High reliability of an algorithm for choice of implants in hip fracture patients

Henrik Palm • Eva Posner • Hans-Ulrik Ahler-Toftehøj •
Peter Siesing • Silas Gylvin • Tobias Aasvang •
Kim Holck • Kenneth Brian Holtz

Received: 3 January 2013 / Accepted: 9 February 2013 / Published online: 28 March 2013
© Springer-Verlag Berlin Heidelberg 2013

Abstract

Purpose Hip fracture treatment is controversial, with high complication rates. An algorithm for hip fracture surgery has shown reduced reoperation rates, but choice of implant is based on the commonly used fracture classifications, which were previously evaluated to be unreliable. The purpose of this study was to investigate the reliability of the algorithm. **Methods** From two hospitals, four observers (orthopaedic consultant, fellow, resident and intern) used the algorithm to classify into 15 hip fracture types [Garden type I–IV femoral neck including posterior tilt, vertical femoral neck, basocervical and Arbeitsgemeinschaft für Osteosynthesefragen (AO)-31 A1.1 to A3.3 trochanteric fractures] and to choose between five surgical procedures [parallel implants, prosthesis, two- or four-hole sliding hip screw (SHS) and intermedullary (IM) nail]. After individual assessment, each hospital made a consensus decision. Observations were performed twice, ten weeks apart, on pelvic, anteroposterior (AP) and axial X-rays from 100 consecutive patients.

Results For fracture classification, mean kappa values were 0.60 for intra and 0.62 for interobserver variation, with interobserver variation between hospitals at 0.65. For posterior tilt, mean intraclass correlation coefficient was 0.91 for intra and 0.87 for interobserver variation. For choice of implant type, mean kappa values were 0.86 for both intra and interobserver variation. The two hospital consensus decisions chose same implant in 91 of 100 patients, giving a kappa value at 0.88.

Conclusion Although hip fracture classification confirmed to be somewhat unreliable in this study, posterior tilt measurement and subsequent choice of implant type by the algorithm was found to be reliable, which opens up the possibility for a more standardized treatment of hip fracture patients between hospitals.

Introduction

Hip fracture treatment is controversial, with high complication rates. However, an easy-to-implement algorithm (Fig. 1), which standardizes choice of implant type, has been shown to reduce reoperation rates [1]. The algorithm is, however, based on the commonly used fracture classification systems, primarily the Garden classification [2], of femoral-neck fractures and the Arbeitsgemeinschaft für Osteosynthesefragen/Orthopaedic Trauma Association (AO/OTA) classification [3] of trochanteric fractures; both previously evaluated to have only a fair to moderate reliability [4–8].

Unreliable classification systems could result in different treatment choices between surgeons and hospitals, even when using this new algorithm. Therefore, we investigated the interobserver and intra-observer reliability for choice of implant type when surgeons from two different hospitals used the algorithm.

Patients and methods

Eight observers were grouped by workplace in two neighbouring suburban hospitals annually treating a total of 1,200 patients with a hip fracture: Hospital 1 consisted of authors from Hvidovre Hospital: SG (intern), EP (resident), HP (fellow) and KH (consultant). Hospital 2

H. Palm (✉) • E. Posner • S. Gylvin • K. Holck
Department of Orthopaedic Surgery, Hvidovre Hospital,
Copenhagen, Denmark
e-mail: hpalm@dadlnet.dk

H.-U. Ahler-Toftehøj • P. Siesing • T. Aasvang • K. B. Holtz
Department of Orthopaedic Surgery, Herlev Hospital,
Copenhagen, Denmark

Fig. 1 Algorithm for choice of implant type in hip fracture surgery is based on fracture classification and patient age

consisted of authors from Herlev Hospital: PS (intern), TA (resident), HA (fellow) and KB (consultant). The algorithm was invented and implemented in Hospital 1, whereas Hospital 2 had a 30 minute introduction with images on 20 patients. All observers marked their choice in an individual pad of numbered printed copies of the algorithm (Fig. 1). The algorithm allows classification into 15 different hip fracture types: the four stages of femoral neck fractures in Gardens classification, which are commonly termed Garden type I–IV fractures [1, 2], the vertical femoral neck fracture

[9, 10], the basocervical fracture [11, 12] and the nine trochanteric fractures 31A1.1 to A3.3 in the AO/OTA classification [3]. If any observer classified a fracture as a Garden type I or II, all observers also measured the posterior tilt [13]. Based on classification, posterior tilt and patient age stated on the radiographs, observers then chose between five surgical procedures (parallel implants, prosthesis, two-hole sliding hip screw, four-hole sliding hip screw or intramedullary nail). After individual assessment of fracture and procedure, each hospital group made a consensus

decision for each patient for simulating a clinical trauma-conference decision.

In both groups, assessments were performed twice in random order exactly ten weeks apart on the same 100 consecutive hip fracture patients (73 women and 27 men, average age 80 years) admitted to Hvidovre Hospital from September to November 2011. Radiographs (pelvic, anteroposterior and lateral) had been obtained on patient admission and stored in the Image Management and Applications Radiology Information Service (IMPAX-RIS) system (Agfa, Köln, Germany). Seated around a conference table, observers watched the images in Digital Imaging and Communications in Medicine (DICOM) format projected to a 2×2,5-m wall screen. Measurement of posterior tilt was performed digitally with the assessor seated in front of a computer screen. All individual assessments were performed blinded for coauthor influence. The study was part of the hip fracture project at Hvidovre Hospital, Copenhagen, Denmark, and was approved by the Danish Data Protection Agency and the Copenhagen Ethical Committee, which concluded that the nature of the study was such that written patient consent was not required.

Statistics

Cohen's unweighted kappa statistics was performed to determine intra and interobserver variation for classifying into the 15 fracture types and deciding between the five different implants. Learning curves have been described previously in hip fracture surgery [14], so improvement between sessions was checked with a paired *t* test of interobserver values for the first and second sessions. Intra- and interclass correlation coefficients (two-way mixed-effect model, absolute agreement definition, single measure) were performed between measurements of the posterior tilt. Values were interpreted following recommendations by Landis and Kock as poor (<0), slight (0–0.2), fair (0.21–0.4), moderate (0.41–0.6), substantial (0.61–0.8) and almost perfect (0.81–1) agreement [4]. The total number of assessments was 4,272, with 2,000 ((eight observers and two consensus decisions)×100 patients×2 sessions) of fracture classification, each followed by a total of 2,000 assessments of choice of implant type, plus 272 (eight observers×17 patients with a Garden type I–II femoral-neck fracture×2 sessions) measurements of posterior tilt. All calculations were performed with SPSS 19.0 statistical software (Chicago, IL, USA).

Results

For classifying into the 15 fracture types in the algorithm, mean kappa values were 0.62 (range 0.54–0.68) for interobserver variation in the first session, 0.60 (range 0.48–0.78) for intra-observer variation, and 0.73 (range 0.62–0.87) for interobserver variation between individual observers and

Table 1 Inter- and intra-observer kappa values in the two hospitals, each with four observers

Classification into 15 fracture types in the algorithm					
Hospital 1, observers	KH	HP	EP	SG	Intra-observer
KH	–	–	–	–	0.57
HP	0.68	–	–	–	0.78
EP	0.66	0.65	–	–	0.61
SG	0.57	0.57	0.64	–	0.48
Consensus	0.74	0.87	0.66	0.62	0.64
Hospital 2, observers	KB	HA	TA	PS	Intra-observer
KB	–	–	–	–	0.58
HA	0.64	–	–	–	0.64
TA	0.64	0.66	–	–	0.51
PS	0.59	0.54	0.59	–	0.59
Consensus	0.81	0.78	0.71	0.64	0.58
Choosing between the five implant types in the algorithm					
Hospital 1, observers	KH	HP	EP	SG	Intra-observer
KH	–	–	–	–	0.89
HP	0.92	–	–	–	0.92
EP	0.92	0.90	–	–	0.86
SG	0.84	0.85	0.85	–	0.77
Consensus	0.96	0.96	0.92	0.85	0.89
Hospital 2, observers	KB	HA	TA	PS	Intra-observer
KB	–	–	–	–	0.89
HA	0.89	–	–	–	0.90
TA	0.85	0.81	–	–	0.77
PS	0.82	0.85	0.78	–	0.85
Consensus	0.94	0.92	0.82	0.85	0.86

Table 2 Consensus for fracture classification between the two hospitals. Agreement in 70 of 100 patients is indicated in bold

		Hospital 2:															
		G-I	G-II	G-III	G-IV	Vert.	Basoc.	A1.1	A1.2	A1.3	A2.1	A2.2	A2.3	A3.1	A3.2	A3.3	Total
Hospital 1:	G-I	1	2	–	–	–	–	–	–	–	–	–	–	–	–	–	3
	G-II	2	10	1	–	–	–	–	–	–	–	–	–	–	–	–	13
	G-III	1	–	18	4	–	–	–	–	–	–	–	–	–	–	–	23
	G-IV	–	–	2	12	–	–	–	–	–	–	–	–	–	–	–	14
	Vert.	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	Basoc.	-	-	1	–	–	1	–	–	–	–	–	–	–	–	–	2
	A1.1	–	–	–	–	–	3	3	1	–	–	1	–	–	1	–	9
	A1.2	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	A1.3	–	–	–	–	–	–	–	–	1	–	–	–	–	–	–	1
	A2.1	–	–	–	–	–	–	–	1	–	–	2	–	–	–	–	3
	A2.2	–	–	–	–	–	–	–	–	–	1	16	2	–	–	–	19
	A2.3	–	–	–	–	–	–	–	–	–	–	–	3	–	–	–	3
	A3.1	–	–	–	–	–	–	–	–	–	–	–	–	1	–	–	1
	A3.2	–	–	–	–	–	–	–	–	–	–	1	–	–	2	1	4
	A3.3	–	–	–	–	–	–	–	–	–	–	1	1	–	1	2	5
	Total	4	12	22	16	–	4	3	2	1	1	21	6	1	4	3	100

G-I to G-IV Garden type I to Garden type IV femoral neck fractures, Vert. Vertical femoral neck fractures, Basoc. Basocervical fractures, A1.1 to A3.3 AO/OTA type 31A1.1 to 31A3.3 trochanteric fractures

hospital consensus decisions (Table 1). Vertical, basocervical and A1.1 trochanteric fractures were most difficult to determine differences [3, 9–12].

Both hospital consensus decisions classified into the same fracture types in 70 of the 100 patients (Table 2), giving a kappa value for interobserver variation between hospitals of 0.65. Except for one patient, hospitals agreed on dividing into intra- and extracapsular fractures. The kappa value was 0.52 within the 46 agreed-upon extracapsular fractures. Within the 53 agreed-upon intracapsular fractures, the kappa value was 0.67 for dividing into all four Garden types and 0.91 if later being grouped into simply undisplaced (Garden type I–II) or displaced (Garden type III–IV) fractures [2, 5].

One or more observer classified fractures as Garden type I–II in 17 patients, in which all observers then had to measure the posterior tilt [13]. The mean individual interclass correlation

coefficient was 0.87 (range 0.74–0.94) for interobserver variation in the first session and 0.91 (range 0.83–0.95) for intraobserver variation. In two patients, different observers measured the posterior tilt to be on each side of the 20° borderline for choice of implant type in the algorithm. All observers chose the correct implant type in the algorithm based on their assessment of fracture classification, measured posterior tilt and patient age. Mean kappa values for choice of implant type were 0.86 (range 0.78–0.92) for interobserver variation in the first session, 0.86 (range 0.77–0.92) for intraobserver variation and 0.90 (range 0.82–0.96) for interobserver between individual observers and hospital consensus decisions (Table 1). No improvement in interobserver values was observed between sessions ($p=0.44$). The two hospital consensus decisions chose the same implant type in 91 of the 100 patients (Table 3), giving a kappa value for interobserver variation between hospitals at 0.88. Except for

Table 3 Consensus for choice of implant type between hospitals. Agreement in 91 of 100 patients is indicated in bold

		Hospital 2				
		Par. Impl.	Prosthesis	2-hole SHS	4-hole SHS	IM-nail
Hospital 1	Par. Impl.	18	–	–	–	–
	Prosthesis	–	35	–	–	–
	2-hole SHS	1	–	1	–	–
	4-hole SHS	–	–	3	6	4
	IMN	–	–	–	1	31
	Total	19	35	4	7	35
						100

Par. Impl. parallel implants,
SHS sliding hip screw,
IM-nail intramedullary nail

one patient, hospitals fully agreed on treatment with parallel implants and prosthesis, whereas most disagreement was seen among treatment with two- and four-hole sliding hip screws.

Discussion

The algorithm (Fig. 1) was proposed for standardizing hip fracture surgery between surgeons and hospitals [1]. This study shows that although the algorithm relies on the commonly used, only moderately reliable, fracture classifications, implants were chosen with an almost perfect agreement between observers [2–8]. Also, the new measurement for posterior tilt [13] revealed to be both assessable and reliable, with doubt regarding subsequent algorithm choice of implant type occurring in only two patients. An optimal classification system should be reliable and useful for differentiating treatment and prognosis in clinical and research situations. However, previous studies have shown the Garden classification to have only a fair reliability, improving to moderate if simply dividing into undisplaced or displaced fracture groups [5]. Also, the AO/OTA classification for trochanteric fractures has shown only a fair reliability, improving to substantial if using the three main groups only [6–8]. Classifying into vertical and basocervical fractures among all types of hip fractures has not previously been evaluated, but the vertical fracture could be seen as the third of the three types in the Pauwel classification, which shows only fair reliability [9–12].

Compared with these studies, our kappa values for fracture classification were higher. This study did, however, differ from previous studies of reliability of classification systems: (1) it included all proximal femoral fractures, (2) observers classified cases directly onto fracture drawings in their individual pad of printed copies of the algorithm, and (3) a formal introduction was given of the algorithm with radiographs of 20 pilot patients. The latter two differences could theoretically improve agreement between observers, the latter by allowing the observers to agree on how to classify borderline situations before the study. This was also allowed during the study in discussions for reaching the consensus decision after each individual observation. The effect of this practice must, however, be minimal, as improvement did not occur between sessions.

Agreement between the two hospitals was also achieved without the two groups of observers ever meeting. Some individual classification outliers were removed in the consensus decisions, but this only reflects the actual clinical trauma conferences. Compared with previous studies, we believe that our study design is more optimal for the purpose of evaluating a clinical algorithm regarding choice of implant type, as it better reflects the actual clinical

situation in a major orthopaedic department with pocket-sized classifications/algorithms, formal introduction and trauma conferences.

Although optimal classification systems are still missing, the algorithm proved to have a close to perfect agreement [4], even among observers not previously familiar with it. Based on well-known classifications, the five different implant types in the algorithm are easily chosen. The simplicity of grouping into fewer choices also improves agreement, as shown in previous studies of the Garden and AO/OTA classifications [5–8]. The main disagreements in the algorithm were seen among vertical, basocervical and A1.1 trochanteric fractures [3, 9–12]. Vertical and basocervical fractures must be separated from the common femoral-neck fracture, as they lack varus calcar support for parallel implants and theoretically need a fixed-angle device [1, 12, 15]. For simplifying the algorithm and potentially raising the agreement even further, treating these fractures with the four-hole sliding hip screw could be considered, despite the longer incision.

The algorithm for hip fracture surgery has previously reduced the reoperation rate and is easy implemented [1]. This study shows it to also be reliable, which opens up the opportunity for a more standardized treatment of hip fracture patients by different surgeons and between different hospitals.

Acknowledgements We thank biostatisticians Janne Petersen and Steen Ladelund, Clinical Research Center, Hvidovre Hospital, for statistical support.

Conflict of interest None.

References

1. Palm H, Krashenninnikoff M, Holck K, Lemser T, Foss NB, Jacobsen S, Kehlet H, Gebuhr P (2012) A new algorithm for hip fracture surgery. *Acta Orthop* 83:26–30
2. Garden RS (1961) Low-angle fixation in fractures of the femoral neck. *J Bone Joint Surg Br* 43:647–663
3. Orthopaedic Trauma Association Classification, Database and Outcomes Committee (2007) Fracture and dislocation classification compendium-2007. *J Orthop Trauma* 10:Suppl.
4. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* 33:159–174
5. van Embden D, Rhemrev SJ, Genelin F, Meylaerts SAG, Roukema GR (2012) The reliability of a simplified Garden classification for intracapsular hip fractures. *Orthop Traumatol Surg Res* 98:405–408
6. van Embden D, Rhemrev SJ, Meylaerts SAG, Roukema GR (2010) The comparison of two classifications for trochanteric femur fractures: The AO/ASIF classification and the Jensen classification. *Injury* 41:377–381
7. Pervez H, Parker MJ, Pryor GA, Lutchman L, Chirodian N (2002) Classification of trochanteric fracture of the proximal femur: a study of the reliability of current systems. *Injury* 33:713–715

8. Schipper IB, Steyerberg EW, Castelein RM, van Vugt AB (2001) Reliability of the AO/ASIF classification for pertrochanteric femoral fractures. *Acta Orthop Scand* 72:36–41
9. Parker MJ (2009) Results of internal fixation of Pauwels type-3 vertical femoral neck fractures. Letter to the editor with reply from Haidukewych GJ. *Joint Surg Am* 91:490–491
10. van Embden D, Roukema GR, Rhemrev SJ, Genelin F, Meylaerts SAG (2011) The pauwel classification for intracapsular hip fractures: is it reliable? *Injury* 42:1238–1240
11. Mallick A, Parker MJ (2004) Basal fractures of the femoral neck: intra or extracapsular. *Injury* 35:989–993
12. Massoud EIE (2010) Fixation of basicervical and related fractures. *Int Orthop* 34:577–582
13. Palm H, Gosvig KK, Krasheninnikoff M, Jacobsen S, Gebuhr P (2009) A new measurement for posterior tilt predicts reoperation in undisplaced femoral neck fractures. *Acta Orthop* 80:303–307
14. Bjorgul K, Novicoff WM, Saleh KJ (2011) Learning curves in hip fracture surgery. *Int Orthop* 35:113–119
15. Alho A, Benterud JG, Roenningen H, Hoeiseth A (1991) Radiographic prediction of early failure in femoral neck fracture. *Acta Orthop* 62:422–426