Classifying Intertrochanteric Fractures of the Proximal Femur: Does Experience Matter?

Wayne Fung^a Anders Jönsson^b Volker Bühren^b Mohit Bhandari^{a, b}

^aDivision of Orthopaedic Surgery, McMaster University and the Clinical Research Group, Hamilton, Ont., Canada;

^bOsteosynthesis and Trauma Care Foundation, Trauma Care Institute, Nice, France

Key Words

Hip · Proximal femur · Intertrochanteric fractures

Abstract

Objective: This prospective study aimed to evaluate inter-observer agreement of the AO/OTA and Evans/Jensen classification systems with an emphasis on the effect of reviewer experience. In addition, the level of agreement in the determination of fracture stability across varying levels of experience was evaluated. **Subject and Methods:** A group of 12 reviewers (6 surgeons, 3 senior residents, 3 junior residents) independently assessed 56 radiographs of intertrochanteric hip fractures and classified them using the AO/OTA and Evans/Jensen systems. Reviewers further assessed the stability of the fracture patterns. **Results:** Overall, higher agreement was obtained for the AO/OTA classification than with the Evans/Jensen classification. However, neither classification system met acceptable thresholds for reliability. Surgeons were unable to reliably determine if fractures were stable or unstable. Increased experience improved AO/OTA classification reliability and assessment of stability but decreased agreement in the Evans/Jensen classification. **Conclusion:** The AO/OTA classification should be used in favor of the Evans/Jensen classification whenever possible.

Our findings suggest that surgeons' perceptions about stability vary to a significant extent thereby necessitating clear definitions of stability.

Copyright © 2007 S. Karger AG, Basel

Introduction

By the year 2040 the number of people aged 65 or older will increase from 34.8 to 77.2 million [1, 2], and the number of hip fractures is likely to exceed 500,000 annually with an estimated annual health care cost of at least USD 9.8 billion [2–4].

Previous studies evaluating alternative implants for managing intertrochanteric fractures have utilized fracture stability or fracture classification systems to include, or exclude, patients from study participation [5]. In addition, the conclusions of a study may further stipulate the class of fracture to which an implant, or technique, may be indicated. However, these recommendations are based on the fundamental assumption that investigators and readers can reliably classify fractures of the hip. In other words, the recommendation that intramedullary hip screws are most useful in unstable intertrochanteric fractures has the fundamental premise that we are able to

reliably agree which fractures are stable and which fractures are unstable.

Ideally, a classification system should easily be applicable, reliable and aid in treatment decision making as well as indicate outcome [6]. In this difficult area it is often best to classify fractures by consensus, which can stimulate debate regarding the appropriate treatment options. In order to do this, any classification system used should aim to possess a high degree of inter-observer reliability.

Several classification systems exist for intertrochanteric fractures – the Evans/Jensen and the AO/OTA have been the most widely used [7–12]. Despite the widespread use of these systems and over 11,000 publications on the topic of hip fractures, few studies have evaluated their reliability [7–12]. Only two previous studies have evaluated whether surgeons can reliably assess fracture stability [11, 12]. A recent study suggests that inexperienced reviewers in previous studies may have biased results and fuelled the concern for otherwise good classification systems [9]. Given the limited reports, the current prospective study aimed to evaluate inter-observer agreement of the AO/OTA and Evans/Jensen classification systems with an emphasis on the effect of reviewer experience. Agreement in the determination of fracture stability across varying levels of experience was also evaluated.

Materials and Methods

Identification of Cases

Fifty-six sets of anteroposterior and lateral radiographs of intertrochanteric fractures were selected from a single institution trauma database of 3,000 intertrochanteric fractures. Radiographs were chosen by 2 authors, independent of the reviewers to include a range of possible fracture types with varying severities. Radiographs were eligible for inclusion in the study if they represented an extracapsular (intertrochanteric) hip fracture and were of sufficient radiographic quality to allow review for classification. The radiographs chosen represented a spectrum of possible fracture types to ensure representation of a sufficient variety of fractures. Although no formal power calculation was conducted, the sample size of 56 radiographs was a compromise, based upon previous studies with sample sizes ranging from 20 to 88 radiographs [7–9].

Reviewers

A group of 12 reviewers was made up of 3 junior orthopaedic surgical trainees (postgraduate years 1–2), 3 senior orthopaedic surgical trainees (postgraduate years 4–5) and 6 orthopaedic traumatologists (2 Canadian, 4 Danish). They independently evaluated the 56 radiographs. The orthopaedic trainees represented 3 different institutions.

The examiners were blinded to patient history, subsequent operative treatment approach, and clinical information for all sets of radiographs. Inter-observer reliability was evaluated by comparing the responses of the 12 observers at the initial viewing of the radiographs.

Radiographic Assessments

Each reviewer was requested to classify the fracture by the AO/OTA and the Evans/Jensen classifications [13, 14]. The AO/OTA classification system is organized into hierarchies including A, B and C types. Intertrochanteric fracture patterns are further subdivided into increasing severities from A1 to A3. The A subtypes are further divided into A1.1, A1.2 and A1.3 to represent increasing severity of the fracture [15]. The modified grading proposed by Jensen and Michaelson in 1975 was intended to improve the predictive value of the Evans system, to indicate which fractures could be reduced anatomically and which were at risk for secondary displacement after fixation [13]. Class I includes two-fragment fractures, which are considered stable. Class II contains Evans type III and type IV fractures, which are difficult to reduce in either the coronal or the sagittal plane, while class III (Evans type V) consists of very unstable fractures, which are difficult to reduce in both planes. Class I and class II fractures are further subdivided into A and B types. All reviewers were given diagrams of the AO and Evans/Jensen classifications for review prior to classifying the 56 radiographs. Reviewers were further requested to identify whether they believed the fractures were stable or unstable. No specific training in the use of the systems or criteria for stability was provided to reviewers. Finally, reviewers were requested to grade the importance of the lateral hip radiograph when classifying hip fractures.

Analysis

Intraclass correlation coefficients (ICC) with 95% confidence intervals were used to measure agreement. The ICC, used to quantify agreement for a continuous variable, is equivalent to the quadratically weighted kappa (κ) for categorical data. Landis and Koch [16] suggest agreement of 0–0.2 represents slight agreement, 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement, and 0.61–0.80 substantial agreement. A value above 0.80 is considered almost perfect agreement [17]. We evaluated overall agreement across 12 reviewers and subdivided agreement by experience categories (junior trainee, senior trainee, and traumatologist). We compared agreement across subgroups of experience using statistical tests for interaction. The value of $p < 0.05$ was taken as statistically significant and all tests were two-tailed.

Results

The overall inter-observer agreement across 12 reviewers for the AO/OTA classification was moderate (ICC = 0.56, 95% confidence interval: 0.47–0.67, table 1). With increase in experience, significant improvement in assessment skill was observed: from junior to senior orthopaedic surgical trainees (ICC = 0.27 and 0.61, respec-

Table 1. Inter-observer agreement across AO/OTA and Evans classifications

	Intraclass correlation coefficients (95% confidence intervals)			
	overall	junior residents (n = 3)	senior residents (n = 3)	staff trauma surgeons (n = 6)
AO/OTA (overall)	0.56 (0.47–0.67)	0.27 (0.009–0.49)	0.61 (0.47–0.73) ⁺	0.49 (0.37–0.62)
Evans (overall)	0.34 (0.24–0.49)	0.52 (0.23–0.71)	0.52 (0.36–0.66)	0.20 (0.11–0.33) ⁺
Stability	0.38 (0.28–0.50)	0.29 (0.03–0.51)	0.37 (0.20–0.54)	0.53 (0.40–0.65)*
Was lateral radiograph useful?	0.07 (0.02–0.14)	0.01 (–0.14 to 0.19)	0.15 (–0.003 to 0.31)	0.02 (–0.01 to 0.09)

* p = 0.08; ⁺ p = 0.02 when compared with junior residents.

Table 2. Interobserver agreement for subgroups of AO/OTA and Evans classifications

	Intraclass correlation coefficients (95% confidence intervals)			
	overall	junior residents (n = 3)	senior residents (n = 3)	staff trauma surgeons (n = 6)
AO/OTA, main group	0.59 (0.48–0.69)	0.25 (–0.06 to 0.47)	0.39 (0.14–0.60)	0.56 (0.41–0.69)
AO/OTA, subgroup	0.14 (0.07–0.24)*	0.35 (0.10–0.56)	0.17 (0.03–0.34)	0.09 (–0.6 to 0.26)*
Evans, main group	0.39 (0.27–0.53)	0.31 (0.06–0.51)	0.38 (0.13–0.58)	0.27 (0.1–0.45)
Evans, subgroup	0.03 (–0.02 to 0.17)*	0.02 (–0.05 to 0.15)*	0.27 (–0.09 to 0.56)	0.04 (–0.1 to 0.25)*

* p < 0.01 when compared to main group.

tively, $p < 0.05$); however, consultant surgeons, although trending towards improved agreement, did not significantly differ from junior residents (table 1). Reviewers performed consistently worse in classifying the subgroups of the AO/OTA classification (table 2).

The Evans/Jensen classification achieved an overall fair agreement (ICC = 0.34, 95% confidence interval: 0.24–0.49, table 1). Increasing experience led to a significant decline in agreement from junior and senior orthopaedic surgical trainees to staff traumatologists (ICC = 0.52 and 0.20, respectively, $p < 0.05$). Reviewers performed consistently worse in classifying the subgroups of the Evans/Jensen classification (table 2).

Reviewers achieved fair agreement in determining stability of the fracture patterns across the 56 radiographs (ICC = 0.38, 95% confidence interval: 0.28–0.50). Increasing experience led to improved ability to assess fracture stability (ICC = 0.29–0.53, $p < 0.05$).

The reviewers achieved poor agreement (ICC = 0.07, 95% confidence interval: 0.02–0.14) in their perceptions about the usefulness of the lateral radiographs in classifying the fractures (table 1).

Neither classification system (AO/OTA or Evans/Jensen) met acceptable thresholds for reliability; surgeons were unable to reliably determine if fractures were stable or not, and increasing experience improved AO/OTA classification reliability and assessment of stability but decreased agreement in the Evans/Jensen classification.

Discussion

Overall, this study has several strengths. First, the large number of reviewers with varying experience levels enabled assessment of learning curves and further improved the generalizability of the findings to all experience levels. The mix of North American and European reviewers further improved the generalizability of the findings. The decision to evaluate surgeons' perceptions of stability further informs the difficulties with current terminologies. Radiographs represented a variety of fracture types taken from a larger dataset.

The sample size of radiographs which were somewhat limited resulted in large confidence intervals around the

Table 3. Relevant literature

Authors	Year	Reviewers	X rays	AO	Evans, κ	Stability	Experience
Our findings	2007	12	56	0.56	0.34	yes	3 groups: junior trainees, senior trainees, trauma surgeons
Wen-Jie et al. [9]	2005	5	40	0.80 (0.38)	0.22	NA	only experienced
Chapman et al. [7]	2003	5	61	0.47	0.23	NA	mixed (2 senior residents/3 consultants)
Pervez et al. [8]	2002	5	88	0.33	0.34	NA	mixed (1 junior/2 senior, 2 surgeons)
Schipper et al. [10]	2001	15	20	0.33	NA	NA	mixed (5 surgeons, 5 radiologists, 5 residents) ^a
Gehrchen et al. [11]	1993	4	52	NA	0.44 ^b	yes ^c	all residents ^a
Andersen et al. [12]	1990	6	49	NA	0.56–0.67	yes ^c	mixed (3 surgeons, 3 residents) ^a

^a Assessed the learning curve.

^b No κ calculated, only the raw agreement presented.

^c Did not specifically ask reviewers to determine whether fracture was stable or unstable/used prespecified classification cutoffs for stable and unstable.

point estimates of agreement. This was a pragmatic decision to provide information without overburdening participants. The decision not to calculate the intra-observer reliability was made since it was thought that for interpretation, achieving high inter-observer agreement was of greater importance. Furthermore, by the nature of the calculations for inter-observer agreement, the intra-observer agreement will always be higher [6]. The calculation for inter-observer agreement included both the between-subject and 'within'-subject variability. Thus, it would always be lower than the intra-observer value [6].

The overall findings of the current study are generally consistent with previous reports (table 3). Similarly to other reports, the reviewers in this study achieved higher agreement for the AO/OTA classification than with the Evans/Jensen classification (table 1). Although four previous studies [7, 8, 10, 11] included a mix of reviewers with varying levels of experience, only two of these provided any analysis of a learning curve [10–12]. Schipper et al. [10] evaluated agreement using the AO/OTA classification in 20 radiographs, across 15 reviewers (5 radiologists, 5 surgeons, and 5 trainees). They reported no appreciable differences in assessment by different professionals (between residents, surgeons and radiologists, $\kappa = 0.69, 0.62$ and 0.65 , respectively) with varying experience. In an earlier study evaluating the Evans/Jensen classification, Gehrchen et al. [11] used 4 reviewers (2 senior residents, 2 junior residents) and 52 radiographs to evaluate agreement. These authors found no difference in agreement with increasing experience.

Contrary to these reports, the findings in this study show that having greater experience improved reliability

in the assessment of the AO/OTA classification of intertrochanteric hip fractures. It remains plausible that the previous studies did not have either a sufficient range of experience among reviewers, or there were too few reviewers in subgroups to identify any differences that might have existed. To overcome the limitations of these previous studies, this study included at least 3 reviewers in each category and professionals with a wider range of experience levels ranging from first year trainees to surgeon traumatologists. Somewhat unexpectedly, it was found that increasing experience led to a decline in agreement for the Evans/Jensen classification. It remains unclear whether or not another classification system is needed. The findings here would suggest that improved training and experience are likely to improve reliability.

Increasing experience led to improved agreement about the stability of the intertrochanteric fractures. However, surgeons achieved only moderate agreement for stability (ICC = 0.53, table 1). Two earlier studies published over a decade ago [11, 12] evaluated agreement for stability in the Evans/Jensen classification. Gehrchen et al. [11] reported that only 44% of 4 reviewers achieved complete agreement for the Evans classification, but 65% agreed when the classification was subdivided into the stable and unstable variants ('the assessment of stability'). Andersen et al. [12], in a study of 6 reviewers and 49 radiographs, recoded responses to either stable or unstable fractures, based upon the Evans/Jensen classification (type I = unstable and II = stable). These investigators reported that only 18% of reviewers agreed upon the overall classification but when subdivided into stable and unstable categories, there was 57% agreement [12]. The as-

assessment of stability was distinctly different from previous reports, in that the surgeons were asked specifically to state whether they judged the fractures were stable or unstable. Previous reports did not ask reviewers but simply recategorized their classifications into stable and unstable categories [11, 12]. This is somewhat misleading as it remains unclear whether the reviewers would have come to the same conclusions as the authors of those papers.

Conclusion

This study highlights the continuing difficulties in the utilization of the current classification systems for intertrochanteric fractures. Readers should be aware of the mix of the individual assessors who classified fractures for eligibility into research studies.

In addition, this study raises concern about the validity of previous study recommendations that focus on implant choice, based on the determination of whether an

intertrochanteric fracture is stable or unstable. Future investigators may overcome the limitations we have identified in this study, if they classify fractures (eligibility into a study, for subgroup analyses of a cohort of patients) with experienced investigators. The AO/OTA classification should be used in favour of the Evans/Jensen classification whenever possible. It is recommended that clear definitions of stability be used since our findings suggest that surgeon's perceptions about stability vary to a significant extent.

Acknowledgements

The authors are grateful to those trainees and surgeons who donated their time as reviewers for our study. Specifically we thank Dr. Olle Svensson, Dr. Kjeld Andersen, Dr. Carl Ekholm, Dr. Lars Borris, Dr. Jaydeep Moro, Dr. Brad Petrisor, Dr. Femi O. Ayeni, Dr. Jamal Rakem, Dr. Aziz Al-Mutair, Dr. Michael Zlodowski, and Dr. Boris Zelle. This study was funded by a research grant from AIOD, Nice, France. Also, we are grateful to Bernard Robionek for his critical review of the manuscript.

References

- 1 Cummings SR, Rubin SM, Black D: The future of hip fractures in the United States. Numbers, costs, and potential effects of postmenopausal estrogen. *Clin Orthopaed Relat Res* 1990;252:163–166.
- 2 Cooper C, Campion G, Melton LJ: Hip fractures in the elderly: a world-wide projection. *Osteoporos Int* 1992;2:2859.
- 3 Wilkins K: Health care consequences of falls for seniors. *Health Rep* 1999;10:47–55.
- 4 AAOS urges hip fracture reform. *Am Acad Orthop Surg Bull* 1999;47:3.
- 5 Parker MJ, Handoll HH: Intramedullary nails for extracapsular hip fractures in adults. *Cochrane Database Syst Rev* 2005;2:CD004961.
- 6 Streiner DL, Norman GR: *Health Measurement Scales: a Practical Guide to Their Development and Use*, ed 3. London, Oxford University Press, 2003.
- 7 Chapman C, Herrera M, Binenbaum BS, et al: Classification of intertrochanteric fractures with computed tomography: a study of intra-observer and inter-observer variability and prognostic value. *Am J Orthop* 2003;32:443–449.
- 8 Pervez H, Parker M, Pryor G, et al: Classification of trochanteric fracture of the proximal femur: a study of the reliability of current systems. *Injury* 2002;33:713–715.
- 9 Wen-Jie J, Li-Yang D, Yi-Min C: Reliability of classification systems for intertrochanteric fractures of the proximal femur in experienced orthopaedic surgeons. *Injury* 2005;36:858–861.
- 10 Schipper IB, Steyerberg W, Castelein R, et al: Reliability of the AO/ASIF classification for peritrochanteric femoral fractures. *Acta Orthop Scand* 2001;72:36–41.
- 11 Gehrchen PM, Neilsen J, Olesen B: Poor reproducibility of the Evan's classification of the trochanteric fracture: assessment of 4 observers in 52 cases. *Acta Orthop Scand* 1993;64:71–72.
- 12 Andersen E, Jorgensen L, Heddam L: Evan's classification of trochanteric fractures: an assessment of the interobserver and intraobserver reliability. *Injury* 1990;21:377–378.
- 13 Jensen JS: Classification of trochanteric fractures. *Acta Orthop Scand* 1980;51:803–810.
- 14 Mueller M, Allgower M, Schneider R, Willenegger H: *Manual of Internal Fixation*. New York, Springer, 1991.
- 15 Fracture and dislocation compendium. Orthopaedic Trauma Association Committee for Coding and Classification. *J Orthop Trauma* 1996;10(suppl 1):v–ix, 1–154.
- 16 Landis JR, Koch GG: The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–174.
- 17 Fleiss JL: *Statistical Methods for Rates and Proportions*, ed 2. New York, Wiley, 1981.