Imaging to improve agreement for proximal humeral fracture classification in adult patient: A systematic review of quantitative studies

Hannah Bougher^{*}, Archana Nagendiram, Jennifer Banks, Leanne Marie Hall, Clare Heal

Mackay Clinical School, James Cook University, Mackay Base Hospital, 475 Bridge Road, Mackay, Queensland, 4740, Australia

ARTICLE INFO

Article history:

Received 25 May 2019

Received in revised form

23 June 2019

Accepted 24 June 2019

Available online 26 June 2019

Keywords:

Inter-observer agreement

Intra-observer agreement

Proximal humeral fracture

Classification

Imaging

ABSTRACT

Proximal humeral fracture classification has low reproducibility. Many studies have tried to increase inter- and intra-observer agreement with more sophisticated imaging. The aim of this review was to determine which imaging modality produces the best inter- and intra-observer agreement for proximal humeral fracture classification in adults and to determine if this varies with observer experience or fracture complexity. OvidMEDLINE, The Cochrane Library, EBSCO CINAHL and Elsevier Scopus were searched on July 22nd, 2018. Quantitative studies comparing at least two imaging modalities for inter- or intra-observer agreement of proximal humeral fracture classification in adults were eligible for inclusion in this systematic literature review. Two reviewers independently screened and extracted data. Study quality was appraised using a modified Downs and Black checklist. The search strategy identified 1987 studies, of which 15 met the eligibility criteria. All included studies addressed inter-observer agreement and 8 provided results for intra-observer agreement. A narrative synthesis was performed. Trends were compared between studies as clinical heterogeneity and the statistical measures used by included studies prevented meta-analysis. Inter- and intra-observer agreement was found to increase from radiographs (x-ray) to two-dimensional (2D) computed tomography (CT) to three-dimensional (3D) CT. 2D and 3D CT may improve inter-observer agreement to a greater extent in less experienced observers and in more complex fractures. Future studies should compare 2D and 3D CT with subgroups categorising surgeon experience and fracture complexity. X-ray should be used for initial assessment; however doctors should have a low threshold for ordering CT.

PROSPERO number: CRD42018094307.

© 2019 Delhi Orthopedic Association. All rights reserved.

1. Introduction

Proximal humeral fractures are the fourth most common fracture in people over the age of 50,¹ affecting Australian men and women at rates of 40.6 and 73.2 per 100,000 person years respectively.² With the ageing population, it is estimated that the incidence may triple by 2030.^{3,4} Numerous classification systems exist for proximal humeral fractures including the Neer, Arbeitsgemeinschaft für Osteosynthesefragen (AO) and Hertel systems. These classify fractures according to the location and number of

fragments.⁵ Ideally, classification systems are clinically relevant; guiding management, predicting complications and providing prognostic information, as well as being accurate, reliable and valid research tools.^{6–11} In the absence of a gold standard for classification, studies use inter- and intra-observer agreement as surrogates for validity and reliability. Classification is invariably based on imaging and is a key determinant of whether a patient receives surgery.^{7,12–14} Accurate classification is paramount as it improves patient outcomes, reduces cost and makes treatment more consistent.^{13,15} Current classification systems have limited inter- and intra-observer agreement.^{11,14–17}

Many studies have compared agreement for proximal humeral fracture classification with radiographs (x-ray), two-dimensional (2D) and three-dimensional (3D) computed tomography (CT). Two systematic reviews^{15,17} synthesising studies from 1988 to 2003, had sections addressing the ability of imaging to improve inter-observer agreement of the Neer classification system. Mahadeva et al.¹⁵ reported that CT does not improve agreement.¹⁵

Abbreviations: X-ray, radiographs; 2D, two-dimensional; CT, computed tomography; 3D, three-dimensional.

^{*} Corresponding author. Mackay Clinical School, James Cook University, Mackay Base Hospital, 475 Bridge Road, Mackay, Queensland, 4740, Australia.

E-mail addresses: hannah.bougher@myjcu.edu.au (H. Bougher), archana.nagendiram@myjcu.edu.au (A. Nagendiram), jennifer.banks2@jcu.edu.au (J. Banks), leanne.hall@jcu.edu.au (L.M. Hall), clare.heal@jcu.edu.au (C. Heal).

Brorson et al.¹⁷ found no clear improvement with 2D or 3D CT. Three narrative reviews,^{10,11,14} of studies from 1987 to 2009, had sections evaluating the capacity of imaging to improve agreement. Robinson et al.¹⁴ found 2D CT provided mixed results and 3D CT yielded no advantage. Carofino et al.¹¹ concluded 2D and 3D CT produced mixed results, while Brorson et al.¹⁰ found 2D and 3D CT improved agreement. Conclusions from past reviews about the capacity of different imaging modalities to improve agreement for proximal humeral fracture classification are inconsistent. Since these reviews were published, 2D and 3D CT resolution has improved, making the images easier to interpret. No reviews were dedicated to exclusively assessing the capacity of imaging to increase agreement and none examined if the benefit of imaging varied with fracture complexity.

1.1. Objectives

The primary objective was to determine which imaging modality produces the best agreement for proximal humeral fracture classification in adult patients. We also aimed to determine if observer experience or fracture complexity impacts on which imaging modality produces greatest agreement.

2. Methods

The protocol was registered prospectively with PROSPERO (registration number: CRD42018094307, accessible at https://www.crd.york.ac.uk/PROSPERO/display_record.php?RecordID=94307).

2.1. Data sources

The search was run in OvidMEDLINE, The Cochrane Library, EBSCO CINAHL and Elsevier Scopus on the July 22nd, 2018. Keywords and MESH headings were searched without limits, for synonyms of “humeral fracture” and “imaging” and either “agreement” or “classification” (see Appendix for details). Reference lists of included articles and review papers were manually searched for additional studies. The authors of one study that indicated an additional relevant paper was published were contacted for further information.

2.2. Study selection

Two authors independently screened all identified titles and abstracts, then full-text manuscripts of potentially eligible papers. At each stage, discrepancies were resolved by discussion. If necessary, a third reviewer was consulted.

Studies were included if they: 1) contained quantitative primary research; 2) compared at least two imaging modalities (x-ray, CT or 3D printed models) before first definitive treatment (first intervention intended to manage the fracture, rather than symptoms or complications); 3) classified proximal humeral fractures, and 4) measured inter- or intra-observer agreement with kappa values, percentage agreement, intraclass correlation coefficients, Bland-Altman limits, percentage correctness or sensitivity and specificity. Classification allocates fractures to pre-described groups based on morphological characteristics. For example, the Neer system classifies fractures by the number of parts displaced by greater than 1 cm or a 45° angle.¹⁸ Inter-observer agreement is a measure of how frequently a group of observers agree about a certain fracture, while intra-observer agreement refers to how often an observer agrees with themselves at different time intervals.

Exclusion criteria were:

1. Full-text manuscript being unavailable in English.
2. Stress-fractures as they have specific, easily recognisable fracture patterns.¹⁹
3. Fractures caused by underlying tumours, because tumours alter fracture morphology.²⁰
4. Patients under 18 years old, as developmental changes to anatomy modify fracture morphology.²¹ Patients were assumed to be adult unless otherwise stated.

2.3. Data extraction

Two authors independently extracted data using a pre-piloted data extraction tool. Fracture and observer details (population, number, selection method, subgroups), imaging details (intervention, control, collection dates, presentation method), classification system, measure of intra- and inter-observer agreement, time between classification sessions, results and conclusions were extracted. One author was contacted but did not provide further details. Bias was assessed using a modified Downs and Black checklist.²² Scores above 18/24 were considered high quality.

2.4. Data synthesis

Narrative synthesis was undertaken by comparing trends between aggregate results for inter- and intra-observer agreement in each study. Where aggregate results were not available, the results were examined by study arm. 2D and 3D CT were referred to as “newer imaging modalities”, except when a study compared only 2D and 3D CT in which case 3D CT alone was considered the “newer imaging modality”. Pre-specified subgroups were observer experience, fracture complexity and Hill-Sachs lesions. Lower quality studies were interpreted with caution.

Meta-analysis of study findings was unfeasible because kappa values are not comparable outside individual studies due to prevalence and bias effects. These effects mean kappa values may have been low when the rate of agreement was high. Prevalence and bias effects result from marginal proportions which are an inherent component of calculating kappa values.^{23,24} Furthermore, heterogeneity between studies prevented meta-analysis.

3. Results

The search strategy identified 1959 articles (Fig. 1). An additional 28 articles were identified from reference lists of included articles. After duplicate removal, 1629 titles and abstracts were screened with 1591 excluded. Full manuscripts of 42 studies were accessed, of which 27 were ineligible (Fig. 1).

All studies were cross sectional surveys. Fractures were classified with newly proposed, Neer, AO and Hertel systems (Table 1). Images captured between 1992 and 2016 were displayed on a variety of media from interactive software to PowerPoint. Assessing whether imaging could improve agreement was the primary aim of eleven studies.^{6,8,9,25–32} No studies classified Hill-Sachs lesions, used 3D printed models or measured agreement with percentage correctness, Bland-Altman limits, sensitivity or specificity. Three studies only provided results by study arm without giving aggregate results for the inter- or intra-observer agreement (Table 2).^{13,25,30}

Eligible texts were of mixed quality scoring between thirteen and twenty out of twenty-four on the modified Downs and Black checklist (Table 1). Only three studies reported sample size calculations,^{13,27,30} five failed to adequately report statistical

Fig. 1. PRISMA flow diagram with arrows indicating how the screening process was undertaken. Adapted from Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. <https://doi.org/10.1371/journal.pmed.1000097>. For more information, visit www.prisma-statement.org.

results,^{5,6,8,9,26} only six presented confidence intervals,^{25,27,29,32–34} and seven did not randomly or consecutively select fractures.^{13,26,27,29–31,34}

3.1. Inter-observer agreement

Ten studies showed newer imaging techniques improved inter-observer agreement (Table 2). Classification based on x-ray most frequently had the lowest inter-observer agreement (Table 3).^{6,8,9,13,25,26,29–34} 2D CT was shown to improve agreement, providing the highest agreement in the majority of studies comparing 2D CT to other modalities for inter-observer agreement.^{5,13,25–31,33,34} and repeatedly outperformed x-ray (Table 4).^{5,13,25,26,28–34} 3D CT also increased agreement, generating the highest inter-observer agreement of all modalities in the majority of studies.^{6,8,9,25,27,28,30} 3D CT was found to regularly produce higher agreement than x-ray.^{6,8,9,25,28–30,32,34} 3D CT was also equivalent to, or improved agreement compared to 2D CT.^{25,27–30,34} Although most studies showed 3D CT was better than 2D CT which was better than x-ray, some studies were not consistent with these trends.

3.1.1. Level of observer experience and inter-observer agreement

Two studies showed newer imaging methods produce a greater improvement in agreement for inexperienced than experienced observers.^{8,25} Two studies found that only inexperienced observers had greater agreement with newer imaging, with 3D CT producing better agreement than 2D CT in less experienced observers, while experienced observers had poorer agreement with 3D CT.^{25,27} In contrast, two studies showed simpler imaging is more beneficial for inexperienced observers, finding that 2D CT results in lower

agreement than x-ray for inexperienced surgeons,²⁶ and that newer imaging only significantly improves agreement in upper extremity subspecialists.²⁹

3.1.2. Fracture complexity and inter-observer agreement

Foroozhar et al.²⁹ found newer imaging modalities improve agreement more as fracture complexity increases. Whilst 3D CT produced the best agreement across all fracture categories, this was only significant in four-part fractures, with progressive improvement from x-ray to 2D CT to 3D CT.²⁹

3.2. Intra-observer agreement

Newer imaging methods were demonstrated to improve intra-observer agreement in all eight studies which examined intra-observer agreement.^{6,9,26,28,31–34} X-ray was shown to generate the lowest agreement in all intra-observer agreement studies with the exception of a single study arm.^{6,9,26,28,31–34} 2D CT was found to improve agreement, achieving the highest agreement of all modalities in most studies which compared it to other imaging methods.^{26,28,31–34} With the exception of one study,³¹ 2D CT increased agreement over x-ray.^{26,28,32–34} 3D CT was also shown to increase intra-observer agreement, producing the highest agreement in most studies comparing it to other modalities.^{6,9,28,32,34} 3D CT also consistently outperformed x-ray with exception of a single study arm,^{6,9,26,28,31–34} and in most cases, 3D CT was found to increase agreement more than 2D CT.^{28,32,34}

3.2.1. Level of observer experience and intra-observer agreement

One study considered the association between observer experience and intra-observer agreement achieved with various

Table 1
Methodology and quality assessment of included studies.

Study	Number of observers	Number of fractures	Intervention	Control	Statistical test	Classification system	Modified Down and Black score
Foroohar et al. ²⁹ 2011	8	16	2D CT.	X-ray.	Kappa with Landis and Kock.	Neer.	20/24
Iordens et al. ³⁴ 2016	4	60	2D CT.	X-ray.	Kappa with Landis and Kock.	Neer.	19/24
Janssen et al. ³⁰ 2016	164	22	2D CT + x-ray.	X-ray.	Interclass correlation co-efficient and kappa with Landis and Kock.	Newly proposed.	19/24
Berkes et al. ²⁵ 2014	5	40	2D CT.	X-ray.	Kappa with Landis and Kock.	Neer.	19/24
Ramappa et al. ¹³ 2014	3	40	2D CT.	X-ray.	Kappa with Landis and Kock.	AO. Newly proposed.	19/24
Bruinsma et al. ²⁷ 2013	135	15	3D CT + x-ray.	2D CT + x-ray.	Kappa with Landis and Kock and percentage agreement.	AO. Newly proposed.	19/24
Brunner et al. ²⁸ 2009	4	40	2D CT + x-ray. Stereo-visualization of 3D CT + 2D CT + x-ray.	X-ray.	Kappa with Landis and Kock.	AO. Neer.	19/24
Meleán et al. ³³ 2017	3	96	2D CT.	X-ray.	Interclass correlation co-efficient and kappa with Landis and Kock.	Newly proposed.	18/24
Sallay et al. ⁸ 1997	9	12	3D CT.	X-ray.	Interclass correlation co-efficient and kappa with Landis and Kock.	Neer.	18/24
Sumrein et al. ⁹ 2018	3	116	3D CT.	X-ray.	Kappa with Landis and Kock.	Neer.	17/24
Bernstein et al. ²⁶ 1996	4	23	2D CT + x-ray.	X-ray.	Kappa with Landis and Kock.	Neer.	16/24
Ohl et al. ³² 2017	3	20	2D CT.	X-ray.	Kappa with Landis and Kock.	Newly proposed.	14/24
Matsushigue et al. ⁶ 2014	4	27	3D CT.	2D CT.	Kappa.	AO. Neer.	14/24
Mora Guix et al. ³¹ 2006	4	X-ray: 30 CT: 22	2D CT + x-ray.	X-ray.	Kappa with Landis and Kock.	Newly proposed.	14/24
Russo et al. ⁵ 2012	2	100	2D CT.	X-ray.	Percentage agreement.	Newly proposed.	13/24

Table note: Two dimensional has been abbreviated to 2D, computed tomography has been abbreviated to CT, radiograph has been abbreviated to x-ray, three dimensional has been abbreviated to 3D and “in addition to” has been abbreviated to +.

imaging modalities. It reported registrars and consultants had similar agreement with 2D CT and x-ray.²⁶

4. Discussion

In summary, newer imaging modalities were demonstrated to improve inter- and intra-observer agreement progressively from x-ray to 2D CT to 3D CT. Generally, newer imaging improved inter-observer agreement more for inexperienced observers, although this finding did not extend to intra-observer agreement. The literature suggests newer imaging improves inter-observer agreement further in more complex fractures.

These findings may reflect the difficulties in understanding complex 3D fracture lines from overlapping shadows seen on x-ray. This builds on the idea that 3D fractures are difficult to understand from x-ray,¹¹ and that overlapping fragments make displacement difficult to quantify.³⁵ 2D CT cross sections allow a better understanding of how fracture lines relate. 3D CT is particularly simple because observers can visualise the relationship between fracture lines. Robinson et al.¹⁴ proposed that 3D CT enabled recognition of additional fracture lines and patterns, although did not connect this to agreement. With an increasing number of fracture lines, complex fractures magnify this issue. Less experience with classifying humeral fractures may bring additional confusion.

Agreement is likely poorer when fractures are borderline between two fracture categories. This is amplified in complex fractures where the extent of bone displacement is more difficult to judge due to the displacement and/or deformation of the adjacent or surrounding bony anatomy. Newer imaging techniques are less ambiguous, therefore have the potential to increase agreement, particularly in

observers less accustomed to determining relevant cut-offs. This explanation is supported by Janssen et al.³⁰ and three reviews^{36–38} that found displacement was best quantified with newer imaging. Our explanation expands on the idea that fractures with marginal displacement require intraoperative classification,³⁵ and numerous fracture lines and osseous densities make displacement difficult to assess on x-ray.¹⁵ Contentious characteristics, including displacement, help determine prognosis and management, therefore should not be removed from classification systems to improve agreement.^{7,12,13} Additionally, studies that have tried this were unsuccessful.^{25,26} Instead, imaging should be used to increase agreement.

All studies measuring intra-observer agreement unanimously confirmed the finding that newer imaging improves agreement.^{6,9,26,28,31–34} However, regarding inter-observer agreement, there was some conflicting findings and three high quality studies disagreed.^{27,29,30} The finding that 3D CT was better than 2D CT is contentious. While 3D CT produced significantly higher inter-observer agreement (two out of three occasions)^{25,27,28} and better intra-observer agreement,^{28,32,34} 2D CT outperformed 3D CT when considering all inter-observer articles comparing the both modalities (five out of nine occasions).^{25,27–30,34}

The high quality study by Foroohar et al.²⁹ disagreed with our findings that newer imaging modalities increase agreement and that inexperienced observers benefit more from newer imaging methods. All images in this study were viewed in PowerPoint, which is similar to viewing x-rays in a clinical setting. In comparison, observers normally scroll through 2D CT and rotate 3D CT. This could explain why simpler x-ray performed best. Potentially, less familiarity with interpreting shoulder CTs amplified this issue, explaining why only experienced surgeons benefitted from CT.

Table 2
Findings for each publication.

Study	Inter-observer agreement			Intra-observer agreement		
	Trends	Statistically significant differences	Newer imaging is better	Trends	Statistically significant differences	Newer imaging is better
Foroohar et al. ²⁹ 2011	X-ray > 3D CT > 2D CT	X-ray > 2D CT.	No	Not measured.	Not applicable.	
Iordens et al. ³⁴ 2016	2D CT > 3D CT > x-ray.	Not significant.	Yes	3D CT > 2D CT > x-ray.	Not significant.	Yes
Janssen et al. ³⁰ 2016	Amount of displacement: 3D CT + 2D CT + x-ray > x-ray > 2D CT + x-ray. 5 mm displacement cut off: x-ray > 2D CT + x-ray > 3D CT + 2D CT + x-ray. Direction of displacement: 2D CT + x-ray > 3D CT + 2D CT + x-ray > x-ray.	Not significant.	No	Not measured.	Not applicable.	
Berkes et al. ²⁵ 2014	Experienced observers: 2D CT > 3D CT > x-ray. In-experienced observers: 3D CT > x-ray > 2D CT.	Experienced observers: 2D CT > x-ray. Inexperienced observers: 3D CT > 2D CT.	Yes	Not published although "did not improve through the use of 3D CT".	Not significant.	
Ramappa et al. ¹³ 2014	New system: 2D CT > x-ray. AO: x-ray > 2D CT.	10/12 aspects of new system: 2D CT > x-ray. AO: x-ray > 2D CT.	Yes	Not measured.	Not applicable.	
Bruinsma et al. ²⁷ 2013	2D CT + x-ray > 3D CT + x-ray.	2D CT + x-ray > 3D CT + x-ray.	No	Not measured.	Not applicable.	
Brunner et al. ²⁸ 2009	3D CT + 2D CT + x-ray > 2D CT + x-ray > x-ray.	3D CT + 2D CT + x-ray > 2D CT + x-ray > x-ray.	Yes	3D CT + 2D CT + x-ray > 2D CT + x-ray > x-ray.	3D CT + 2D CT + x-ray > 2D CT + x-ray > x-ray.	Yes
Meleán et al. ³³ 2017	2D CT > x-ray.	Not significant.	Yes	2D CT > x-ray.	Not significant.	Yes
Sallay et al. ⁸ 1997	3D CT > x-ray.	Not measured.	Yes	Not given separately by imaging method.	Not measured.	
Sumrein et al. ⁹ 2018	X-ray > x-ray + 3D CT > 3D CT.	Not measured.	No	X-ray + 3D CT > 3D CT > x-ray.	Not measured.	Yes
Bernstein et al. ²⁶ 1996	X-ray > 2D CT + x-ray.	Not measured.	No	2D CT + x-ray > x-ray.	Not measured.	Yes
Ohl et al. ³² 2017	2/8 aspects of new system: 3D CT and 2D CT > x-ray.	Not reported.	Yes	2D CT > 3D CT > x-ray.	Not reported.	Yes
Matsushigue et al. ⁶ 2014	3D CT > x-ray.	Not measured.	Yes	Neer: 3D CT > x-ray. AO: X-ray > 3D CT.	Not measured.	Yes
Mora Guix et al. ³¹ 2006	2D CT + x-ray > x-ray.	2D CT + x-ray > x-ray.	Yes	X-ray > 2D CT + x-ray.	2D CT + x-ray > x-ray.	Yes
Russo et al. ⁵ 2012	2D CT > x-ray.	Not measured.	Yes	Not measured.	Not applicable.	

Table note: Radiograph has been abbreviated to x-ray, "produces greater agreement than" has been abbreviated to >, three dimensional has been abbreviated to 3D, computed tomography has been abbreviated to CT, two dimensional has been abbreviated to 2D, "in addition to" has been abbreviated to + and Arbeitsgemeinschaft für Osteosynthesefragen has been abbreviated to AO.

Table 3
Number of occasions each imaging modality produced the highest, middle and lowest agreement.

	Highest agreement	Middle agreement	Lowest agreement
Inter-observer agreement			
X-ray	3 studies. ^{9,26,29} 2 study arms. ^{13,30}	2 study arms. ^{25,30}	7 studies. ^{6,8,31–34} 3 study arms. ^{13,25,30}
2D CT	5 studies. ^{5,27,31,33,34} 3 study arms. ^{13,25,30}	1 study. ²⁸ 1 study arm. ³⁰	2 studies. ^{26,29} 3 study arms. ^{13,25,30}
3D CT	3 studies. ^{6,8,28} 2 study arms. ^{25,30}	2 studies. ^{29,34} 2 study arms. ^{25,30}	2 studies. ^{9,27} 1 study arm. ³⁰
Intra-observer agreement			
X-ray	1 study. ³¹ 1 study arm. ⁶	0 occasions.	6 studies. ^{9,26,28,32–34} 1 study arm. ⁶
2D CT	3 studies. ^{8,26,32}	2 studies. ^{28,34}	1 study. ³¹
3D CT	3 studies. ^{9,28,34} 1 study arm. ⁶	1 study. ³²	1 study arm. ⁶

Table note: Radiograph has been abbreviated to x-ray, two dimensional has been abbreviated to 2D, computed tomography has been abbreviated to CT and three dimensional has been abbreviated to 3D.

Table 4
Comparison of agreement between each imaging modality.

	Inter-observer agreement	Intra-observer agreement
X-ray > 2D CT	2 studies. ^{26,29} 4 study arms. ^{13,25,30}	1 study. ³¹
2D CT > x-ray	6 studies. ^{5,28,31–34} 3 study arms. ^{13,25,30}	5 studies. ^{26,28,32–34}
3D CT > x-ray	6 studies. ^{6,8,25,28,32,34} 2 study arms. ³⁰ 2 studies. ^{9,29} 1 study arm. ³⁰	4 studies. ^{9,28,32,34} 1 study arm. ⁶ 1 study arm. ⁶
X-ray > 3D CT	2 studies. ^{28,29} 2 study arms. ^{25,30}	2 studies. ^{28,34}
3D CT > 2D CT	2 studies. ^{27,34} 3 study arms. ^{25,30}	1 study. ³²

Table note: Radiograph has been abbreviated to x-ray, "produces greater agreement than" has been abbreviated to >, two dimensional has been abbreviated to 2D, computed tomography has been abbreviated to CT and three dimensional has been abbreviated to 3D.

The finding that newer imaging methods improve inter-observer agreement to a greater extent in inexperienced observers was supported whether examining high quality studies alone, just studies which produced statistically results or all identified studies.^{8,25–27,29} However, significant findings in one high quality study disagreed.²⁹ The only study which examined intra-observer agreement by level of experience also did not support this finding, although it was of lower quality and the lack of aggregate results for each study arm made the results difficult to interpret.²⁶

The finding that newer imaging techniques are more useful in complex fractures is based on statistically significant results in only one high quality study.²⁹ This finding was supported by seven other studies that subdivided agreement by classification category, which generally concurred that newer imaging produced higher agreement for difficult characteristics such as displacement.^{8,13,27,30–33} However, these studies did not rank categories by complexity, so did not contribute to our conclusion.

This was the first review with the primary objective of determining whether imaging could improve agreement; it included more studies and more recently published articles than previous reviews with sections addressing this. This likely explains why this systematic review is more supportive of newer imaging modalities than previous narrative and systematic reviews. All previous literature reviews agreed x-ray should be used as a first line for imaging,^{10,11,14} with some arguing x-ray precludes the need for CT.³⁷ Most previous reviews agreed 2D CT has a role if x-ray is inadequate, however not because it improves agreement as we have suggested.^{15,36,38,39} Previous reviews were less supportive of 3D CT than our review, concluding it did not improve agreement,¹⁴ has little role,³⁶ or only has a role in pre-surgical planning.^{14,17,39} Possibly this is because 3D CT had limited quality 10 years ago and pixelated images could be more confusing than useful. The evidence indicating 3D CTs can improve agreement found in this current systematic review was mostly from studies published in the last decade.

4.1. Limitations of review

Limiting this review to English may have excluded relevant studies which could have strengthened our findings, however there is nothing to suggest language would introduce selection bias.

Publication bias may have caused overstatement of our conclusion that newer imaging techniques improve agreement, as it is possible studies favouring older imaging or showing no differences between imaging modalities may have been published less frequently.

This review inadvertently entailed selection bias towards complex fractures. By requiring articles to compare modalities, all studies included CT, a modality generally ordered for complex fractures.^{32,37} This may have exaggerated our conclusion that newer imaging improves agreement, given it provides greater advantage for complex fractures.

Selection bias within studies may also have impacted our findings. Seven authors either did not specify their method of selecting images, or specified a non-rational selection method,^{5,9,13,26,27,29,31} including three of five studies which found simpler imaging methods were superior.^{26,27,29} The remainder used rational selection methods of consecutive or random selection. Given agreement is possibly lower for borderline fractures, selection bias towards easily classified fractures may have muted our finding that newer imaging increases agreement. Only two studies disclosed how observers were selected.^{27,30} Twelve studies included more consultants than registrars.^{5,6,8,9,13,27,28,30,33,34} Hence, selection bias within studies may have undermined the finding that inexperienced observers benefit more from newer imaging.

The definition of “experienced observer” varied between studies from the level of training to the level of subspecialisation. Most

studies which found less experienced observers performed better with newer imaging defined “experienced observers” as upper limb subspecialists, while “less experienced” groups consisted mostly of other orthopaedic consultants. In contrast, studies which found experienced observers benefited more newer imaging defined an “experienced observer” as a consultant and an “inexperienced observer” as a registrar. In this way, varying definitions between studies may have biased our finding that less experienced observers receive greater benefit from newer imaging.

Nine articles did not reported statistically significant findings.^{5,6,8,9,26,30,32–34} This may have been a result of studies being underpowered, given twelve did not report sample size calculations.^{5,6,8,9,25,26,28,29,31–34} Nine studies did not report confidence intervals for the main study findings,^{5,6,8,9,13,26,28,30,31} making the results difficult to interpret. The conclusions need to be viewed in the context of these limitations.

4.2. Implications for research

Classification systems are likely intrinsically unreliable given attempts to improve agreement have only been partially successful.^{11,14,16,17} Despite newer imaging methods, suboptimal agreement is likely to continue with current classification systems. Therefore, development of new classification systems which result in better agreement should be considered. The success of imaging seen in this review warrants further research into improving agreement of current classification systems with innovative imaging methods until a newly developed classification system consistently produces higher agreement than current systems.

Although our search terms specifically sought 3D printed models as an imaging modality, no studies were identified. Foroohar et al.²⁹ proposed 3D structures are difficult to interpret via a 3D image on a 2D screen. Given newer imaging methods increase agreement and stereo-visualization of 3D CT improves agreement,²⁸ 3D printed models should be investigated. Models avoid the need to interpret the anatomy from imaging by reproducing the fracture. Most classification systems were designed to be applied after examining intraoperative anatomy.^{14,18} 3D models simulate the intraoperative findings more closely than imaging viewed through a 2D screen. Therefore, 3D models would allow classification systems to be applied more similarly to how they were designed, so may increase further agreement.

Future research should also investigate if 3D CT is beneficial over 2D CT and if newer imaging increases agreement further in complex fractures and inexperienced observers given the limitations of these findings. Kappa values correct for agreement by chance but entail prevalence and bias effect as described in section 2.4,^{23,24} whereas percentage agreement does not have prevalence and bias effects but also does not account for chance. Therefore, new studies should apply both measures.

4.3. Implications for clinical practice

While CT appeared to improve agreement, especially in complex fractures, it exposes patients to more radiation and cost than x-ray.⁴⁰ Therefore, x-ray should remain first line imaging. Pain from repositioning often prevents collection of adequate x-ray views, whereas CT avoids repositioning.^{14,15,36,39} By improving agreement about classification, CT likely produces more consistent decision making and better patient outcomes, thereby mitigating legal ramifications and lowering overall cost.^{13,15} CT also assists with preoperative planning.^{14,15,17,36,38,39} Therefore, if classification is not obvious from x-ray, 2D CT is a worthwhile second imaging modality and provides additional value for planning of operative management. 3D CT adds modest cost and no radiation,⁴⁰ and is also useful

for surgical planning, so is justifiable when 2D CT is insufficient. These recommendations support current practice in which patients initially receive an x-ray and frequently receive subsequent 2D and 3D CT scans as additional information is required. As newer imaging has been found to improve inter-observer agreement further for complex fractures, CT should be ordered once a complex fracture has been identified on x-ray. Given inexperienced observers benefit more from newer imaging modalities, these recommendations have special significance in public hospitals where a greater proportion of doctors are registrars.

5. Conclusions

This was the first systematic review with the primary objective of evaluating whether imaging could improve agreement about proximal humeral fracture classification. This systematic review also examined if this varied with fracture complexity and observer experience.

Conclusions are as follows:

1. Inter- and intra-observer agreement improves from x-ray to 2D CT to 3D CT.
2. Newer imaging modalities appear to improve inter-observer agreement more in less experienced observers.
3. Newer imaging may improve inter-observer agreement further in more complex fractures.

This was the first review to suggest that 3D CT improves agreement over 2D CT and that newer imaging methods may improve agreement more for inexperienced observers and more complex fractures, although support for these findings is limited. To confirm these findings, well powered future studies should compare 2D CT and 3D CT with subgroups categorising surgeon

experience and fracture complexity. These findings support current practice in which x-ray is first line with 2D and 3D CT frequently ordered when additional clarity is required, especially for complex fractures.

Declaration of interest

None.

Funding source

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Hannah Bougher: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization, Project administration. **Archana Nagendiram:** Validation, Investigation, Data curation. **Jennifer Banks:** Conceptualization, Methodology, Validation, Writing - review & editing, Supervision. **Leanne Marie Hall:** Writing - review & editing, Supervision. **Clare Heal:** Conceptualization, Methodology, Writing - review & editing, Supervision.

Acknowledgements

None.

Appendix. Search strategy used in Medline

- 1 exp Humeral Fractures/
- 2 exp Shoulder Fractures/
- 3 Fractures, Bone/
- 4 exp HUMERUS/
- 5 exp Shoulder Joint/
- 6 exp SHOULDER/
- 7 3 and 4
- 8 3 and 5
- 9 3 and 6
- 10 1 or 2 or 7 or 8 or 9
- 11 (humer* or shoulder* or glenohumer*).mp. [mp = title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
- 12 (break* or broken or fract* or injury).mp. [mp = title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
- 13 11 and 12
- 14 10 or 13
- 15 *Classification/
- 16 **Severity of Illness Index"/
- 17 *trauma severity indices/
- 18 15 and 16 and 17
- 19 (Neer* or classif* or Taxonom* or assess* tool* or sever* rating* or sever* indic* or sever* index* or sever* scale* or sever* score* or clinical diagnostic criteria or clinical diagnostic tool or clinical diagnostic scale or clinical assessment tool or clinical assessment criteria or clinical assessment scale or diagnostic rat* scale*).mp. [mp = title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
- 20 18 or 19
- 21 exp "Reproducibility of Results"/
- 22 exp Data Accuracy/
- 23 exp Observer Variation/
- 24 21 or 22 or 34
- 25 (reliab* or valid* or reproduc* or kappa* or accura* or inter-observ* or inter observ* or interobserv* or intra observ* or intraobserv* or inter rater* or interrater* or intra rater* or intrarater* or inter coder* or intercoder* or intra coder* or intracoder* or (Landis and Kock*) or test retest or observer varia* or varia* observ*).mp. [mp = title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
- 26 24 or 25

(continued)

- 27 exp X-rays/
 28 exp Tomography, X-Ray Computed/
 29 exp Imaging, Three-Dimensional/
 30 exp Magnetic Resonance Imaging/
 31 exp Printing, Three-Dimensional/
 32 exp Tomography, X-Ray/
 33 27 or 28 or 29 or 30 or 31 or 32
 34 (X-ray* or x-ray* or x ray* or X ray* or xray* or Xray* or X-radi* or X-radi* or x radi* or X radi* or xradi* or Xradi* or cat scan* or CT or ct or (compute* adj5 tomograph*) or tomodensitometry or (3 d adj5 imag*) or (3 D adj5 imag*) or (3-d adj5 imag*) or (3-D adj5 imag*) or (3d adj5 imag*) or (3D adj5 imag*) or (three dimensional adj5 imag*) or (three-dimensional adj5 imag*) or 3 dimensional NEAR imag* or (3-dimensional adj5 imag*) or magneti* resonance imaging or MRI* or mri* or (mr adj5 tomography) or (magnet* resonance adj5 tomography) or magnet* resonance spectrometry or (3 d adj5 print*) or (3 D adj5 print*) or (3-d adj5 print*) or (3-D adj5 print*) or (3d adj5 print*) or (3D adj5 print*) or (three dimensional adj5 print*) or (three-dimensional adj5 print*) or (3 dimensional adj5 print*) or (3-dimensional adj5 print*)).mp. [mp = title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
 35 33 or 34
 36 14 and 20 and 35
 37 14 and 26 and 35
 38 36 or 37

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Med* 6(7): e1000097. <https://doi.org/10.1371/journal.pmed1000097>.

References

- Court-Brown CM, Caesar B. Epidemiology of adult fractures: a review. *Injury*. 2006;37(8):691–697. <https://doi.org/10.1016/j.injury.2006.04.130>.
- Holloway KL, Bucki-Smith G, Morse AG, et al. Humeral fractures in south-eastern Australia: epidemiology and risk factors. *Calcif Tissue Int*. 2015;97(5): 453–465. <https://doi.org/10.1007/s00223-015-0039-9>.
- Palvanen M, Kannus P, Niemi S, Parkkari J. Update in the epidemiology of proximal humeral fractures. *Clin Orthop Relat Res*. 2006;442:87–92. <https://doi.org/10.1097/01.blo.0000194672.79634.78>.
- Bell JE, Leung BC, Spratt KF, et al. Trends and variation in incidence, surgical treatment, and repeat surgery of proximal humeral fractures in the elderly. *J Bone Joint Surg Am*. 2011;93(2):121–131. <https://doi.org/10.2106/JBJS.I.01505>.
- Russo R, Cautiero F, Rotonda GD. The classification of complex 4-part humeral fractures revisited: the missing fifth fragment and indications for surgery. *Musculoskelet Surg*. 2012;96(Suppl 1):S13–S19. <https://doi.org/10.1007/s12306-012-0195-2>.
- Matsushige T, Franco VP, Pierami R, Tamaoki MJS, Netto NA, Matsumoto MH. Do computed tomography and its 3D reconstruction increase the reproducibility of classifications of fractures of the proximal extremity of the humerus? *Rev*. 2014;49(2):174–180. <https://doi.org/10.1016/j.rboe.2014.03.016>.
- Resch HM. Proximal humeral fractures: current controversies. *J Shoulder Elb Surg*. 2011;20(5):827–832. <https://doi.org/10.1016/j.jse.2011.01.009>.
- Sallay PI, Pedowitz RA, Mallon WJ, Vandemark RM, Dalton JD, Speer KP. Reliability and reproducibility of radiographic interpretation of proximal humeral fracture pathoanatomy. *J Shoulder Elb Surg*. 1997;6(1):60–69. [https://doi.org/10.1016/S1058-2746\(97\)90072-0](https://doi.org/10.1016/S1058-2746(97)90072-0).
- Sumrein BO, Mattila VM, Lepola V, Laitinen MK, Launonen AP, NITEP Group. Intraobserver and interobserver reliability of recategorized Neer classification in differentiating 2-part surgical neck fractures from multi-fragmented proximal humeral fractures in 116 patients. *J Shoulder Elb Surg*. 2018;27(10): 1756–1761. <https://doi.org/10.1016/j.jse.2018.03.024>.
- Brorson S, Bagger J, Sylvest A, Hróbjartsson A. Diagnosing displaced four-part fractures of the proximal humerus: a review of observer studies. *Int Orthop*. 2009;33(2):323–327. <https://doi.org/10.1007/s00264-008-0591-2>.
- Carofino BC, Leopold SS. Classifications in brief: the Neer classification for proximal humerus fractures. *Clin Orthop Relat Res*. 2013;471(1):39–43. <https://doi.org/10.1007/s11999-012-2454-9>.
- Okike K, Lee OC, Mankanji H, Harris MB, Vrahas MS. Factors associated with the decision for operative versus non-operative treatment of displaced proximal humerus fractures in the elderly. *Injury*. 2013;44(4):448–455. <https://doi.org/10.1016/j.injury.2012.09.002>.
- Ramappa AJ, Patel V, Goswami K, et al. Using computed tomography to assess proximal humerus fractures. *Am J Orthoped*. 2014;43(3):e43–47. <http://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=med8&AN=24660183>. Accessed August 1, 2018.
- Robinson BC, Athwal GS, Sanchez-Sotelo J, Rispoli DM. Classification and imaging of proximal humerus fractures. *Orthop Clin N Am*. 2008;39(4):393–403. <https://doi.org/10.1016/j.jocl.2008.05.002>.
- Mahadeva D, Mackay DC, Turner SM, Drew S, Costa ML. Reliability of the Neer classification system in proximal humeral fractures: a systematic review of the literature. *Eur J Orthop Surg Traumatol*. 2008;18(6):415–424. <https://doi.org/10.1007/s00590-008-0325-6>.
- Brorson S. Fractures of the proximal humerus. *Acta Orthop*. 2013;84(Suppl. 351):1–32. <https://doi.org/10.3109/17453674.2013.826083>.
- Brorson S, Hróbjartsson A. Training improves agreement among doctors using the Neer system for proximal humeral fractures in a systematic review. *J Clin Epidemiol*. 2008;61(1):7–16. <https://doi.org/10.1016/j.jclinepi.2007.04.014>.
- Neer CS. Displaced proximal humeral fractures: part I. classification and evaluation. *Clin Orthop Relat Res*. 2006;442:77–82. <https://doi.org/10.1097/01.blo.0000198718.91223.ca>.
- Fayad LM, Kamel IR, Kawamoto S, Bluemke DA, Frassica FJ, Fishman EK. Distinguishing stress fractures from pathologic fractures: a multimodality approach. *Skeletal Radiol*. 2005;34(5):245–259. <https://doi.org/10.1007/s00256-004-0872-9>.
- Zarin F, Kazemi T, Vakili P. A review of malignant bone tumors. *Res J Pharmaceut Biol Chem Sci*. 2015;6(2):884–893. [https://www.rjpbcs.com/pdf/2015_6\(2\)/128j.pdf](https://www.rjpbcs.com/pdf/2015_6(2)/128j.pdf). Accessed August 28, 2018.
- Lefèvre Y, Journeau P, Angelliaume A, Bouty A, Dobremez E. Proximal humerus fractures in children and adolescents. *Orthop Traumatol Surg Res*. 2014;100(1): S149–S156. <https://doi.org/10.1016/j.otsr.2013.06.010>.
- Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *J Epidemiol Community Health*. 1998;52(6):377–384. <https://doi.org/10.1136/jech.52.6.377>.
- Hoehler FK. Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. *J Clin Epidemiol*. 2000;53(5):499–503. [https://doi.org/10.1016/S0895-4356\(99\)00174-2](https://doi.org/10.1016/S0895-4356(99)00174-2).
- Feinstein AR, Cicchetti DV. High agreement but low kappa: the problems of two paradoxes. *J Clin Epidemiol*. 1990;43(6):543–549. [https://doi.org/10.1016/0895-4356\(90\)90158-L](https://doi.org/10.1016/0895-4356(90)90158-L).
- Berkes MB, Dines JS, Little MT, 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(15):1281–1286. <https://doi.org/10.2106/JBJS.M.00199>.
- 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(9):1371–1375. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1027.4889&rep=rep1&type=pdf>. Accessed August 28, 2018.
- 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(17):1600–1604. <https://doi.org/10.2106/JBJS.L.00586>.
- Brunner A, Honigsmann P, Treumann T, Babst R. The impact of stereo-visualisation of three-dimensional CT datasets on the inter- and intra-observer reliability of the AO/OTA and Neer classifications in the assessment of fractures of the proximal humerus. *J Bone Joint Surg Br*. 2009;91(6):766–771. <https://doi.org/10.1302/0301-620X.91B6.22109>.
- 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(1):38. <https://doi.org/10.1186/1749-799X-6-38>.
- Janssen SJ, Hermanussen HH, Guitton TG, van den Bekerom MPJ, van Deuren DFF, Ring D. Greater tuberosity fractures: does fracture assessment and treatment recommendation vary based on imaging modality. *Clin Orthop Relat Res*. 2016;474(5):1257–1265. <https://doi.org/10.1007/s11999-016-4706-6>.
- Mora Guix JM, Gonzalez AS, Bruggala JV, Carril EC, Banos FG. Proposed protocol for reading images of humeral head fractures. *Clin Orthop Relat Res*. 2006;448: 225–233. <https://doi.org/10.1097/01.blo.0000205899.28856.98>.
- Ohl X, Mangin P, Barbe C, Brun V, Nerot C, Sirveaux F. Analysis of four-fragment fractures of the proximal humerus: the interest of 2D and 3D imagery and

- inter- and intra-observer reproducibility. *Eur J Orthop Surg Traumatol*. 2017;27(3):295–299. <https://doi.org/10.1007/s00590-017-1911-2>.
33. Meleán P, Munjin A, Pérez A, Rojas JT, Cook E, Fritis N. Coronal displacement in proximal humeral fractures: correlation between shoulder radiographic and computed tomography scan measurements. *J Shoulder Elb Surg*. 2017;26(1):56–61. <https://doi.org/10.1016/j.jse.2016.05.016>.
  34. Iordens GI, 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(5):596–602. <https://doi.org/10.1016/j.jos.2016.05.011>.
  35. Neer CS. Four-segment classification of proximal humeral fractures: purpose and reliable use. *J Shoulder Elb Surg*. 2002;11(4):389–400. <https://doi.org/10.1067/mse.2002.124346>.
  36. Green A, Izzi J. Isolated fractures of the greater tuberosity of the proximal humerus. *J Shoulder Elb Surg*. 2003;12(6):641–649. <https://doi.org/10.1016/S1058274602868112>.
  37. Gruson KI, Ruchelsman DE, Tejawani NC. Isolated tuberosity fractures of the proximal humeral: current concepts. *Injury*. 2008;39(3):284. <https://doi.org/10.1016/j.injury.2007.09.022>.
  38. Iyengar JJ, Ho J, Feeley BT. Evaluation and management of proximal humerus fractures. *Phys Sportsmed*. 2011;39(3):52–61. <https://doi.org/10.3810/psm.2011.02.1862>.
  39. Handford C, Nathoo S, Porter K, Kalogrianitis S. A review of current concepts in the management of proximal humerus fractures. *Trauma*. 2015;17(3):181–190. <https://doi.org/10.1177/1460408614555069>.
  40. The Australian Government: Department of Health. Search the MBS. Medicare benefits scheme online website. <http://www9.health.gov.au/mbs/search.cfm>. Accessed September 1, 2018. Updated August 31, 2018.