Clin Orthop Relat Res (2022) 480:150-159
DOI 10.1097/CORR.0000000000001921

Clinical Orthopaedics and Related Research®
A Publication of The Association of Bone and Joint Surgeons®

Clinical Research

3D-printed Handheld Models Do Not Improve Recognition of Specific Characteristics and Patterns of Three-part and Four-part Proximal Humerus Fractures

Reinier W. A. Spek MD[1], Bram J. A. Schoolmeesters MD[1], Jacobien H. F. Oosterhoff MD[2,3], Job N. Doornberg MD, PhD[4], Michel P. J. van den Bekerom MD, PhD[5,6], Ruurd L. Jaarsma MD, PhD, FRACS[1], Denise Eygendaal MD, PhD[7], Frank IJpma MD, PhD[8], and the Traumaplatform 3D Consortium[a]

Received: 13 April 2021 / Accepted: 9 July 2021 / Published online: 23 August 2021

Abstract

Background Reliably recognizing the overall pattern and specific characteristics of proximal humerus fractures may aid in surgical decision-making. With conventional onscreen imaging modalities, there is considerable and undesired interobserver variability, even when observers receive training in the application of the classification systems used. It is unclear whether three-dimensional (3D) models, which now can be fabricated with desktop printers at relatively little cost, can decrease interobserver variability in fracture classification.

Questions/purposes Do 3D-printed handheld models of proximal humerus fractures improve agreement among residents and attending surgeons regarding (1) specific fracture characteristics and (2) patterns according to the Neer and Hertel classification systems?

Methods Plain radiographs, as well as two-dimensional (2D) and 3D CT images, were collected from 20 patients (aged 18 years or older) who sustained a three-part or four-part proximal humerus fracture treated at a Level I trauma center between 2015 and 2019. The included images were chosen to comprise images from patients whose fractures were considered as difficult-to-classify, displaced fractures. Consequently, the images were assessed for eight fracture characteristics and categorized according to the Neer and Hertel classifications by four orthopaedic residents and four attending orthopaedic surgeons during two separate sessions. In the first session, the assessment was performed with conventional onscreen imaging (radiographs and 2D and 3D CT images). In the second session, 3D-printed handheld models were used for assessment, while onscreen imaging was also available. Although proximal humerus classifications such as the Neer classification have, in the past, been shown to have low interobserver reliability, we theorized that by receiving direct tactile and visual feedback from 3D-printed handheld fracture models, clinicians would be able to recognize the complex 3D aspects of classification systems reliably. Interobserver agreement was determined with the multirater Fleiss kappa and scored according to the categorical rating by Landis and Koch. To determine whether there was a difference between the two sessions, we calculated the delta (difference in the) kappa value with 95% confidence intervals and a two-tailed p value. Post hoc power analysis revealed that with the current sample size, a delta kappa value of 0.40 could be detected with 80% power at alpha = 0.05.

Results Using 3D-printed models in addition to conventional imaging did not improve interobserver agreement of the following fracture characteristics: more than 2 mm medial hinge displacement, more than 8 mm metaphyseal extension, surgical neck fracture, anatomic neck fracture, displacement of the humeral head, more than 10 mm lesser tuberosity displacement, and more than 10 mm greater tuberosity displacement. Agreement regarding the presence of a humeral head–splitting fracture was improved but only to a level that was insufficient for clinical or scientific use (fair to substantial, delta kappa = 0.33 [95% CI 0.02 to 0.64]). Assessing 3D-printed handheld models adjunct to onscreen conventional imaging did not improve the interobserver agreement for pattern recognition according to Neer (delta kappa = 0.02 [95% CI -0.11 to 0.07]) and Hertel (delta kappa = 0.01 [95% CI -0.11 to 0.08]). There were no differences between residents and attending surgeons in terms of whether 3D models helped them classify the

Wolters Kluwer

fractures, but there were few differences to identify fracture characteristics. However, none of the identified differences improved to almost perfect agreement (kappa value above 0.80), so even those few differences are unlikely to be clinically useful.

Conclusion Using 3D-printed handheld fracture models in addition to conventional onscreen imaging of three-part and four-part proximal humerus fractures does not improve agreement among residents and attending surgeons on specific fracture characteristics and patterns. Therefore, we do not recommend that clinicians expend the time and costs needed to create these models if the goal is to classify or describe patients' fracture characteristics or pattern, since doing so is unlikely to improve clinicians' abilities to select treatment or estimate prognosis.

Level of Evidence Level III, diagnostic study.

Introduction

Recognizing the overall pattern and specific characteristics of proximal humerus fractures may aid in decision-making and determining prognosis. However, there is considerable and undesired interobserver variability, even when observers receive training in the application of the classification systems used [6]. Because the relationship between fracture lines and displacement can be difficult to assess on plain radiographs [16], two-dimensional (2D) and three-dimensional (3D) CT images are part of the routine diagnostic workup in many institutions. 2D and 3D CT images result in better intersurgeon reliability than radiographs and are particularly valuable for assessing more severe fracture configurations (such as head-splitting fractures and three-part and four-part fractures) [4, 10]. Despite the improvements seen with the use of 2D and 3D CT onscreen imaging, overall agreement on fracture patterns between attending surgeons remains low (slight-to-fair concordance) [4]. Another contentious issue is the value of 3D CT images for attending surgeons with different levels of experience; although one study concluded that residents benefit the most from using 3D CT images [2], other studies found improvement among specialists only [6, 10].

Printing of 3D models for diagnostic assessment and surgical planning of fractures is now widely available using freely available software and relatively inexpensive

[a]Members of the Traumaplatform 3D Consortium are listed in an Appendix at the end of this article.

One author (RWAS) has received payments, during the study period, in an amount of less than USD 10,000 from the Michael van Vloten Foundation (Rotterdam, the Netherlands).
One author (BJAS) received a payment of less than USD 10,000 from the Traumaplatform Foundation and Anna Foundation.
Three authors (RWAS, BJAS, JHFO) received a payment of between USD 10,000 and USD 100,000 from Prins Bernhard Cultuurfonds.
One author (RLJ) received personal fees from Smith & Nephew and DePuy Synthes in the form of paid lectures.
One author (DE) received an institutional research grant from Zimmer Biomet and Stryker and received educational research support from Lima.
All ICMJE Conflict of Interest Forms for authors and *Clinical Orthopaedics and Related Research*® editors and board members are on file with the publication and can be viewed on request.
Ethical approval was obtained from Flinders Medical Centre, Adelaide, Australia (reference number 50.19).
This work was performed at Amphia Hospital, Breda, the Netherlands, and Flinders Medical Centre, Adelaide, Australia.

[1]Department of Orthopaedic Surgery, Flinders University and Flinders Medical Centre, Adelaide, Australia

[2]Department of Orthopaedic Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

[3]Amsterdam University Medical Center, University of Amsterdam, Department of Orthopaedic Surgery, Amsterdam Movement Sciences, Amsterdam, the Netherlands

[4]Department of Orthopaedic Surgery, University of Groningen and University Medical Centre Groningen, Groningen, the Netherlands

[5]Shoulder and Elbow Expertise Centre, Department of Orthopaedic Surgery, OLVG, Amsterdam, the Netherlands

[6]Department of Human Movement Sciences, Faculty of Behavioral and Movement Sciences, Vrije Universiteit Amsterdam, Amsterdam Movement Sciences, Amsterdam, the Netherlands

[7]Department of Orthopaedic Surgery, Amsterdam University Medical Centre, Amphia Hospital, Breda, the Netherlands

[8]Department of Trauma Surgery, University of Groningen and University Medical Centre Groningen, Groningen, the Netherlands

R. W. A. Spek ✉, Department of Orthopaedic Surgery, Flinders Medical Centre, Flinders Dr, Bedford Park, Adelaide, SA 5042, Australia, Email: reinierspek@gmail.com

desktop 3D printers, without the need to rely on commercial vendors [19]. In distal humerus fractures, 3D-printed models have been demonstrated to improve intersurgeon agreement in determining fracture characteristics [5]. However, the clinical value of 3D printing for diagnostic workup of proximal humerus fractures, as well as the potential value in aiding residents to recognize patterns, has yet to be determined. To date, one study found that agreement improved regarding the choice of treatment (nonoperative versus osteosynthesis versus arthroplasty) when proximal humerus fractures were assessed with 3D-printed models [8]. Nonetheless, two studies showed that 3D-printed models improved agreement for the Neer and AO classification systems among both residents and attending surgeons, but did not reveal a difference between both groups [3, 9]. Although they conducted valuable work, they did not account for characterization and other fracture classification systems. Therefore, it remains unclear whether 3D-printed models can decrease interobserver variability in fracture assessment.

To fill this knowledge gap, we asked: Do 3D-printed handheld models of proximal humerus fractures improve agreement among residents and attending surgeons regarding (1) specific fracture characteristics and (2) patterns according to the Neer and Hertel classification systems?

Patients and Methods

Setting and Study Design

This diagnostic study was performed between August 2019 and June 2020 in a Level I trauma center in Australia and a Level II trauma center in the Netherlands. During this period, four orthopaedic residents and four attending orthopaedic surgeons (DE, three Traumaplatform 3D Consortium members) assessed 20 proximal humerus fractures for eight specific fracture characteristics and the full Neer [21] and Hertel [12] fracture patterns during two separate observation sessions with a minimum interval of 1 month between reads. As all participants were involved in the treatment of hundreds of trauma patients monthly, it was assumed that a 1-month interval would be sufficient to minimize information bias. The Neer classification categorizes proximal humerus fractures into four groups (minimally displaced, two-part, three-part, and four-part fractures) while distinguishing four anatomic segments (the shaft, articular segment, lesser tuberosity, and greater tuberosity). The segments are considered as a separate part if they are displaced more than 1 centimeter or angulated more than 45°. If not, the fracture part is considered minimally displaced. The classification also accounts for the presence of dislocation and head-splitting fractures. Altogether, 16 different categories can be chosen [22]. The Hertel classification consists of 12 different fracture patterns, which are determined by identifying the fracture planes between the greater tuberosity, humeral head, lesser tuberosity, and the shaft. Unlike the Neer classification, this system does not consider displacement or angulation between any of the segments. This classification is illustrated by LEGO bricks, and can be found in the original study by Hertel et al. [13]. Despite the poor intersurgeon agreement of the Neer classification (kappa = 0.07; 18 observers; used modality = 3D CT reconstruction images) [10] and the relatively low agreement of the Hertel classification (kappa = 0.44; four observers; used modality = rapid sequence prototype models) [18], this study incorporated both fracture patterns in the assessment.

Both classification systems have limited value for clinical decision-making, but they are still widely used to report outcomes of proximal humerus fractures in conjunction with specific fracture characteristics. For this reason, we wanted to establish how reliably these injuries could be assessed: If clinicians cannot agree on fracture characteristics and classification, it will be challenging to study results of proximal humerus fractures. The 3D-printed fracture models were designed to be held in the hand and freely rotated in space in every direction. We theorized that observers could move one step closer to reality by handling the models, allowing them to better determine angulation, displacement, and recognize the anatomic parts (such as the lesser tuberosity). Although proximal humerus classifications such as the Neer classification have, in the past, been shown to have low interobserver reliability, and in particular, the Neer classification is a complex classification system that requires 3D understanding of the fracture morphology, we wondered whether 3D-printed models could decrease its high interobserver variability.

In the first session, assessment was completed with conventional imaging, which comprised standard trauma radiographs (AP and Y-view) and 2D and 3D CT images. During the second session, the same proximal humerus fractures were evaluated, but now a 3D-printed handheld model was used in adjunct to conventional imaging (Fig. 1). Conventional imaging was presented in RadiAnt DICOM viewer (Medixant, version 2020.1). With this software, participants could toggle through the radiographs, scroll through the various 2D CT slices, and rotate the 3D CT reconstructions over the x- and y-axes. Tools to perform measurements, and the option to adjust contrast and brightness, were also available. Participants were not allowed to discuss cases; in both sessions they completed the assessment on their own.

Study Patients

We considered patients potentially eligible to have their images included if they were aged 18 years or older,

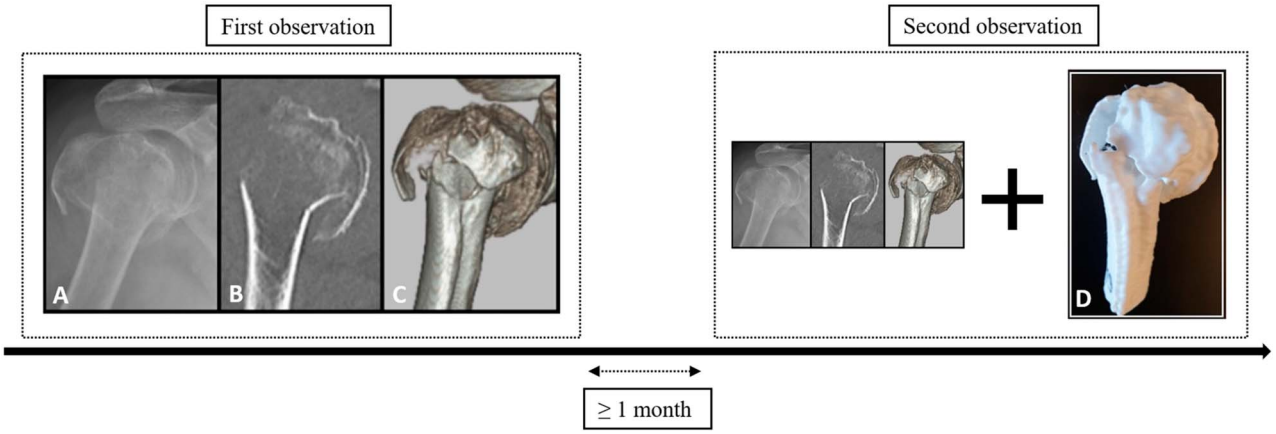

Fig. 1 During the first observation, proximal humerus fractures were assessed with conventional onscreen imaging. During the second observation, 3D-printed models were added. The image labeled with the letter A represents the trauma radiograph, B the 2D CT image (coronal plane), C the 3D CT image (anterolateral aspect), and D the 3D-printed handheld model. A color image accompanies the online version of this article.

sustained a three-part or four-part proximal humerus fracture between 2015 and 2019 that was treated at the Level I trauma center, and if they received a series of plain radiographs (AP and Y-view) and a 2D CT scan with 3D CT reconstruction images. All images, as well as the 3D CT reconstructions, were obtained as part of routine patient care and retrospectively collected from the medical imaging system Carestream Vue PACS (Carestream Health). In our Level I trauma center, it was standard practice to perform a CT scan in patients with a displaced three- or four-part proximal humerus fracture. All 3D-printed fracture models were fabricated specifically for this study. Images were collected by the second author (BJAS) who reviewed all CT shoulder scans between January 1, 2015 and July 1, 2019. Within this period there were 77 three- and four-part proximal humerus fractures. Of those fractures, the second author (BJAS) included conventional imaging from 20 patients who were considered as having especially difficult-to-classify fractures. Availability of 2D CT image data in DICOM format was a prerequisite to create 3D-printed handheld models; thus, patients without CT images or those with poor-quality CT images were excluded. For each patient, trauma radiographs and CT images were downloaded and saved as a DICOM file and subsequently anonymized with a DICOM Cleaner (PixelMed Publishing, LLC). No attending surgeons who were part of the original care of these patients were involved in the study.

Description of 3D Printing

All CT images were uploaded into a 3D slicer (The Slicer Community, version 4.10.2) for preprocessing of the 3D-printed model. To develop these skills, we followed online tutorials on the 3D slicer website [1]. Because CT images included the entire shoulder complex and part of the thorax, the proximal part of the humerus had to be cropped and contoured. This was done with a volume rendering tool by indicating the region of interest. After this, the shoulder was further segmented using thresholding. Thresholding is a semiautomatic segmentation process that selects areas based on signal intensity. The threshold used in this study to select the proximal humerus while minimizing adjacent tissues or structures was between 250 pixels and 300 pixels for the lowest volume intensity and 2002 pixels for the highest volume intensity. All surrounding bone structures were removed with the island feature. Successively, the 3D surface model was built and exported to Ultimaker Cura (version 4.6, Ultimaker B.V.) as an OBJ file (file format for 3D images containing all necessary object data and coordinates). In this software, the models were sliced and subsequently printed with a standard nozzle (diameter of 0.4 mm) on a 1:1 scale with the layer height at 0.15 mm, infill density at 20%, and printing speed at 100 mm per second. All models were printed with support material that was manually removed after the printing was finished. The prints were made with an Ultimaker 2D + 3D printer (Ultimaker BV). The costs of 3D printing depend on the preprocessing time, type of printer, and printing material [20]. Preprocessing of the models required 45 minutes, the actual printing process required approximately 6 to 8 hours, and removal of support material required less than 15 minutes. In this study, we used a printer valued at USD 2650 with polylactic acid as the printing material. Polylactic acid costs approximately USD 30 per kg and labor of a resident at the start of his/her training approximately USD 26.75 per hour; thus, considering an average of 35 g needed per 3D-

printed model, the cost of one model was USD 160 (printer = USD 132.50 [2650 ÷ 20], material = USD 1.05 [0.035 x 30], labor = USD 26.75).

Variables and Outcome Measures

The primary outcome was interobserver reliability, and observers with different levels of experience were included in this study to represent a group of clinicians working within an orthopaedic department. The participants were two orthopaedic residents who just started their training, two residents who were halfway through their training, three attending orthopaedic surgeons: two with 11 to 20 years of experience, one who was within 5 years of finishing orthopaedic training (these residents and attending orthopaedic surgeons were members of the Traumaplatform 3D Consortium), and one attending orthopaedic surgeon with fellowship training in upper extremity surgery (DE, > 21 years of experience). All participants assessed the presence or absence of the following fracture characteristics on two occasions with an interval of at least 1 month: humeral head split, more than 2 mm medial hinge displacement, more than 8 mm metaphyseal extension, surgical neck fracture, anatomic neck fracture, displacement of the humeral head (varus, valgus, or no displacement), more than 10 mm lesser tuberosity displacement, and more than 10 mm greater tuberosity displacement. Observers also classified the fractures according to the full Neer classification (16 options) and the Hertel binary LEGO description system (12 options). Answers to each question were provided on questionnaires and captured via REDCap (Vanderbilt University Medical Center) [11, 12]. Participants were able to spend as much time on the assessment as they wished. Before each session, every participant was trained by two study authors (RWAS, BJAS) using a sheet of paper with figures depicting all respective fracture classifications. Observers were allowed to keep these sheets during the assessments.

Ethical Approval

This study was approved by the institutional review board at Flinders Medical Centre, Adelaide, Australia (reference number 50.19).

Statistical Analysis

The statistical analysis was performed with Stata Statistical Software (Release 16, StataCorp LLC). The interobserver variability was determined with a multirater Fleiss kappa using bootstrapping with 1000 iterations and scored according to the Landis and Koch rating with the following categories: poor (kappa < 0.00), slight (kappa 0.00-0.20), fair (kappa 0.21-0.40), moderate (kappa 0.41-0.60), substantial (kappa 0.61-0.80), and almost perfect (kappa 0.81-1.00) [17]. If values were missing, all ratings within a participant were excluded. The multirater Fleiss kappa values are provided with corresponding 95% confidence intervals. To determine whether there was a difference between the sessions, delta (difference in the) kappa was calculated with a 95% CI and a two-tailed p value. A p value less than 0.05 was considered statistically significant. A post hoc power analysis was conducted in PASS (version 21.0.2, NCSS LLC) by comparison of two independent proportions. Although this test could not control for the number of observers, it revealed that with 20 images, a delta (difference in the) kappa of 0.40 could be established between the group with and without the 3D-printed handheld model at 80% power with alpha = 0.05.

We note that eight fracture-assessment questions were not completed. As the multirater Fleiss kappa analysis cannot handle missing data, these 8 of 400 fracture assessments were excluded through listwise deletion (20 patients assessed with conventional imaging, 20 patients with 3D-printed models, eight specific fractures characteristics, and two fracture patterns; [20 + 20] * [8 +2] = 400). The missing values were only present among residents.

Finally, a kappa value less than 0 indicates poor agreement. If the groups with and without the 3D-printed handheld model are compared, a delta kappa value less than 0 indicates that agreement for conventional imaging with 3D models is lower than for conventional imaging only. If attending surgeons are compared with residents, it means that agreement for residents is lower than for attending surgeons.

Results

Agreement on Fracture Characteristics and Classification

Among the eight observers (four orthopaedic residents and four attending surgeons), assessment by 3D-printed handheld models together with onscreen imaging did not improve agreement regarding the following fracture characteristics: more than 2 mm of medial hinge displacement, more than 8 mm of metaphyseal extension, surgical neck fracture, anatomic neck fracture, displacement of the humeral head, more than 10 mm of lesser tuberosity displacement, and more than 10 mm of greater tuberosity displacement. Interobserver agreement for the presence of a humeral head–splitting fracture improved to a level that was still inadequate for clinical use (fair to substantial, delta kappa = 0.33 [95% CI 0.02 to 0.64]). The

Table 1. Agreement for conventional imaging and 3D printed models among all eight observers

	Conventional			Conventional + 3D					
Parameter	**Kappa**		**Agreement**	**Kappa**		**Agreement**	**Δ kappa**		**p value**
Characteristic									
Humeral head split	0.39	(0.17-0.60)	Fair	0.72	(0.50-0.94)	Substantial	0.33	(0.02-0.64)	0.04
Medial hinge displacement > 2 mm	0.19	(0.00-0.38)	Slight	0.35	(0.06-0.64)	Fair	0.16	(-0.19 to 0.51)	0.36
Metaphyseal extension > 8 mm	0.14	(0.00-0.31)	Slight	0.28	(0.12-0.44)	Fair	0.14	(-0.09 to 0.36)	0.24
Surgical neck fracture	0.10	(0.00-0.26)	Slight	0.27	(0.05-0.50)	Fair	0.17	(-0.10 to 0.45)	0.21
Anatomic neck fracture	0.16	(0.01-0.31)	Slight	0.29	(0.10-0.47)	Fair	0.13	(-0.11 to 0.37)	0.29
Displacement of humeral head	0.44	(0.22-0.65)	Moderate	0.35	(0.16-0.53)	Fair	0.09	(-0.37 to 0.19)	0.55
LT displacement > 10 mm	0.03	(-0.05 to 0.12)	Slight	0.16	(0.02-0.30)	Slight	0.13	(-0.04 to 0.29)	0.13
GT displacement >10 mm	0.13	(0.01-0.25)	Slight	0.16	(0.00-0.36)	Slight	0.03	(-0.21 to 0.26)	0.81
Fracture pattern									
Neer classification	0.13	(0.06-0.20)	Slight	0.11	(0.05-0.17)	Slight	0.02	(-0.11 to 0.07)	0.67
Hertel classification	0.14	(0.06-0.22)	Slight	0.13	(0.07-0.19)	Slight	0.01	(-0.11 to 0.08)	0.81

A kappa value less than 0 indicates poor agreement; a Δ kappa value less than 0 indicates that agreement for conventional imaging with 3D models is lower than conventional imaging only; LT = lesser tuberosity; GT = greater tuberosity.

interobserver agreement for the Neer fracture patterns using conventional imaging was 0.13 (95% CI 0.06 to 0.20) and did not improve when assessed with 3D-printed models (delta kappa = 0.02 [95% CI -0.11 to 0.07]). Similarly, the agreement on the Hertel fracture patterns using conventional imaging was 0.14 (95% CI 0.06 to 0.22), and additional 3D-printed models did not result in improvement (delta kappa = 0.01 [95% CI -0.11 to 0.08]) (Table 1).

Agreement Among Residents and Attending Surgeons

Among residents, additional 3D-printed handheld models did not improve agreement regarding fracture characteristics and patterns (Table 2). Among attending surgeons, only agreement on lesser tuberosity displacement more than 10 mm improved from poor to slight (delta kappa = 0.22 [95% CI 0.01 to 0.42]), which was still insufficient for clinical use. Thus, adding 3D-printed handheld models to the diagnostic process likewise did not improve concurrence among attending surgeons (Table 3). There were no differences between residents and attending surgeons in terms of whether 3D models helped them to classify the fractures, and there were few differences in terms of whether the 3D models helped them to identify fracture characteristics. However, none of the identified differences improved to almost perfect agreement (kappa value above 0.80), so we do not see even those few differences as likely to be clinically useful (Table 4).

Discussion

Recognizing the overall pattern and specific characteristics of proximal humerus fractures may aid in decision-making and determining prognosis. However, there is considerable and undesired interobserver variability, even when observers receive training in the application of the classification systems used. Both the Neer and Hertel classifications are routinely reported in research studies, so to enhance our knowledge, we wanted to evaluate how reliably these injuries can be assessed with the assistance of 3D models. We therefore sought to determine whether cutting-edge technology (3D-printed fracture models), which now can be fabricated with desktop printers at relatively little cost, could deliver its promise and reduce the great undesired interobserver variability in fracture classification and characterization. If clinicians cannot agree, it will be challenging to evaluate results of proximal humeral fractures based on these classification schemes. In summary, we found that using 3D-printed handheld models

Table 2. Agreement for conventional imaging and 3D printed models among four residents

	Conventional			Conventional + 3D				
Parameter	**Kappa**		**Agreement**	**Kappa**		**Agreement**	**Δ kappa**	**p value**
Characteristic								
Humeral head split	0.48	(0.21-0.74)	Moderate	0.66	(0.37-0.95)	Substantial	0.18 (-0.21 to 0.58)	0.18
Medial hinge displacement > 2 mm	0.02	(-0.15 to 0.20)	Slight	0.07	(-0.18 to 0.33)	Slight	0.05 (-0.25 to 0.35)	0.74
Metaphyseal extension > 8 mm	0.27	(0.02-0.52)	Fair	0.28	(0.08-0.48)	Fair	0.01 (-0.32 to 0.33)	0.48
Surgical neck fracture	0.02	(-0.20 to 0.24)	Slight	0.13	(-0.15 to 0.40)	Slight	0.11 (-0.25 to 0.46)	0.28
Anatomic neck fracture	0.17	(-0.03 to 0.38)	Slight	0.28	(0.00-0.55)	Fair	0.10 (-0.24 to 0.45)	0.57
Displacement of the humeral head	0.29	(0.07 to 0.52)	Fair	0.1	(-0.07 to 0.26)	Slight	0.19 (-0.79 to 0.40)	0.52
LT displacement >10 mm	0.18	(-0.06 - 0.41)	Slight	0.17	(-0.06 to 0.40)	Slight	0.01 (-0.33 to 0.32)	0.97
GT displacement >10 mm	0.16	(-0.06 to 0.38)	Slight	0.13	(-0.13 to 0.39)	Slight	0.03 (-0.37 to 0.31)	0.43
Fracture pattern								
Neer classification	0.14	(0.03-0.25)	Slight	0.04	(-0.05 to 0.13)	Slight	0.10 (-0.25 to 0.04)	0.16
Hertel classification	0.14	(0.00-0.27)	Slight	0.13	(0.03-0.23)	Slight	0.01 (-0.17 to 0.15)	0.46

A kappa value less than 0 indicates poor agreement; a Δ kappa value less than 0 indicates that agreement for conventional imaging with 3D models is lower than for conventional imaging only; LT = lesser tuberosity; GT = greater tuberosity.

with conventional imaging to assess three-part and four-part proximal humerus fractures did not improve agreement for fracture characteristics to a level that was adequate for clinical or scientific use. No improvement in agreement on fracture pattern recognition according to Neer and Hertel was established by using 3D-printed models together with onscreen conventional imaging. Residents did not seem to benefit more from 3D-printed handheld models than attending surgeons did. Hence, we do not recommend using these models in clinical practice if the goal is to improve classification reliability or to describe patients' fracture patterns or characteristics.

Limitations

Several limitations must be considered when interpreting our findings. First, we included only eight observers, which resulted in wide 95% CIs. However, as 3D-printed models are still relatively expensive and time-consuming, they must show strong value to be incorporated in clinical practice. Therefore, even a small study like ours should have been able to demonstrate the added value of 3D-printed models. For this reason, it was powered to detect profound differences between conventional imaging and 3D-printed models and not to detect subtle changes (post hoc power analysis revealed that with 20 images, a delta kappa of 0.40 could be detected). Second, results should be inferred considering differences in experience among residents and attending surgeons. One may argue that residents and attending surgeons do not have the same knowledge level compared with an upper extremity expert. This could have decreased the agreement; however, our goal was to include an observer panel that would represent typical orthopaedic practice in public hospitals. Third, in our observer cohort, there were missing values. To address this, proximal humerus fractures were listwise excluded from the analysis. This was 2% of the total number of proximal humerus fractures and could therefore have influenced our 95% confidence intervals. Notably, missing values only occurred among residents and mainly in the 3D group.

In addition, we did not analyze intraobserver reliability. Classifying proximal humerus fractures is challenging; so much so that even advanced technology such as the 3D models used in our study could not improve agreement. We therefore argue that the classification is the main flaw and must be revised. We also note that diagnostic parameters, such as accuracy, were not included in this study. Ideally, this would be established intraoperatively, but because not all fractures were treated surgically, this was not feasible. A potential limitation here was that a cost analysis showing at

Table 3. Agreement for conventional imaging and 3D printed models among four attending surgeons

	Conventional			Conventional + 3D				
Parameter	**Kappa**		**Agreement**	**Kappa**		**Agreement**	**Δ kappa**	**p value**
Characteristic								
Humeral head split	0.37	(0.05-0.69)	Fair	0.75	(0.48-1.00)	Substantial	0.38 (-0.04 to 0.79)	0.08
Medial hinge displacement > 2 mm	0.38	(0.05-0.70)	Fair	0.62	(0.24-1.00)	Substantial	0.24 (-0.26 to 0.74)	0.34
Metaphyseal extension > 8 mm	0.00	(-0.18 to 0.17)	Slight	0.16	(-0.02 to 0.33)	Slight	0.16 (-0.09 to 0.40)	0.20
Surgical neck fracture	0.22	(-0.16 to 0.59)	Fair	0.54	(0.10-0.98)	Moderate	0.33 (-0.25 to 0.90)	0.27
Anatomic neck fracture	0.13	(-0.09 to 0.34)	Slight	0.42	(0.17-0.66)	Moderate	0.29 (-0.03 to 0.61)	0.08
Displacement of the humeral head	0.58	(0.36-0.81)	Moderate	0.69	(0.41-0.97)	Substantial	0.11 (-0.25 to 0.47)	0.55
LT displacement > 10 mm	-0.11	(-0.23 to 0.02)	Poor	0.11	(-0.05 to 0.28)	Slight	0.22 (0.01-0.42)	0.04
GT displacement > 10 mm	-0.01	(-0.15 to 0.13)	Poor	0.13	(-0.14 to 0.40)	Slight	0.14 (-0.16 to 0.45)	0.36
Fracture pattern								
Neer classification	0.13	(-0.01 to 0.27)	Slight	0.05	(-0.04 to 0.13)	Slight	0.08 (-0.25 to 0.08)	0.32
Hertel classification	0.09	(-0.02 to 0.20)	Slight	0.06	(-0.01 to 0.19)	Slight	0.03 (-0.18 to 0.12)	0.68

A kappa value less than 0 indicates poor agreement; a Δ kappa value less than 0 indicates that agreement for conventional imaging with 3D models is lower than conventional imaging only; LT = lesser tuberosity; GT = greater tuberosity.

what price such models would become cost-effective was not reported in this work. Something that is not effective cannot be cost effective; we therefore decided that a cost analysis should not be performed, given that effectiveness (in other words, interobserver reliability) was not established. Lastly, the Hertel fracture patterns were classified according to the original binary description system comprising 12 different categories [14] and thus without the two humeral-head split fracture types [13].

Agreement on Fracture Characteristics and Classification

This study revealed that using 3D-printed handheld models in adjunct to onscreen imaging did not improve agreement regarding fracture characteristics. Based on these results, we cannot recommend using these models in the diagnostic workup of patients with proximal humerus fractures, especially because these models require time, materials, and money to produce. Only one study reported on the use of 3D-printed models to assess the characteristics of proximal humerus fractures [7]. They retrospectively compared preoperative planning with conventional imaging, virtual planning, and 3D-printed models in patients undergoing internal fixation with locking plates and assessed clinical outcomes and the accuracy of fracture characteristics. Their results were based on intraoperative findings as the reference standard, and they therefore determined diagnostic parameters and not interobserver agreement. However, consistent with our study, they did not reveal any differences between 3D-printed models and conventional imaging.

The kappa value for fracture patterns according to the Neer classification in prior studies ranges from 0.07 to 0.14 (16 observers) [10], and from 0.39 to 0.60 (four observers) for the Hertel classification [15] with the availability of radiographs and 2D and 3D CT images. In our study, fracture pattern recognition according to the Neer and Hertel classifications had low interobserver agreement in all imaging modalities despite 3D-printed modeling (Neer: kappa = 0.11, Hertel: kappa = 0.13). One study demonstrated fair-to-moderate agreement for the simplified Neer classification (three categories: two-part, three-part, and four-part fractures) among 20 residents (kappa = 0.40) and 20 attending surgeons (kappa = 0.50) when using 3D-printed models only (without additional imaging). This supports that 3D models are not clinically useful for classifying proximal humerus fractures but the question remained unanswered if other classifications, such as the Hertel LEGO description system, or specific fracture characteristics would improve with 3D modeling [9]. Another study found moderate agreement (kappa = 0.47)

Table 4. Comparison of agreement between four residents and four attending surgeons using 3D printed models

	Residents (conventional + 3D)			Attending surgeons (conventional + 3D)				
Parameter	Kappa		Agreement	Kappa		Agreement	Δ kappa	p value
Characteristic								
Humeral head split	0.66	(0.37-0.95)	Substantial	0.75	(0.48-1.00)	Substantial	0.09 (-0.31 to 0.49)	0.65
Medial hinge displacement > 2 mm	0.07	(-0.18 to 0.33)	Slight	0.62	(0.24-1.00)	Substantial	0.54 (0.09-0.99)	0.02
Metaphyseal extension > 8 mm	0.28	(0.08-0.48)	Fair	0.16	(-0.02 to 0.33)	Slight	0.12 (-0.39 to 0.14)	0.36
Surgical neck fracture	0.13	(-0.15 to 0.40)	Slight	0.54	(0.10-0.98)	Moderate	0.42 (-0.10 to 0.94)	0.11
Anatomic neck fracture	0.28	(0.00-0.55)	Fair	0.42	(0.17-0.66)	Moderate	0.14 (-0.23 to 0.51)	0.45
Displacement of the humeral head	0.10	(-0.07 to 0.26)	Slight	0.69	(0.41-0.97)	Substantial	0.60 (0.27-0.92)	< 0.001
LT displacement > 10 mm	0.17	(-0.06 to 0.40)	Slight	0.11	(-0.05 to 0.28)	Slight	0.06 (-0.34 to 0.22)	0.68
GT displacement > 10 mm	0.13	(-0.13 to 0.39)	Slight	0.13	(-0.14 to 0.40)	Slight	0.00 (-0.38 to 0.38)	> 0.99
Fracture pattern								
Neer classification	0.04	(-0.05 to 0.13)	Slight	0.05	(-0.04 to 0.13)	Slight	0.01 (-0.12 to 0.13)	0.90
Hertel classification	0.13	(0.03-0.23)	Slight	0.06	(-0.01 to 0.19)	Slight	0.07 (-0.22 to 0.07)	0.30

A kappa value less than 0 indicates poor agreement; A Δ kappa value less than 0 indicates that agreement for residents is lower than for attending surgeons; LT = lesser tuberosity; GT = greater tuberosity.

among 14 assessors, but they also simplified the Neer classification to three categories (two-part, three-part, and four-part fractures) and assessed the 3D-printed models without additional radiographs or CT images [3]. Again, the question whether 3D fracture models would be useful for characterization and assessment of other fracture classification systems was left open. Combining these studies with our data, it seems justifiable to say that the utility of 3D models in determining fracture assessment of proximal humerus fractures is negligible. Nevertheless, 3D models may help create surgical strategies and approaches, such as guides to place K-wires and screws. They may also be valuable for educational purposes (such as teaching medical students or explaining surgical plans preoperatively), but well-designed follow-up studies are needed to identify any potential benefits.

Agreement Among Residents and Attending Surgeons

There were no important differences between residents and attending surgeons in whether 3D models helped them to classify or describe the fractures, and the few observed differences were not sufficiently large to be clinically useful (Table 4). These findings were in line with another study that did not find any differences in agreement between residents and attending surgeons [9]. It is likely that because of the complexity of three-part and four-part proximal humerus fractures, assessment is difficult and debatable for both residents and attending surgeons. It also confirms that the hallmarks of proximal humerus fractures are seen differently and subjectively by observers, and that they are difficult to categorize in any classification scheme. Considering this, we do not recommend using the currently available classification systems for supporting clinical decisions or to report on patient outcomes. Time-consuming interventions like the 3D-printed models used in this study did not overcome the shortcomings of difficult-to-use classifications; keeping those classifications as simple as possible therefore seems important.

Conclusion

Using 3D-printed handheld models with onscreen conventional imaging (radiographs and 2D and 3D CT images) to assess three-part and four-part proximal humerus fractures did not improve agreement regarding fracture characteristics and patterns. Therefore, we cannot recommend that clinicians expend the time and costs needed to create these models if the goal is to classify or describe patients' fracture characteristics. Future studies are needed to

establish the value of 3D modeling in practicing fracture fixation and templating a preoperative plan.

Acknowledgments We thank Barbara Tosun MSc, senior research fellow biostatistician at Flinders University in Adelaide, Australia, for her statistical advice on kappa calculations and Pawel Skuza PhD, a statistician at Flinders University, for his statistical advice on the sample size analysis.

Group Authorship

The members of Traumaplatform 3D Consortium include Nick Assink, Wouter J. Bekkers, Christiaan J.A. van Bergen, Ronald Boer, Lars Brouwers, Tijan Gajic, Michiel M.A. Janssen, Paul C. Jutte, Laura J. Kim, Lisette C. Langenberg, Anne M.L. Meesters, Patrick Nieboer, Sjoerd P.F.T. Nota, Karsten D. Ottink, Bertram The, Anne Veldhuizen, and Hanneke Weel.

References

1. 3D Slicer. Documentation/4.10/Training - Slicer Wiki. Available at: https://www.slicer.org/wiki/Documentation/4.10/Training#Introduction:_Slicer_4.10_Tutorials. Accessed March 24, 2021.
2. Berkes MB, Dines JS, Little MTM, et al. The impact of three-dimensional CT imaging on intraobserver and interobserver reliability of proximal humeral fracture classifications and treatment recommendations. *J Bone Joint Surg Am*. 2014;96:1281-1286.
3. Bougher H, Buttner P, Smith J, et al. Interobserver and intraobserver agreement of three-dimensionally printed models for the classification of proximal humeral fractures. *JSES Int*. 2021;5:198-204.
4. Bougher H, Nagendiram A, Banks J, Hall LM, Heal C. Imaging to improve agreement for proximal humeral fracture classification in adult patient: a systematic review of quantitative studies. *J Clin Orthop Trauma*. 2020;11:S16-S24.
5. Brouwer KM, Lindenhovius AL, Dyer GS, Zurakowski D, Mudgal CS, Ring D. Diagnostic accuracy of 2- and 3-dimensional imaging and modeling of distal humerus fractures. *J Shoulder Elbow Surg*. 2012;21:772-776.
6. Bruinsma WE, Guitton TG, Warner JJP, Ring D. Interobserver reliability of classification and characterization of proximal humeral fractures. *J Bone Joint Surg Am*. 2013;95:1600-1604.
7. Chen Y, Jia X, Qiang M, Zhang K, Chen S. Computer-assisted virtual surgical technology versus three-dimensional printing technology in preoperative planning for displaced three and four-part fractures of the proximal end of the humerus. *J Bone Joint Surg Am*. 2018;100:1960-1968.
8. Cocco LF, Aihara AY, Franciozi C, Dos Reis FB, Luzo MVM. Three-dimensional models increase the interobserver agreement for the treatment of proximal humerus fractures. *Patient Saf Surg*. 2020;14:33.
9. Cocco LF, Yazzigi JA, Kawakami EFKI, Alvachian HJF, Dos Reis FB, Luzo MVM. Inter-observer reliability of alternative diagnostic methods for proximal humerus fractures: a comparison between attending surgeons and orthopedic residents in training. *Patient Saf Surg*. 2019;13:1-13.
10. 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.
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.
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;42:377-381.
13. Hertel R, Hempfing A, Stiehler M, Leunig M. Predictors of humeral head ischemia after intracapsular fracture of the proximal humerus. *J Shoulder Elbow Surg*. 2004;13:427-433.
14. Hertel R, Mees C, Scholl E, Ballmer FT, Siebenrock K. Morphologic classification of fractures of the proximal humerus. A validated, teachable and practicable alternative. In: *Eighth International Conference on Shoulder Surgery* (ICSS). 2001:23-26.
15. Iordens GIT, Mahabier KC, Buisman FE, 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.
16. Janssen SJ, Hermanussen HH, Guitton TG, van den Bekerom MPJ, van Deurzen DFP, Ring D. Greater tuberosity fractures: does fracture assessment and treatment recommendation vary based on imaging modality? *Clin Orthop Relat Res*. 2016;474:1257-1265.
17. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159-174.
18. Majed A, Macleod I, Bull AMJ, et al. Proximal humeral fracture classification systems revisited. *J Shoulder Elbow Surg*. 2011;20:1125-1132.
19. Mitsouras D, Liacouras P, Imanzadeh A, et al. Medical 3D printing for the radiologist. *Radiographics*. 2015;35:1965-1988.
20. Morgan C, Khatri C, Hanna SA, Ashrafian H, Sarraf KM. Use of three-dimensional printing in preoperative planning in orthopaedic trauma surgery: a systematic review and meta-analysis. *World J Orthop*. 2020;11:57-67.
21. Neer CS. Displaced proximal humeral fractures. I. Classification and evaluation. *J Bone Joint Surg Am*. 1970;52:1077-89.
22. Sumrein BO, Mattila VM, Lepola V, et al. Intraobserver and interobserver reliability of recategorized Neer classification in differentiating 2-part surgical neck fractures from multifragmented proximal humeral fractures in 116 patients. *J Shoulder Elbow Surg*. 2018;27:1756-1761.