). If the prevalence of an outcome is low, it causes an imbalance in the marginal totals, generating a lower kappa than one might expect. Consequently, we plan to study the ability of observers to distinguish among a few of the most common fracture types in future studies. Second, the observers had no patient-specific information (e.g., comorbidities, level of activity, and age) on which to base treatment recommendations. Third, we did not train the observers or provide them with measurement tools (for instance, no criteria were specified for greater tuberosity displacement or head splitting), as we were interested in the surgeons’ general impressions based on their experience and training. Fourth, there may be important differences between the web interface that we utilized and the usual way in which doctors view radiographs and CT scans. Finally, for practical purposes we chose to limit the study to two classifications systems even though others merit testing. For instance, one study suggested that the classification of Hertel and colleagues—which is widely used in Europe—may have better interobserver agreement, although that agreement was still only moderate¹¹.

The Neer classification system was originally intended as a classification based on intraoperative findings⁵, and the low

reliability found in previous studies might be explained by the attempt to classify fractures on the basis of preoperative imaging. Also, the observers in the present study were not able to navigate the pull-down menu for the Neer classification; instead of choosing the number of "parts" between one and four and subsequently choosing the structures involved from the corresponding pull-down menu, they chose the number of parts from each pull-down menu. Only seventeen observers used this menu correctly, and this might emphasize the merit of training in the use of the Neer classification.


To date, the consistently low interobserver agreement of proximal humeral fracture classification has not been consistently or substantially improved by more sophisticated imaging or training^{1,3,6,12-16}. One exception involves the study by Brunner et al., who described an improvement in agreement (from moderate to good) on the AO and Neer classifications with the use of stereoscopic visualization and interactive 3-D imaging by four independent observers classifying forty proximal humeral fractures¹⁷.

Brorson et al. demonstrated that training of observers holds promise. They randomized observers to either two training sessions or no training, then had them classify forty-two fractures according to the Neer system. The mean difference in kappa between the groups with and without training was 0.30 (95% confidence interval [CI], 0.10 to 0.50; $p = 0.006$). The mean kappa for interobserver variation improved from 0.27 (95% CI, 0.24 to 0.31) to 0.62 (95% CI, 0.57 to 0.67) with the ad-

dition of training¹⁸. Training also improves the reliability of determining characteristics of other fractures, such as the diagnosis of scaphoid fracture displacement¹⁹.

Although the Neer and AO classifications are helpful conceptually, a reliable classification is necessary in order to be confident that the fractures treated in one trial are comparable with those treated in another. Perhaps a simpler classification would have better interobserver reliability. In future studies, we plan to test the ability of surgeons to reliably distinguish among a few common fracture patterns.

Appendix

 Tables showing interobserver agreement according to country of origin, years in independent practice, and specialization are available with the online version of this article as a data supplement at jbjs.org. ■

Wendy E. Bruinsma, MD
Thierry G. Guitton, MD, PhD
Jon J.P. Warner, MD
David Ring, MD, PhD
Department of Orthopaedic Surgery,
Massachusetts General Hospital,
55 Fruit Street,
Boston, MA 02114.
E-mail address for D. Ring: dring@partners.org

References

- Bernstein J, Adler LM, Blank JE, Dalsey RM, Williams GR, Iannotti JP. Evaluation of the Neer system of classification of proximal humeral fractures with computerized tomographic scans and plain radiographs. *J Bone Joint Surg Am*. 1996 Sep;78(9):1371-5.
- Siebenrock KA, Gerber C. The reproducibility of classification of fractures of the proximal end of the humerus. *J Bone Joint Surg Am*. 1993 Dec;75(12):1751-5.
- Sallay PI, Pedowitz RA, Mallon WJ, Vandemark RM, Dalton JD, Speer KP. Reliability and reproducibility of radiographic interpretation of proximal humeral fracture pathoanatomy. *J Shoulder Elbow Surg*. 1997 Jan-Feb;6(1):60-9.
- Kristiansen B, Andersen UL, Olsen CA, Varmarken JE. The Neer classification of fractures of the proximal humerus. An assessment of interobserver variation. *Skeletal Radiol*. 1988;17(6):420-2.
- Neer CS 2nd. Four-segment classification of proximal humeral fractures: purpose and reliable use. *J Shoulder Elbow Surg*. 2002 Jul-Aug;11(4):389-400.
- Sjödén GO, Movin T, Güntner P, Aspelin P, Ahrengart L, Ersmark H, Sperber A. Poor reproducibility of classification of proximal humeral fractures. Additional CT of minor value. *Acta Orthop Scand*. 1997 Jun;68(3):239-42.
- Foroohar A, Tosti R, Richmond JM, Gaughan JP, Ilyas AM. Classification and treatment of proximal humerus fractures: inter-observer reliability and agreement across imaging modalities and experience. *J Orthop Surg Res*. 2011;6:38. Epub 2011 Jul 29.
- Müller ME, Nazarian S, Koch P, Schatzker J. The comprehensive classification of fractures of long bones. Berlin: Springer; 1990.
- Siegel S, Castellan NJ. Nonparametric statistics for the behavioral sciences. 2nd ed. New York: McGraw-Hill; 1988.
- Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977 Mar;33(1):159-74.
- Majed A, Macleod I, Bull AM, Zyto K, Resch H, Hertel R, Reilly P, Emery RJ. Proximal humeral fracture classification systems revisited. *J Shoulder Elbow Surg*. 2011 Oct;20(7):1125-32. Epub 2011 Apr 09.
- Castagno AA, Shuman WP, Kilcoyne RF, Haynor DR, Morris ME, Matsen FA. Complex fractures of the proximal humerus: role of CT in treatment. *Radiology*. 1987 Dec;165(3):759-62.
- Mora Guix JM, Gonzalez AS, Brugalla JV, Carril EC, Baños FG. Proposed protocol for reading images of humeral head fractures. *Clin Orthop Relat Res*. 2006 Jul;448:225-33.
- Sjödén GO, Movin T, Aspelin P, Güntner P, Shalabi A. 3D-radiographic analysis does not improve the Neer and AO classifications of proximal humeral fractures. *Acta Orthop Scand*. 1999 Aug;70(4):325-8.
- Brorson S, Bagger J, Sylvest A, Hróbjartsson A. Low agreement among 24 doctors using the Neer-classification; only moderate agreement on displacement, even between specialists. *Int Orthop*. 2002;26(5):271-3. Epub 2002 Jun 08.
- Brorson S, Bagger J, Sylvest A, Hróbjartsson A. Diagnosing displaced four-part fractures of the proximal humerus: a review of observer studies. *Int Orthop*. 2009 Apr;33(2):323-7. Epub 2008 Jun 07.
- Brunner A, Honigsmann P, Treumann T, Babst R. 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. *J Bone Joint Surg Br*. 2009 Jun;91(6):766-71.
- Brorson S, Bagger J, Sylvest A, Hróbjartsson A. Improved interobserver variation after training of doctors in the Neer system. A randomised trial. *J Bone Joint Surg Br*. 2002 Sep;84(7):950-4.
- Buijze GA, Guitton TG, van Dijk CN, Ring D; Science of Variation Group. Training improves interobserver reliability for the diagnosis of scaphoid fracture displacement. *Clin Orthop Relat Res*. 2012 Jul;470(7):2029-34.