Original Article

Clin Shoulder Elbow 2022;25(4):304-310
https://doi.org/10.5397/cise.2022.01025

CiSE
Clinics in Shoulder and Elbow

eISSN 2288-8721

What is the interobserver agreement of displaced humeral surgical neck fracture patterns?

Reinier W. A. Spek[1,2,3,*], Laura J. Kim[4,*], Traumaplatform 3D Consortium

[1]*Department of Orthopaedic Surgery, Flinders Medical Centre and Flinders University, Adelaide, Australia*
[2]*Department of Orthopaedic Surgery, OLVG, Amsterdam, The Netherlands*
[3]*Department of Orthopaedic Surgery, University Medical Centre Groningen and University of Groningen, Groningen, The Netherlands*
[4]*Department of Trauma Surgery, University Medical Centre Groningen and University of Groningen, Groningen, The Netherlands*

Background: The Boileau classification distinguishes three surgical neck fracture patterns: types A, B, and C. However, the reproducibility of this classification on plain radiographs is unclear. Therefore, we questioned what the interobserver agreement and accuracy of displaced surgical neck fracture patterns is categorized according to the modified Boileau classification. Does the reliability to recognize these fracture patterns differ between orthopedic residents and attending surgeons?
Methods: This interobserver study consisted of a randomly retrieved series of 30 plain radiographs representing clinical practice in a level 1 and a level 2 trauma center. Radiographs were included from patients (≥18 years) who sustained an isolated displaced surgical neck fracture if they were taken ≤1 week after initial injury. A ground truth was established by consensus among three senior orthopedic surgeons. All images were assessed by 17 orthopedic residents and 17 attending orthopedic trauma surgeons.
Results: Agreement for the modified Boileau classification was fair (κ=0.37; 95% confidence interval [CI], 0.36–0.38) with an accuracy of 62% (95% CI, 57%–66%). Comparison of interobserver variability between residents and attending surgeons revealed a significant but clinically irrelevant difference in favor of attending surgeons (0.34 vs. 0.39, respectively, Δ κ=0.05, 95% CI, 0.02–0.07).
Conclusions: The modified Boileau classification yields a low interobserver agreement with an unsatisfactory accuracy in a panel of orthopedic residents and attending surgeons. This supports the hypothesis that surgical neck fractures are challenging to categorize and that this classification should not be used to determine prognosis if only plain radiographs are available.

Keywords: Surgical neck fractures; Proximal humerus fracture; Shaft translation; Boileau classification; Interobserver variability

INTRODUCTION

Two-part surgical neck fractures of the humerus entail 28% of proximal humerus fractures and can be treated nonoperatively or by several surgical modalities (e.g., plate fixation and intramedullary nailing) [1-3]. However, substantial treatment variability is

Received: October 6, 2022 Revised: October 6, 2022 Accepted: October 6, 2022
Correspondence to: Reinier W. A. Spek
Department of Orthopaedic Surgery, Flinders Medical Centre and Flinders University, Dr. Bedford Park SA 5042, Adelaide, Australia
Tel: +61-08-8204 4289, Fax: +61-08-8374 0832, E-mail: reinierspek@gmail.com, ORCID: https://orcid.org/0000-0002-7509-6508
*These authors contributed equally to this work.

Financial support: The corresponding author (RWAS) has received payments during the study period, in amounts of less than USD 10,000 from Michael van Vloten Fonds (Rotterdam, The Netherlands), less than USD 10,000 from Anna Fonds NOREF (Mijdrecht, The Netherlands), less than USD 10,000 from Flinders Foundation (Adelaide, Australia), and between USD 10,000 and USD 100,000 from Prins Bernhard Cultuurfonds (Amsterdam, The Netherlands). No funding was received to carry out this specific study.
Conflict of interest: None.

observed between clinicians, hospitals, and even among countries [4]. Among other things, classification of the fracture is important for determining the optimal treatment [5]. Ideally, classification should guide the surgeons' decision-making and be taken into account to determine the optimal treatment for proximal humerus fractures.

Currently available classification systems for surgical neck fractures are the fracture patterns according to Neer [6] and Arbeitsgemeinschaft für Osteosynthesefragen (AO) [7]. Neer created three subgroups (impacted angulated, separated, and comminuted two-part surgical neck fractures), while the AO created two subgroups (impacted and non-impacted two-part surgical neck fractures). Nevertheless, clinical implications of these distinct fracture patterns are unclear.

To determine the optimal entry point for intramedullary nailing, Boileau et al. [8] developed a new classification system which categorized displaced surgical neck fractures into three types: type A, partial medial shaft translation with valgus humeral head angulation; type B, entire medial shaft translation without humeral head tilt or angulation; and type C, lateral shaft translation with varus humeral head angulation. Although numerous studies have investigated the agreement on the full array of two-, three-, and four-part proximal humerus fractures, no interobserver study has been carried out regarding surgical neck fracture patterns in particular [9,10]. A reproducible fracture classification is a prerequisite to comparing patient outcomes of different clinical trials [5]. Moreover, if a high level of agreement can be reached, fracture patterns could potentially influence surgical decision-making and might predict prognosis.

The Boileau classification was originally based on radiographs and computed tomography (CT) scans, but as CT scans are not routinely available for every patient, this study aimed to assess its reproducibility on plain radiographs. The following research questions were asked: what is the interobserver agreement and accuracy of displaced surgical neck fracture patterns categorized according to the modified Boileau criteria? And does the reliability to recognize these fracture patterns differ between orthopedic residents and attending surgeons?

METHODS

Ethical approval was received from OLVG (Amsterdam, The Netherlands, No. 19.135) and Flinders Medical Centre (Adelaide, Australia, No. 234.19). Informed consent from patients was waived.

Setting and Study Design

This is an interobserver study in which 30 radiographs were assessed and categorized according to the modified Boileau classification of displaced surgical neck fractures [8]. The study was carried out in March and April 2021, and an observer panel was created with participants from the orthopedic and trauma units of four different teaching hospitals. The panel consisted of 17 orthopedic residents and 17 attending orthopedic trauma surgeons with different levels of experience and subspecialties.

Images

Anteroposterior (true or standard) and lateral radiographic views were included from patients (≥18 years) who sustained an isolated displaced surgical neck fracture which could be classified according to the Boileau classification. Patients were deemed eligible irrespective of the treatment provided; thus, trauma radiographs of both non-operatively treated patients and surgically-treated patients were included. Patients were excluded if they presented to the emergency department more than 1 week after the initial injury or had a concomitant fracture (Hill-Sachs lesion, proximal humerus, humeral shaft, or pathologic fracture).

Classification

Boileau et al. [8] developed this classification system to categorize displaced surgical neck fractures into three types: type A, partial medial shaft displacement with valgus humeral head angulation; type B, entire medial shaft translation without humeral head tilt; and type C, lateral shaft displacement with varus humeral head angulation. A fracture was considered displaced if it was translated >25% of the humeral midshaft width. Displacement was measured from the outer cortex of the most proximal part of the humeral shaft fragment to the outer cortex of the most distal humeral head fragment. To cover all displaced surgical neck fractures, an additional category was incorporated in this study: "non-classifiable." This meant that the head angulation and humeral shaft translation did not match Boileau criteria (e.g., partial lateral humeral shaft translation without head angulation). Therefore, four categories could be chosen by the observers: type A, type B, type C, or non-classifiable (Fig. 1).

Selection of Radiographs

Radiographs of eligible patients were collected from a level 1 trauma center in Australia (March 1, 2016, to July 31, 2020) and a level 2 trauma center in the Netherlands (January 1, 2004, to June 30, 2018). A total of 614 surgical neck fractures were identified, of which 236 patients had a displaced fracture. Among these displaced fractures, 121 patients could be classified according to Boileau classification (type A, n=41; type B, n=20; type C, n=60). While maintaining this mutual distribution between the

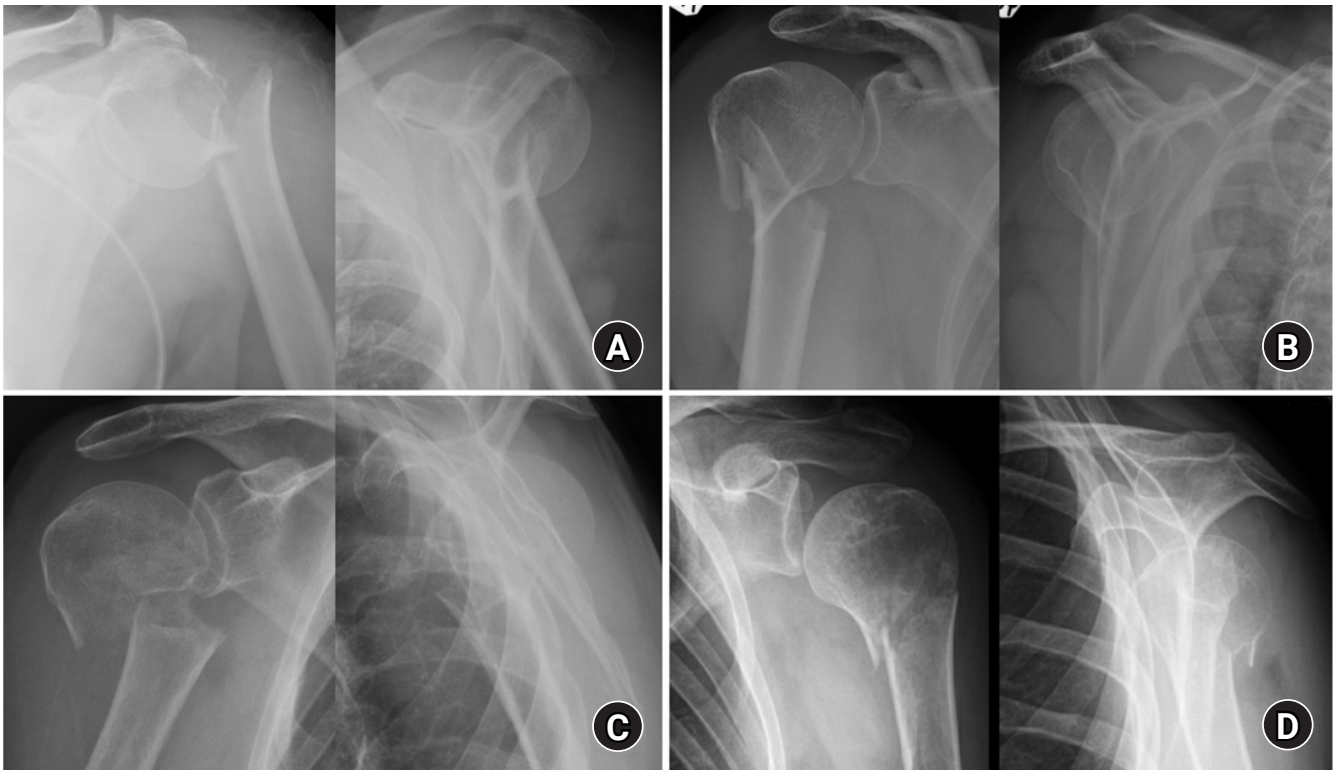

Fig. 1. The modified Boileau classification covers four options: type A, type B, type C, and non-classifiable displaced surgical neck fractures. (A) Type A: medial shaft translation with valgus humeral head tilt. (B) Type B: entire medial (or ventral) shaft translation without humeral head tilt. (C) Type C: lateral shaft displacement with varus angulation of the head. (D) Non-classifiable: shaft translation and/or head angulation do not match with Boileau classification. In this example, there is no varus angulation of the head, meaning it could not be classified according to Boileau. Type A and C were used for training; type B and the non-classifiable radiograph were used for the actual assessments.

three Boileau types, we randomly selected 9 type A fractures, 5 type B, 11 type C, and 5 non-classifiable fractures. The number selected for the non-classifiable category was equal to that of the group with the lowest number (i.e., type B fractures). Randomization was carried out in Microsoft Excel version 2102 (Microsoft Corp., Redmond, WA, USA) by assigning a randomization number which was sorted from low to high. Cases with the lowest randomization number were selected until the predefined sample size (n = 30) was reached. The mean age (range) of included patients was 72.4 years (29–96 years), and the majority were females (80%).

Ground Truth

A ground truth was generated by consensus among three senior orthopedic attending surgeons (two with > 20 years of experience and one with > 15 years of experience after finishing their training). Each of these orthopedic surgeons completed the study prior to the consensus meeting, so they classified all fractures independently before answers were compared. The meeting was led by the first author (RWAS), and discrepancies were resolved by discussion.

Observer Panel

The observer panel consisted of 34 participants: 17 orthopedic residents and 17 attending orthopedic surgeons. Six attending orthopedic surgeons had < 5 years of experience. All other attending surgeons had > 5 years of experience: five were seniors (> 20 years of experience), three were shoulder specialists (they completed fellowship training on the upper extremity), two were dedicated attending trauma surgeons, and one was an orthopedic oncologist. All attending surgeons had substantial experience in treating trauma, and years of experience was defined as years in clinical practice after finishing the training program.

Training and Assessment

Prior to assessment, each observer received training in recognizing the fracture patterns according to Boileau classification. The first part of the training consisted of an explanation of the fracture patterns and the following rules: (1) dorsal head angulation is not considered (e.g., medial translation with valgus head angulation and dorsal head angulation should be classified as a type A fracture) and (2) type B fractures require entire medial or entire ventral humeral shaft translation. It was also emphasized that

https://doi.org/10.5397/cise.2022.01025

both head angulation and shaft displacement had to match Boileau criteria (e.g., medial humeral shaft translation with varus angulation should be categorized as non-classifiable). Following this, four training cases were provided (one case covering each category) (Fig. 2). At the discretion of observers, training was provided either face-to-face (by RWAS or LK) or as self-study via REDCap [11,12]. Face-to-face training was provided to 73.5% of observers, and 26.5% followed the self-study on REDCap.

Each observer classified 30 displaced surgical neck fractures with both anteroposterior and lateral views. Questions and radiographs were both presented on-screen. Illustration sheets depicting the classification system were displayed during the observation. There was no time limit on assessment, and radiographs were presented in the identical order for each observer. Observers could not use radiographic measurement tools. However, they could go back if needed and adjust their answer for each radiograph.

Statistical Analysis

IBM SPSS software ver. 27 (IBM Corp., Armonk, NY, USA) was used for statistical analysis. To determine interobserver variability, the multi-rater Fleiss' kappa (κ) was calculated. Values were interpreted according to Landis and Koch: κ <0.00 (poor), κ=0.00–0.20 (slight), κ=0.21–0.40 (fair), κ=0.41–0.60 (moderate), κ=0.61–0.80 (substantial), and κ=0.81–1.00 (almost perfect) [13]. Accuracy was defined as the degree to which each given answer corresponded with the ground truth and expressed as a percentage from 0 to 100. If the accuracy was 0%, no cases were classified the same as the ground truth. If the accuracy was 100%, all cases were classified the same as the ground truth. To calculate accuracy, the accuracy per observer was determined and subsequently averaged across all participants. To compare residents versus attending surgeons, delta (Δ) κ was computed and depicted with a two-tailed p-value. Accuracy among residents and attending surgeons was compared with an independent samples t-test. Multi-rater Fleiss' κ as well as accuracy was displayed with a 95% confidence interval (CI).

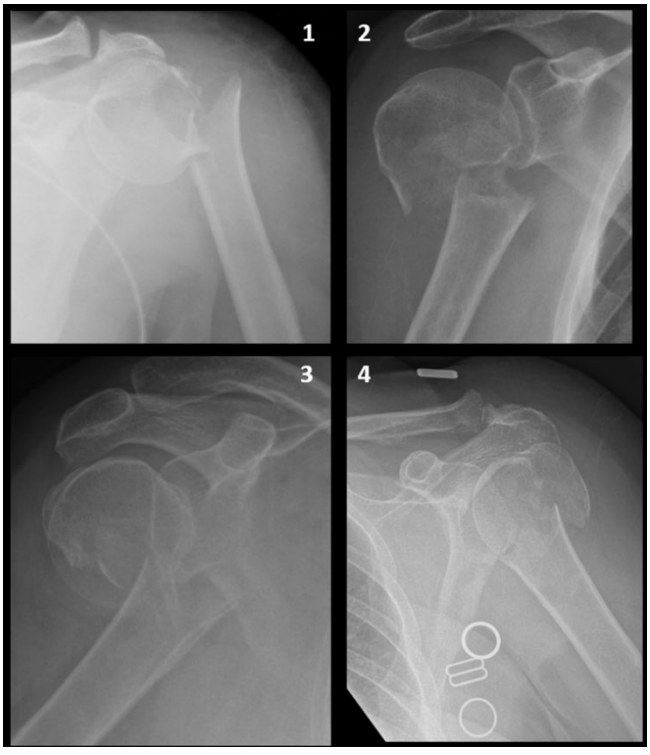

Fig. 2. Radiographs used for training, shown in order from 1 to 4, with 1=type C, 2=type A, 3=type B, and 4=non-classifiable. Although present on image 3 and 4, fracture dislocations and concomitant greater tuberosity fractures were not included in the actual assessment. This was explained to the observers accordingly.

RESULTS

Interobserver Variability and Accuracy

Interobserver agreement to classify fractures according to the modified Boileau criteria among all observers was fair (κ=0.37; 95% CI, 0.36–0.38) (Fig. 3). In type A and C fractures, concordance was moderate (κ=0.42; 95% CI, 0.40–0.44 and κ=0.58; 95% CI, 0.57–0.59, respectively). Observers disagreed the most on type B (κ=0.23; 95% CI, 0.21–0.25) and non-classifiable fractures (κ=0.18; 95% CI, 0.16–0.20). Accuracy amongst all participants was 62% (95% CI, 57%–66%) and the highest for type C fractures, 79% (95% CI, 74%–85%) (Table 1).

Residents vs. Attending Surgeons

Comparison of interobserver variability between residents and

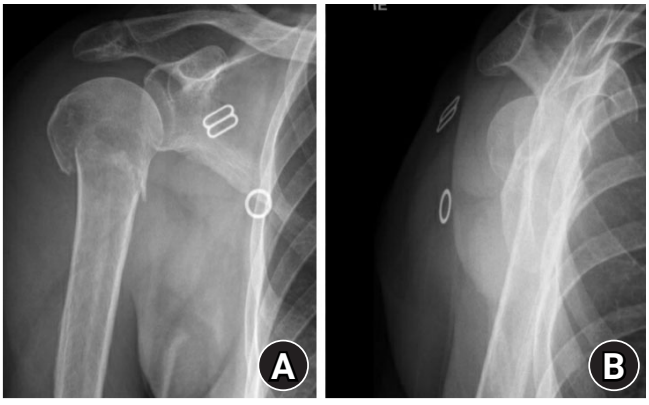

Fig. 3. Assessment of a radiograph with substantial variability amongst the observers: 53% classified this as type A (18 observers), 3% as type B (1 observer), 3% as type C (1 observer), and 41% as "non-classifiable" (14 observers). (A) Standard anterior-posterior view. (B) Lateral view.

Table 1. Agreement and accuracy among all observers

Category	Kappa (95% CI)	Agreement	Accuracy (95% CI), %
Overall	0.37 (0.36–0.38)	Fair	62 (57–66)
Type A	0.42 (0.40–0.44)	Moderate	64 (57–71)
Type B	0.23 (0.21–0.25)	Fair	69 (59–79)
Type C	0.58 (0.57–0.59)	Moderate	79 (74–85)
Non-classifiable	0.18 (0.16–0.20)	Slight	57 (49–65)

Type A: medial shaft translation with valgus humeral head tilt, Type B: entire medial (or ventral) shaft translation without humeral head tilt, Type C: lateral shaft displacement with varus angulation of the head, Non-classifiable: shaft translation and/or head angulation do not match with Boileau classification.
CI: confidence interval.

attending surgeons revealed a significant but intuitively clinically irrelevant difference in favor of attending surgeons (fair vs. fair, Δ κ = 0.05; 95% CI, 0.02–0.07). Residents showed an accuracy of 60% (95% CI, 55–65) in correctly classifying the fractures, whereas attending surgeons revealed an accuracy of 63% (95% CI, 55%–72%). No statistically significant difference was found between both groups (Δ κ = 0.03; 95% CI, –0.06 to 0.12) (Table 2).

Table 2. Agreement and accuracy compared between 17 residents and 17 attending surgeons

Parameter	Kappa (95% CI)	Agreement	Accuracy (95% CI), %
Resident	0.34	Fair	60%
Surgeon	0.39	Fair	63%
Delta	0.05		3%
p-value	< 0.001		0.47

CI: confidence interval.

DISCUSSION

Boileau classification is a recently introduced classification to enhance the humeral nail entry point in treatment for displaced surgical neck fractures. Its inter-surgeon reliability on plain radiographs is unclear, hence our aim was to assess the interobserver variability and accuracy. This study revealed an overall kappa of 0.37 with 62% accuracy for the modified Boileau classification on radiographs. The interobserver variability is a measure that represents the extent of variation between observers for the same radiographs expressed as the kappa coefficient and should be considered together with accuracy. A kappa value of 0.38 is relatively low and implies strong variability in classification, which can lead to misdiagnosis and a potential delay in best treatment. In other words, our study demonstrated that 62% of radiographs were classified correctly, but there was substantial disagreement in the misclassified radiographs.

The interobserver reliability of the general AO and full Neer classification systems has been studied intensively. However, many of these studies had a limited number of observers, which could result in overestimation of agreement, and the question remained unanswered as to the interobserver agreement for the subgroups of surgical neck fractures (Neer included three subgroups, and AO included two subgroups) [14,15]. Regarding the AO classification, the largest study included 46 observers and found a kappa of 0.18 [10]. Another study included 18 observers and investigated the agreement on two-, three-, and four-part fractures according to Neer. They revealed a kappa ranging from 0.03 to 0.07 for classifying two-part fractures [9]. Additionally, kappa values do not improve when fractures are assessed with CT scans [8,9,14,16]. Our study therefore demonstrated a better kappa (0.38); however, this is still inadequate for clinical use. Furthermore, the low interobserver agreement of Boileau classification has implications for surgical decision-making in clinical practice: it is unlikely that surgeons can solely rely on radiographs for surgical planning of humeral nailing.

Assessment of three- and four-part proximal humerus fractures is thought to be better among shoulder specialists compared to general orthopedic surgeons [9]. Additionally, some studies advocate that attending surgeons outperform residents [16]. In this study, we did not find a clinically relevant difference between assessments by residents compared to attending surgeons. As opposed to three- and four-part fractures, this study therefore suggests that two-part displaced surgical neck fractures do not require a certain level of expertise, potentially due to their less complex nature or due to the matter that nobody had any experience with this classification.

It has yet to be established whether or not Boileau classification has clinical implications aside from humeral nailing, and if it can determine prognosis. Nevertheless, one could argue that this classification may be useful for decision-making. For instance, in type B fractures, the entire shaft is translated, which, in our experience, may require surgical intervention. Moreover, type C fractures are likely to respond well to non-operative treatment due to traction of the pectoralis major muscle while wearing a collar and cuff. Decision-making in type A fractures could depend on the degree of valgus angulation, as patients with ≥ 160° may be better off with surgical fixation [17].

This work reconfirms the challenges clinicians are facing to improve interobserver agreement for proximal humerus fracture patterns. As the era of artificial intelligence is approaching, it is speculated that we should make a transition to data-driven care: potentially, an algorithm trained on fracture classification by the input of senior surgeons could neutralize current misconceptions

https://doi.org/10.5397/cise.2022.01025

and observation bias [17].

Several shortcomings should be considered: firstly, the quality of radiographs varied as not all radiographs were taken with similar radiographic imaging settings. In some, the patients' true anteroposterior radiographic views were not obtained, which may have changed the perception of humeral shaft translation as well as head angulation. Additionally, internal humeral head rotation makes it difficult to assess head deformity as the greater tuberosity is not well profiled. However, our aim was to evaluate the classification on radiographs, which would reflect the hospital setting well: in clinical practice, it is well known that radiographic quality can be low, and that patients retain their shoulders in internal rotation due to pain. As opposed to the original classification, CT scans were not used for this study. The rationale for assessing this classification was to assess whether it could be applied to all patients presenting at the emergency department, and as CTs are not routinely performed for these patients, this was not feasible. Hence, we coined it the modified Boileau classification: a fourth category (non-classifiable) was added to cover all displaced surgical neck fractures. One could argue that by mitigating these factors, interobserver variability could improve. Secondly, in clinical practice, radiographs are usually discussed between colleagues (e.g., between orthopedic residents and attending surgeons). This is a limitation for interobserver studies in general so it would be interesting to assess its impact on agreement. For instance, during the consensus meeting there was hardly any significant dispute on radiographs even though the attending surgeons classified 12 radiographs differently during initial assessment. This underscores the suggestion that group discussion might improve agreement. Thirdly, the intra-observer agreement was not evaluated.

One of the study strengths was the representativeness of the observer panel, which was a good reflection of potential users of this classification. Displaced surgical neck fractures are hard to classify on plain radiographs: the modified Boileau classification yields a poor interobserver agreement with an accuracy of 62% in a panel of orthopedic residents and attending surgeons with different levels of experience. This suggests that two-part displaced surgical neck fractures do not require a certain level of expertise, and that surgeons cannot rely solely on radiographs for surgical planning of humeral nailing.

ACKNOWLEDGMENTS

The traumaplatform 3D study collaborative:
Henrik Åberg, Anushka Abeywickrama, Michel P. J. van den Bekerom, Wael Chiri, Samantha Damude, Marion M. Deken, Ron L. Diercks, Derek F. P. van Deurzen, Job N. Doornberg, Nathan Eardley-Harris, Anne T. Fokkema, Tom J. Gieroba, H. S. Femke Hagenmaier, Sharon Hendriks, Tanneke I. Herklots, Genevieve S. Hernández, Lotje A. Hoogervorst, Frank F. A. IJpma, Ruurd L. Jaarsma, Bhavin Jadav, Paul C. Jutte, Bas Keizers, Simone F. Kleiss, Maarten C. Koper, Borg Leijtens, Hamid Lutfi, Shoumit Mukhopadhaya, Arthur van Noort, Pradeep M. Poonnoose, Tim Ramsey, Jai Rawat, Jack Richards, Mieke van Suijlichem, Hugo C. van der Veen, Klaus W. Wendt, Roy Zuidema.
Representative trauma platform study Collaborative: Job N. Doornberg.

ORCID

Reinier W. A. Spek https://orcid.org/0000-0002-7509-6508
Laura J. Kim https://orcid.org/0000-0003-1783-0679

REFERENCES

1. Yoon RS, Dziadosz D, Porter DA, Frank MA, Smith WR, Liporace FA. A comprehensive update on current fixation options for two-part proximal humerus fractures: a biomechanical investigation. Injury 2014;45:510-4.
2. Setaro N, Rotini M, Luciani P, Facco G, Gigante A. Surgical management of 2- or 3-part proximal humeral fractures: comparison of plate, nail and K-wires. Musculoskelet Surg 2022;106:163-7.
3. Court-Brown CM, Garg A, McQueen MM. The epidemiology of proximal humeral fractures. Acta Orthop Scand 2001;72:365-71.
4. Launonen AP, Sumrein BO, Reito A, et al. Operative versus non-operative treatment for 2-part proximal humerus fracture: a multicenter randomized controlled trial. PLoS Med 2019;16:e1002855.
5. Handoll HH, Brorson S. Interventions for treating proximal humeral fractures in adults. Cochrane Database Syst Rev 2015;(11):CD000434.
6. Neer CS 2nd. Displaced proximal humeral fractures. I. Classification and evaluation. J Bone Joint Surg Am 1970;52:1077-89.
7. Meinberg EG, Agel J, Roberts CS, Karam MD, Kellam JF. Fracture and Dislocation Classification Compendium-2018. J Orthop Trauma 2018;32 Suppl 1:S1-170.
8. Boileau P, d'Ollonne T, Bessière C, et al. Displaced humeral surgical neck fractures: classification and results of third-generation percutaneous intramedullary nailing. J Shoulder Elbow Surg 2019;28:276-87.
9. 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.
10. Bruinsma WE, Guitton TG, Warner JJ, Ring D; Science of Variation Group. Interobserver reliability of classification and characterization of proximal humeral fractures: a comparison of two and three-dimensional CT. J Bone Joint Surg Am 2013;95:1600-4.
11. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019;95:103208.
12. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap): a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009;42:377-81.
13. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.
14. Iordens GI, Mahabier KC, Buisman FE, Schep NW, Muradin GS, Beenen LF, et al. The reliability and reproducibility of the Hertel classification for comminuted proximal humeral fractures compared with the Neer classification. J Orthop Sci 2016;21:596-602.
15. Marongiu G, Leinardi L, Congia S, Frigau L, Mola F, Capone A. Reliability and reproducibility of the new AO/OTA 2018 classification system for proximal humeral fractures: a comparison of three different classification systems. J Orthop Traumatol 2020;21:4.
16. 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;78:1371-5.
17. Chung SW, Han SS, Lee JW, et al. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. Acta Orthop 2018;89:468-73.

https://doi.org/10.5397/cise.2022.01025