# Classification and measurement of displacement of isolated greater tuberosity fractures: Intra and interobserver reliability

Ghiath Ismayl[1] (iD), Catherine Ogbechie[2], Samuel Goundry[2], Luke Budworth[3], Ikechukwu Ejiofor[1], Hassaan Sheikh[4], Paul McCormack[5], Charlotte Tunstall[1], Mark Philipson[1] and Paul Cowling[1] (iD)

## Abstract

**Background:** Literature demonstrates variability in the amount of displacement of isolated greater tuberosity (GT) fractures and measurement techniques that orthopaedic surgeons deem warrant surgical intervention. This study aims to assess the intra and interobserver reliability for classifying and measuring the displacement amount for isolated GT fractures.

**Methods:** Eight surgeons, consisting of four shoulder specialists and four trainee surgeons, reviewed 25 plain radiographs on two separate occasions, 3 months apart. They were required to morphologically classify the GT fracture, measure the displacement distance on anteroposterior and axillary views, calculate the GT displacement ratio, and state whether they would offer surgical treatment.

**Results:** There was a lack of good reliability for the classification of the depression and avulsion fracture types. There was good intraobserver but poor interobserver consistency in classifying the split-type fractures. The measurement of the displacement distance showed good intraobserver reliability, but not as good interobserver agreement. Also, the displacement ratio calculated revealed poor consistency. We found good agreement between and within the raters for the treatment decision. No significant difference was noted when comparing the senior surgeons to the junior surgeons.

**Conclusions:** This study has revealed ongoing inconsistency in the classification and measurement of isolated GT fractures.

## Background

The greater tuberosity (GT) is the bony prominence of the proximal humerus that acts as an attachment point for three of the four rotator cuff muscles. It therefore plays an integral role in sustaining shoulder stability and movement.[1] Isolated GT fractures amount to ~20% of all proximal humerus fractures.[2]

The measurement of isolated GT fracture displacement plays an important role in the evaluation and treatment of these injuries. Depending on the amount of displacement, these injuries can be managed operatively or nonoperatively. However, there is controversy on the level of displacement that orthopaedic surgeons deem to warrant surgical intervention.

[1]Trauma and Orthopaedics Department, Leeds Teaching Hospitals NHS Trust, Leeds General Infirmary, Leeds, UK
[2]University of Leeds, Leeds, UK
[3]Leeds Institute of Rheumatic and Musculoskeletal Medicine, University of Leeds, Leeds, UK
[4]Trauma and Orthopaedics Department, Barnsley Hospital NHS Foundation Trust, Barnsley, UK
[5]Trauma and Orthopaedics Department, Airedale NHS Foundation Trust, Keighley, UK

**Corresponding author:**
Ghiath Ismayl, Trauma and Orthopaedics Department, Leeds Teaching Hospitals NHS Trust, Leeds General Infirmary, Great George Street, Leeds, LS1 3EX, UK.
Email: ghiath.ismayl@nhs.net

Figures ranging from 3 to 10 mm have been reported in the literature, but 5 mm displacement is the most commonly used threshold.[1,3]

Unfortunately, it has been shown to be difficult to measure the amount of displacement with this degree of precision and reliability on plain radiographs.[1,2,4] To address this subject, Mutch et al.[5] described a measurement method which involves calculating the GT fracture displacement ratio. Mutch et al.[6] further presented a new classification system in 2014, dividing GT fractures into one of three morphological types; avulsion, split and depression.

We aim in this study to evaluate the intra and interobserver reliability for the assessment of isolated GT fractures on plain radiographs. This includes the morphological classification, measurement of displacement distance and ratio, and whether surgical treatment is required.

## Method

### Study design

Eight surgeons working at our institution were chosen for the study. The plain radiographs of 25 patients with isolated GT fractures of the proximal humerus were retrospectively taken from the institution's medical records over the previous 10 years. The radiographs chosen were the images performed in the emergency department of the first presentation of the GT fracture and therefore were the images from which the decision regarding management was made by the treating orthopaedic team. The radiographs were conducted using a standard imaging protocol. A recent study used 25 patients to assess the intra and interobserver reliability of follow-up (FL) plain radiographs for isolated GT fractures.[7] Hence the decision was made to include 25 patients in our study.

The eight surgeons were independently provided with the 25 plain radiographs to review. The images were anonymized and presented in a random order for each rater. Assessment of the images was conducted on PACS digital imaging software within the institution utilizing high-resolution radiology screens to enhance the quality of the radiographs. Following the first review by each rater, a second review was conducted by the same raters for the same images 3 months later. No additional material or training was provided to the raters during these 3 months that may affect their second review. At the second review, the same radiographs were randomized in a different order to the original series, again to attempt to reduce bias.

A standardized guide was distributed to each rater prior to the review to ensure a uniform approach to the assessment. They were required to morphologically classify the GT fracture, measure the displacement on anteroposterior (AP) and axillary views, calculate the GT fracture displacement ratio, and state whether they would offer surgical treatment. For the fracture classification, diagrammatic illustrations and corresponding plain radiographs from Mutch's original description were provided to demonstrate the three groups of GT fractures: avulsion, split and depression fracture types.[6] For the measurement of displacement on AP and axillary views, no specific directions were supplied. This lack of guidance was deliberate, as it was noticeable that a number of published papers describing their management of GT fractures using displacement distance failed to describe the method of measurement.[1,3] The measurement was conducted using a PACS digital ruler in millimetres to 2 decimal places. However, guidance on calculating the GT fracture displacement ratio was provided and came directly from the original description of this method of measurement, along with the diagrammatical guide provided in that publication.[5] Finally, the raters were required to decide whether they would recommend operative or nonoperative treatment for the GT fracture based on radiological findings only: this was a binary question posed 'On the basis of these images, would you offer this patient surgery? Yes/No'. The patient's age and gender were visible on all the plain radiographs during the assessment, but no other patient information was provided to assist in the decision-making.

### Raters

The eight raters comprised four senior and four junior surgeons. The senior surgeons were consultants specialized in shoulder surgery with more than 5 years of experience at the consultant level. The four junior surgeons were trainees in their early postgraduate years.

### Statistical analysis

Statistical analysis was performed using RStudio software version 1.4.1564. For binary variables, in the partially nested dataset, the agreement package was first used to calculate and plot a series of agreement metrics ($\alpha$, $\gamma$, Ir2, $\kappa$, $\pi$, and S) between the raters at each time point. These were interpreted as per standard kappa ($\kappa$) coefficient conventions to evaluate the interrater variability only, with a value >0.69 generally being considered 'good'.[8]

To further evaluate binary and continuous variables for both intra and interrater reliability, the fully nested dataset was used to calculate the intraclass correlation coefficient (ICC) and associated 95% confidence intervals (CIs) for a two-way random effect, absolute agreement, and single rater/measurement model. An ICC >0.69 was deemed to constitute good reliability on the intra and interobserver analysis for 95% CI.[9]

A stratified analysis was conducted to compare the intra and interobserver reliability between the four senior and four junior surgeons. A $p$-value <0.05 represented statistical significance for a 95% CI.

The results of the analysis are reported in the text in the form of ICC and kappa (κ), with both baseline (B) and FL figures for κ. κ baseline (κB) represents the first review and κFL represents the second review conducted 3 months later.

## Results

All eight surgeons completed the assessment for the 25 plain radiographs at both time points. There was variability in the timing of when the fractures took place, the patient's age, gender and degree of GT fracture displacement in the cohort of plain radiographs selected for the study, as well as the treatment performed. Of the 25 plain radiographs, 12 were male patients and 13 were female patients. The mean age was 64 years old (range 30–89 years). The most common mechanism of injury was a fall from standing height, followed by high-energy trauma. All results are summarized in Tables 1 and 2.

### Classification

There was good intraobserver reliability in classifying the split-type GT fractures (intra-ICC 0.72), but there was poor interobserver reliability (inter-ICC 0.44, κB 0.50 and κFL 0.36).

There was weak consistency in the intra and interobserver classification of the GT fracture for the depression type (inter-ICC 0.22, κB 0.23, κFL 0.19 and intra-ICC 0.45) and interobserver classification for the avulsion type (inter-ICC 0.37, κB 0.37 and κFL 0.35), but moderate reliability for the intraobserver classification of the avulsion type (intra-ICC 0.65).

**Table 1.** Summary of statistical analysis (*k*) for the binary outcomes.

| Outcome | Time | *k* | L95% | U95% |
| --- | --- | --- | --- | --- |
| Classification: split | B | 0.50 | 0.35 | 0.65 |
| | FL | 0.36 | 0.23 | 0.49 |
| Classification: depression | B | 0.23 | −0.02 | 0.43 |
| | FL | 0.19 | 0.00 | 0.32 |
| Classification: avulsion | B | 0.37 | 0.22 | 0.50 |
| | FL | 0.35 | 0.20 | 0.48 |
| Surgery recommended | B | 0.82 | 0.70 | 0.92 |
| | FL | 0.73 | 0.54 | 0.87 |

k: kappa; L95%: lower limit 95% confidence interval; U95%: upper limit 95% confidence interval; B: baseline, represents the first review; FL: follow-up, represents the second review conducted 3 months later.

*Measurement of displacement.* The measurement of the GT fracture displacement showed good intraobserver consistency on both AP and axillary views (intra-ICC 0.81 and intra-ICC 0.70, respectively), but not as good interobserver consistency (inter-ICC 0.69 and inter-ICC 0.60, respectively).

*Fracture displacement ratio.* GT fracture displacement ratio calculated revealed poor reliability (intra-ICC 0.24 and inter-ICC 0.18).

*Decision to operate.* The decision to operate on the GT fracture based on the radiograph alone demonstrated good reliability between and within the raters (inter-ICC 0.76, κB 0.82, κFL 0.73 and intra-ICC 0.77).

*Comparison of senior and junior surgeons.* The stratified analysis comparing the senior surgeons to the junior surgeons yielded results with a very wide 95% CI across all variables. The results of this analysis are summarized in Tables 3 and 4.

There was a slightly better intra and interobserver consistency for the senior surgeons in the classification of the GT fracture and decision to operate. The junior surgeons had higher intraobserver reliability but lower interobserver reliability for the measurement of fracture displacement on AP and axillary plain radiograph views. The GT displacement ratio calculated had better intra and inter-rater consistency with the junior surgeons.

However, no statistical significance was noted on the κ and ICC coefficients for a 95% CI across all parameters between the two groups.

## Discussion

Our results demonstrated ongoing inconsistency in the classification and measurement of isolated GT fractures. However, we did find good agreement on whether to offer surgery or nonoperative management.

Several classification systems have been developed for proximal humerus fractures and GT fractures. Neer's classification was one of the earliest systems and was based on the number of displaced fragments in the proximal humerus. It divided the proximal humerus into four segments, consisting of the GT, lesser tuberosity, articular surface, and humeral diaphysis. A segment with >1 cm separation or 45° angulation was considered a displaced part.[10] AO later categorized GT fractures into displaced, nondisplaced, and those associated with shoulder dislocation. The fracture displacement was defined as a separation of 5 mm or more.[11] However, both Neer and AO classification systems have achieved less than ideal interobserver reliability on plain radiographs.[12] The addition of computerized tomography (CT) scan and 3D CT reconstruction did not significantly improve the interobserver reliability.[13–15]

**Table 2.** Summary of statistical analysis (ICC) for the binary and continuous outcomes.

| Outcome | Intra-ICC | L95% | U95% | Inter-ICC | L95% | U95% |
|---|---|---|---|---|---|---|
| **Binary** | | | | | | |
| Classification: split | 0.72 | 0.59 | 0.82 | 0.44 | 0.30 | 0.56 |
| Classification: depression | 0.45 | 0.04 | 0.72 | 0.22 | 0.00 | 0.37 |
| Classification: avulsion | 0.65 | 0.54 | 0.75 | 0.37 | 0.23 | 0.49 |
| Surgery recommended | 0.77 | 0.64 | 0.86 | 0.76 | 0.62 | 0.86 |
| **Continuous** | | | | | | |
| Measurement AP view | 0.81 | 0.58 | 0.91 | 0.69 | 0.46 | 0.76 |
| Measurement axillary view | 0.70 | 0.59 | 0.77 | 0.60 | 0.46 | 0.70 |
| GT displacement ratio | 0.24 | 0.12 | 0.40 | 0.18 | 0.07 | 0.28 |

ICC: intraclass correlation coefficient; L95%: lower limit 95% confidence interval; U95%: upper limit 95% confidence interval; AP: anteroposterior; GT: greater tuberosity.

**Table 3.** Summary of statistical analysis (k) for the binary outcomes stratified by surgeons' seniority.

| Outcome | Time | Junior doctors | | | Senior doctors | | | $k_1-k_2$ |
|---|---|---|---|---|---|---|---|---|
| | | k | L95% | U95% | k | L95% | U95% | |
| Classification: split | B | 0.348 | 0.157 | 0.546 | 0.683 | 0.46 | 0.86 | −0.335 |
| | FL | 0.076 | −0.087 | 0.221 | 0.551 | 0.35 | 0.75 | −0.475 |
| Classification: depression | B | 0.35 | −0.048 | 0.706 | 0 | – | – | 0.35 |
| | FL | 0.227 | 0 | 0.49 | 0 | – | – | 0.227 |
| Classification: avulsion | B | 0.285 | 0.113 | 0.465 | 0.49 | 0.25 | 0.69 | −0.205 |
| | FL | 0.014 | −0.139 | 0.173 | 0.543 | 0.34 | 0.72 | −0.529 |
| Surgery recommended | B | 0.711 | 0.506 | 0.879 | 0.943 | 0.79 | 1 | −0.232 |
| | FL | 0.695 | 0.459 | 0.878 | 0.763 | 0.54 | 0.94 | −0.068 |

k: kappa; L95%: lower limit 95% confidence interval; U95%: upper limit 95% confidence interval; B: baseline, represents the first review; FL: follow-up, represents the second review conducted 3 months later.

Whereas the use of stereo-visualisation of 3D volume-rendering CT datasets was found to increase the intra and interobserver agreement for both classifications.[16]

Mutch et al. established a morphology-based classification system in 2014, dividing GT fractures into one of three types; avulsion, split and depression. Avulsion fractures result from the pull effect of the rotator cuff muscles on the GT. Vertical shear causes split-type fractures, and fragment impaction produces depression-type fractures.[6] In contrast to Neer and AO, the Mutch classification has shown good intra and interobserver reliability for isolated GT fractures on plain radiographs and CT imaging.[17] Our study demonstrated good intraobserver agreement for classifying the split-type fractures and moderate intraobserver agreement for classifying the avulsion type. However, the interobserver reliability for all three fracture types and intraobserver reliability for the depression type revealed weak consistency among the raters.

Accurate measurement of the fracture displacement is paramount for the management of isolated GT fractures. As little as 3 to 5 mm of displacement may negatively affect rotator cuff biomechanics and result in subacromial impingement.[3] However, it has proven to be a challenge to measure the amount of displacement precisely and

**Table 4.** Summary of statistical analysis (ICC) for the binary and continuous outcomes stratified by surgeons' seniority.

| Outcome | Junior surgeons | | | | | | Senior surgeons | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Intra-ICC | L95% | U95% | Inter-ICC | L95% | U95% | Intra-ICC | L95% | U95% | Inter-ICC | L95% | U95% | Intra1–Intra2 | Inter1–Inter2 |
| **Binary** | | | | | | | | | | | | | | |
| Classification: split | 0.655 | 0.444 | 0.832 | 0.233 | 0.081 | 0.382 | 0.767 | 0.63 | 0.876 | 0.633 | 0.47 | 0.78 | −0.112 | −0.4 |
| Classification: depression | 0.328 | 0.023 | 0.612 | 0.328 | 0 | 0.612 | 1 | 1 | 1 | 0 | 0 | 0 | −0.672 | 0.328 |
| Classification: avulsion | 0.58 | 0.418 | 0.731 | 0.204 | 0.069 | 0.339 | 0.718 | 0.57 | 0.84 | 0.532 | 0.34 | 0.695 | −0.138 | −0.328 |
| Surgery recommended | 0.704 | 0.541 | 0.822 | 0.704 | 0.541 | 0.822 | 0.835 | 0.69 | 0.94 | 0.832 | 0.68 | 0.939 | −0.131 | −0.128 |
| **Continuous** | | | | | | | | | | | | | | |
| Measurement AP view | 0.829 | 0.642 | 0.918 | 0.5 | 0.424 | 0.698 | 0.803 | 0.55 | 0.904 | 0.791 | 0.49 | 0.892 | 0.026 | −0.291 |
| Measurement axillary view | 0.731 | 0.596 | 0.815 | 0.532 | 0.385 | 0.649 | 0.679 | 0.54 | 0.782 | 0.617 | 0.46 | 0.718 | 0.052 | −0.085 |
| GT displacement ratio | 0.488 | 0.242 | 0.703 | 0.218 | 0.048 | 0.397 | 0.063 | 0 | 0.204 | 0.063 | 0 | 0.179 | 0.425 | 0.155 |

ICC: intraclass correlation coefficient; L95%: lower limit 95% confidence interval; U95%: upper limit 95% confidence limit; AP: anteroposterior; GT: greater tuberosity.

reliably on plain radiographs.[1,4] Van Wier et al.[7] noted poor interobserver agreement on measuring the GT fracture displacement on plain radiographs, but there was less interobserver variation when subsequent FL radiographic images were used, though the overall reliability remained less than ideal. Likewise, our study has found weak interobserver reliability for measuring the displacement distance on AP and axillary plain radiograph views, but there was good intraobserver consistency. This is likely due to the variability in the measurement method, landmarks used and interpretation of the images between the raters. The GT fracture fragment and footprint were sometimes difficult to visualize, complicating the measurement technique, particularly in cases where the GT fragment displaced posteriorly, the shoulder joint had an abnormal rotation, or the orientation of the image taken was not ideal.

Some have advocated for further imaging in the form of a CT scan when the assessment of the fracture is inconclusive, but this involves a significant amount of radiation.[1,2] In a recent study, CT imaging was not found to improve the interobserver agreement for measuring GT fracture displacement, but the surgeons felt slightly more confident about their treatment recommendation with the addition of a CT scan.[18] Mutch et al. established a measurement technique which involves calculating the GT fracture displacement ratio. This ratio separated isolated GT fractures into operative (ratio = 0.50 or more) and nonoperative (ratio = 0.00 or less) groups, with the intermediate ratios (ratio = 0.00–0.50) benefiting from a CT scan.[5] However, the intra and interobserver agreement for this measurement technique has proven to be poor in our study. We noted particular difficulty with the technique when the fracture had displaced posteriorly rather than superiorly.

Despite the lack of consistency in measuring the displacement distance and ratio, there was unexpectedly good intra and interrater agreement for the decision on whether to operate or not on the GT fracture. Indeed, this was the strongest area of agreement between all raters. It therefore seems that merely reviewing the images without any form of measurement seems to provide the best consistency in deciding treatment.

The senior surgeons had a slightly higher intra and interobserver reliability for classifying the fracture and decision to operate and better interobserver consistency for the measurement of displacement on AP and axillary plain radiograph views. However, the junior surgeons had better intraobserver consistency for the displacement measurement, and higher intra and interobserver reliability for the GT displacement ratio calculation. The results, however, did not yield any statistically significant difference between the two groups across all parameters. Furthermore, the presence of aberrant figures such as absolute '0' and '1', and the very wide margins of the CIs indicate significant imprecision in the results from this comparison, hindering the reliability and validity of its findings. This is likely attributed to the low number of raters involved in each comparative group.

## Limitations

Limitations should be accounted for when implementing the study findings. This is mainly attributed to the number of patients and raters included. The stratified analysis comparing the senior to junior surgeons demonstrated imprecise findings as a result of the low number of raters in each group. There is also the risk of bias in the selection of the raters and patients for the study.

## Conclusion

The best intra and interobserver consistency was when surgeons decided whether they would offer surgery based on the images provided. Otherwise, this study has revealed ongoing inconsistency in the classification and measurement of isolated GT fractures. Future studies on GT fractures should recognize the inconsistencies associated with using displacement distance as the sole criteria for treatment and should define clearly the classification method and measurement technique used to allow for more appropriate implementation in clinical practice and comparison of studies.

Research for identifying a more reliable classification system and measurement method is further recommended.

### Contributorship

All authors reviewed and approved the final version of the manuscript. GI researched the literature, wrote the initial and final versions of the manuscript. CO participated in the data collection and writing the initial version of the manuscript. SG participated in the data collection and writing the initial version of the manuscript. LB performed the data analysis and prepared a draft for the results section of the manuscript. IE participated in the data collection. HS participated in the data collection and analysis and writing the initial version of the manuscript. PM participated in the data collection. CT participated in the data collection and writing the initial version of the manuscript. MP participated in the data collection and writing the initial version of the manuscript. PC supervised the study, set the study design, participated in the writing the initial and final versions of the manuscript.

## ORCID iDs

Ghiath Ismayl https://orcid.org/0000-0002-8546-7499
Paul Cowling https://orcid.org/0000-0003-3656-8549

## References

1. White EA, Skalski MR, Patel DB, et al. Isolated greater tuberosity fractures of the proximal humerus: Anatomy, injury patterns, multimodality imaging, and approach to management. *Emerg Radiol* 2018; 25: 235–246.
2. Gruson KI, Ruchelsman DE and Tejwani NC. Isolated tuberosity fractures of the proximal humeral: Current concepts. *Injury* 2008; 39: 284–298.
3. Rouleau DM, Mutch J and Laflamme GY. Surgical treatment of displaced greater tuberosity fractures of the humerus. *J Am Acad Orthop Surg* 2016; 24: 46–56.
4. Parsons BO, Klepps SJ, Miller S, et al. Reliability and reproducibility of radiographs of greater tuberosity displacement. A cadaveric study. *J Bone Joint Surg Am* 2005; 87: 58–65.
5. Mutch JA, Rouleau DM, Laflamme GY, et al. Accurate measurement of greater tuberosity displacement without computed tomography: Validation of a method on plain radiography to guide surgical treatment. *J Orthop Trauma* 2014; 28: 445–451.
6. Mutch J, Laflamme GY, Hagemeister N, et al. A new morphological classification for greater tuberosity fractures of the proximal humerus: Validation and clinical implications. *Bone Joint J* 2014; 96B: 646–651.
7. Van Wier MF, Amajjar I, Hagemeijer NC, et al. Follow-up radiographs in isolated greater tuberosity fractures lead to a change in treatment recommendation; an online survey study. *Orthop Traumatol Surg Res* 2020; 106: 255–259.
8. McHugh ML. Interrater reliability: The kappa statistic. *Biochem Med (Zagreb)* 2012; 22: 276–282.
9. Koo TK and Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016; 15: 155–163.
10. Neer CS. 2nd Displaced proximal humeral fractures. I. Classification and evaluation. *J Bone Joint Surg Am* 1970; 52: 1077–1089.
11. Muller ME, Nazarian S, Koch P, et al. *The comprehensive classification of fractures of long bones*. Berlin: Springer Verlag, 1990, pp. 120–121.
12. Papakonstantinou MK, Hart MJ, Farrugia R, et al. Interobserver agreement of neer and AO classifications for proximal humeral fractures. *ANZ J Surg* 2016; 86: 280–284.
13. Sjödén GO, Movin T, Aspelin P, et al. 3D-radiographic Analysis does not improve the neer and AO classifications of proximal humeral fractures. *Acta Orthop Scand* 1999; 70: 325–328.
14. Foroohar A, Tosti R, Richmond JM, et al. Classification and treatment of proximal humerus fractures: Inter-observer reliability and agreement across imaging modalities and experience. *J Orthop Surg Res* 2011; 6: 38.
15. Majed A, Macleod I, Bull AM, et al. Proximal humeral fracture classification systems revisited. *J Shoulder Elbow Surg* 2011; 20: 1125–1132.
16. Brunner A, Honigmann P, Treumann T, et al. 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; 91: 766–771.
17. Razaeian S, Askittou S, Wiese B, et al. Inter- and intraobserver reliability of morphological mutch classification for greater tuberosity fractures of the proximal humerus: A comparison of X-ray, two-, and three-dimensional CT imaging. *PLoS ONE* 2021; 16: e0259646.
18. Janssen SJ, Hermanussen HH, Guitton TG, et al. Greater tuberosity fractures: Does fracture assessment and treatment recommendation vary based on imaging modality? *Clin Orthop Relat Res* 2016; 474: 1257–1265.