

Preoperative classification assessment reliability and influence on the length of intertrochanteric fracture operations

Jing Shen · FangKe Hu · LiHai Zhang · PeiFu Tang · ZhengGang Bi

Received: 25 October 2012 / Accepted: 2 December 2012 / Published online: 21 December 2012
© Springer-Verlag Berlin Heidelberg 2012

Abstract

Purpose The accuracy of intertrochanteric fracture classification is important; indeed, the patient outcomes are dependent on their classification. The aim of this study was to use the AO classification system to evaluate the variation in classification between X-ray and computed tomography (CT)/3D CT images. Then, differences in the length of surgery were evaluated based on two examinations.

Methods Intertrochanteric fractures were reviewed and surgeons were interviewed. The rates of correct discrimination and misclassification (overestimates and underestimates) probabilities were determined. The impact of misclassification on length of surgery was also evaluated.

Results In total, 370 patents and four surgeons were included in the study. All patients had X-ray images and 210 patients had CT/3D CT images. Of them, 214 and 156 patients were treated by intramedullary and extramedullary fixation systems, respectively. The mean length of surgery was 62.1 ± 17.7 min. The overall rate of correct discrimination was 83.8 % and in the classification of A1, A2 and A3

were 80.0, 85.7 and 82.4 %, respectively. The rate of misclassification showed no significant difference between stable and unstable fractures (21.3 vs 13.1 %, $P=0.173$). The overall rates of overestimates and underestimates were significantly different (5 vs 11.25 %, $P=0.041$). Subtracting the rate of overestimates from underestimates had a positive correlation with prolonged surgery and showed a significant difference with intramedullary fixation ($P<0.001$).

Conclusions Classification based on the AO system was good in terms of consistency. CT/3D CT examination was more reliable and more helpful for preoperative assessment, especially for performance of an intramedullary fixation.

Introduction

The number of intertrochanteric fractures has increased rapidly due to the aging population [1]. Surgical fixation has been demonstrated to result in better outcomes [2, 3]. The most commonly used surgical fixations are the intramedullary (IM) fixation system and the extramedullary (EM) fixation system [4]. Which is better—IM or EM—has long been debated [5–7]. IM has been used increasingly in recent years, but there is no substantial evidence to support its use [4].

Also, the fracture classification system—as a tool—should provide the surgeon with a reasonably precise estimation of the likely outcome [8]. The type of classification can have a great effect on patient outcome [9]. If the preoperative classification is not correct, the usefulness of this prognostic formula is limited. Thus adequate preoperative evaluation of bone fragment conditions is important. Various systems have been used to classify intertrochanteric fractures. Of them, the AO classification system has been used widely in recent years; it was proposed by Muller and colleagues in the 1980s [10].

J. Shen · Z. Bi (✉)
Orthopedic Department, The First Affiliated Hospital of Harbin
Medical University, Harbin Medical University,
Harbin, People's Republic of China
e-mail: bizheng@gmail.com

F. Hu
Orthopedic Department, Tianjin Hospital, Tianjin, People's
Republic of China

J. Shen · L. Zhang · P. Tang (✉)
Orthopedic Department, Chinese PLA General Hospital, Beijing,
People's Republic of China
e-mail: peifutang@gmail.com

With advances in radiography, traditional classification methods based on plain radiographs are now considered insufficient and are associated with marked variation according to the level of experience of the surgeon [11–14]. Currently, computed tomography (CT) and three-dimensional reconstruction CT techniques (3D CT) are used for preoperative assessment at our institute. However, few reported studies have addressed whether these advanced examinations are, in fact, more useful.

The aim of this study was to evaluate variation in the AO classification between X-ray and CT/3D CT examinations. Furthermore, the length of surgery with classification as a related factor was assessed to evaluate the effectiveness of CT/3D CT examinations.

Participants and methods

Intertrochanteric fracture patients were reviewed continuously from 1 July 2008 to 1 July 2010 at our institution. Inclusion criteria were an intertrochanteric fracture within two weeks, treated by open reduction with internal fixation. Exclusion criteria were conservative treatment, multiple fractures, treated by external fixation, a second operation and an operative procedure interrupted by complications. Length of surgery and the identity of the surgeon were obtained from medical records. Radiology images were obtained from the Radiology Department. The most frequently involved surgeons were invited to participate in the study. Patients not treated by the invited surgeons were then excluded. The AO classification system was used in this study.

Consistency test Patients were divided into two groups. Group 1 contained patients who underwent both X-ray and CT/3D CT examinations before surgery. Group 2 contained the remaining patients who were operated on by the invited surgeons and underwent X-rays but not a CT/3D CT examination before surgery.

Surgeons were invited to classify the fracture types. An example of the flowsheet used is shown in Fig. 1. First, X-ray and CT/3D CT images from group 1 were randomly shown unpaired and were classified by the surgeons independently. Classification B1 was based on X-ray examination while B2 was based on CT/3D CT. Statistical analyses were used to evaluate the consistency of the surgeons. The surgeons were also asked about procedure length. A surgeon would be excluded if the agreement value was less than 0.80 and their procedure length differed significantly from the others. Also, patient images were removed if the patient was treated by an excluded surgeon. The remaining patients in group 1 became group 3 and the remaining patients in group 2 became group 4.

Standard data of classification Next, the included surgeons first estimated the images of group 1 together, based on

paired X-ray and CT/3D CT images. These data were called classification ‘S’ and we considered them to be the standard classification. Another data set consisted of estimates using the X-ray images of groups 1 and 4 together. These we called data set C based on group 1 and D based on group 4.

Impact of misclassification Data sets C and S were compared and the rates of correct discrimination and misclassification were determined. Misclassification was divided into overestimates and underestimates. The procedure length and the influence of misclassification in the same classification type between groups 3 and 4 were compared.

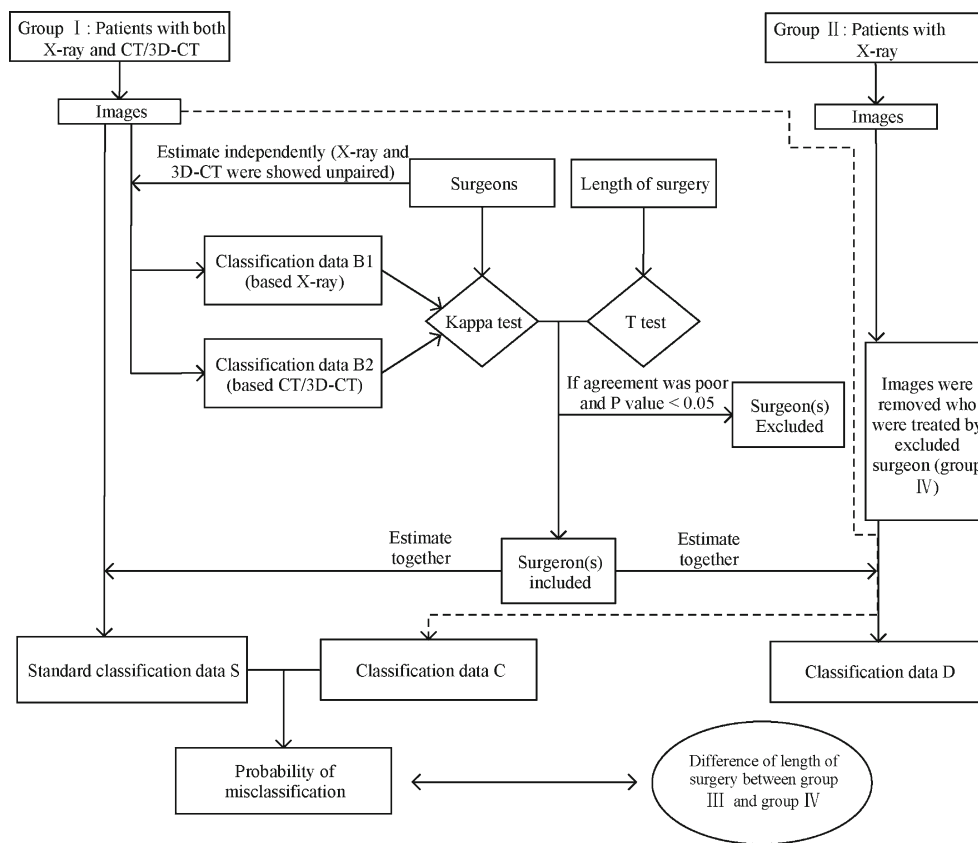
Statistical analysis Student’s *t* test or the Wilcoxon rank-sum test was performed on numerical data among groups. The chi-square test was used for categorical data. Data sets B1 and B2 were analysed using the κ coefficient of agreement to quantify agreement between observers. The κ coefficient value is in the range of -1 to $+1$; -1 indicates complete disagreement, 0 is the level of agreement expected by chance and $+1$ is complete agreement. The Landis and Koch guideline provided the basis for our interpretation of the reliability estimate (greater than 0.80 represents almost perfect agreement) [15]. The level of significance was set at $P < 0.05$. All analyses were performed using the SPSS software (version 13.0, SPSS Inc., IBM, Chicago, IL, USA).

Results

In total, 424 patients were reviewed and 32 were excluded. Reasons for exclusion were: four had a fracture over two weeks earlier, six received conservative treatment, seven had multiple fractures, three were treated with an external fixation system, ten were second operations and two procedures were interrupted by intraoperative complications. Of the remaining patients, 370 patients were treated by four surgeons (103 by chief Dr. Tang, 90 by associate chief Dr. Zhang, 90 by associate chief Dr. Liang and 87 by associate chief Dr. Tao) and accounted for 94.4 % (370/392) of the cases.

There were 160 patients in group 1 and 210 in group 2. Of them, 214 and 156 patients were treated with the IM and EM systems, respectively. In group 1, 87 (54.4 %) and 73 (45.6 %) patients were treated with the IM and EM systems, respectively. In group 2, 125 (59.5 %) and 85 (40.5 %) patients were treated with the IM and EM systems, respectively. The overall length of surgery was 62.1 ± 17.7 min (in group 1 and group 2, 60.4 ± 17.1 min and 63.4 ± 18.1 min, respectively). In group 1, the time was 62.3 ± 20.2 min for IM and 57.7 ± 10.7 min for EM (Wilcoxon test, $z = -0.809$, $P = 0.418$, > 0.05). In group 2, the time was 67.5 ± 20.9 min for IM and 57.2 ± 10.4 min for EM (Wilcoxon test,

Fig. 1 Flowsheet of this study



$z=-3.279$, $P=0.001$). The procedure length did not differ significantly among the four surgeons in group 2, by either IM or EM (Wilcoxon test, each $P>0.05$; Table 1). In group 1, IM and EM showed a statistically significant difference ($P=0.024$, 0.022 and 0.030, respectively). After multiple comparisons, only Dr. Tao showed a significant difference from Drs. Zhang and Liang (Wilcoxon test, each $P<0.05$).

The numbers of classifications in data sets B1 and B2 are shown in Fig. 2. The κ coefficient values based on X-ray images (data set B1) were 0.815 between Drs. Tang and Zhang, 0.801 between Drs. Tang and Liang, 0.787 between Drs. Tang and Tao, 0.815 between Drs. Zhang and Liang, 0.688 between Drs. Zhang and Tao and 0.808 between Drs. Liang and Tao. The κ coefficient values based on CT/3D CT images (data set

B2) were 0.900 between Drs. Tang and Zhang, 0.909 between Drs. Tang and Liang, 0.886 between Drs. Tang and Tao, 0.914 between Drs. Zhang and Liang, 0.924 between Drs. Zhang and Tao and 0.873 between Drs. Liang and Tao.

Dr. Tao was excluded from the analysis due to the low agreement and a procedure length that differed significantly from the other surgeons. Thus, Drs. Tang, Zhang and Liang were included in subsequent analyses. The patients operated on by Dr. Tao were removed from groups 3 and 4. Thus, the numbers of patients in groups 3 and 4 were 124 and 159, respectively (Drs. Tang, Zhang and Liang performed 44, 40 and 40 surgeries in group 3, respectively, and 59, 50 and 50 in group 4, respectively). The numbers of patients treated with IM and EM were 67 and 57 in group 3 and 96 and 63 in group 4, respectively. The mean procedure lengths in

Table 1 Difference in length of surgery by different surgeons

Variables	Length of surgery (min)				P value
	Dr. Tang	Dr. Liang	Dr. Zhang	Dr. Tao	
Group 1	57.67±16.63	64.39±14.59	62.13±17.68	56.39±17.14	0.024
IM	59.52±20.73	66.50±18.72	66.25±20.76	56.82±20.03	0.022
EM	55.91±11.71	62.38±9.17	55.94±9.17	55.71±11.91	0.213
Group 2	62.20±17.89	66.80±20.67	64.60±17.38	60.10±16.11	0.384
IM	65.00±21.93	71.45±23.71	68.83±19.24	65.00±18.13	0.508
EM	58.13±8.18	59.21±11.34	58.25±11.95	53.64±10.14	0.278
Total	60.29±17.43	64.72±19.44	64.51±16.09	58.56±16.55	0.030

EM extramedullary fixation, IM intramedullary fixation

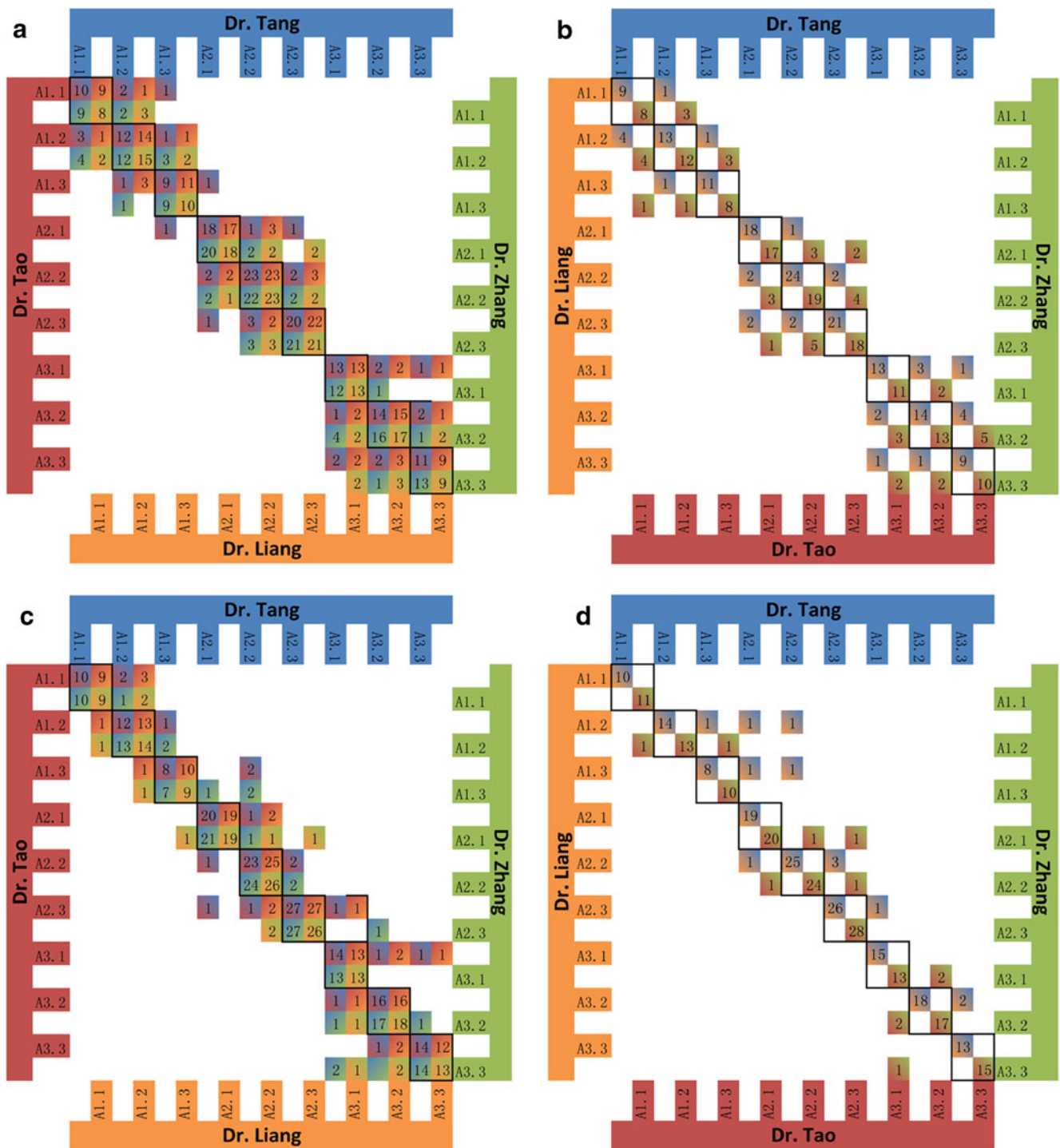


Fig. 2 Classification data B1 and data B2. **a** and **b** show data B1 based on X-ray images. **c** and **d** show data B2 based on CT/3D CT images. **a** and **c** show the number of patients in classification between Drs. Tang and Tao, Drs. Tang and Zhang, Drs. Liang and Tao, Drs. Liang and

Zhang. **b** and **d** show the number of patients in classification between Drs. Tang and Liang, Drs. Tao and Zhang. *Black square frames* are the number of patients in agreement between two surgeons

groups 3 and 4 were 61.6±16.8 min and 64.4±18.6 min (Wilcoxon test, $z=-1.058$, $P=0.290$).

For the three included surgeons, the estimates and numbers of correct determinations based on X-ray images are shown in

Table 2. The rate of correct discrimination was 83.8 % (134/160) and in each classification were A1 80.0 % (28/35), A2 85.7 % (78/91) and A3 82.4 % (28/34). The rates of misclassification in each subgroup were A1.1 16.7 % (2/12), A1.2

Table 2 Number of misclassifications based on data S and data C

Classification data (no.)	Data S (based on both X-ray and 3D CT images)										Total	
	A1.1	A1.2	A1.3	A2.1	A2.2	A2.3	A3.1	A3.2	A3.3	MC ^a		
Data C (based on X-ray images)	A1.1	10	1	1							2	12
	A1.2		11	1	1						2	13
	A1.3	1	1	7	1						3	10
	A2.1				20	3	2	1			6	26
	A2.2				1	27	2		1		4	31
	A2.3				1	1	31			1	3	34
	A3.1							8	1	1	2	10
	A3.2							1	11	1	2	13
	A3.3							1	1	9	2	11
Total		11	13	9	24	31	35	11	14	12		160

^aMC Number of misclassifications

15.4 % (2/13), A1.3 30.0 % (3/10), A2.1 23.1 % (6/26), A2.2 12.9 % (4/31), A2.3 8.8 % (3/34), A3.1 20.0 % (2/10), A3.2 15.4 % (2/13) and A3.3 18.2 % (2/11). The rates of misclassification for stable and unstable fracture were 21.3 % (13/61) and 13.1 % (13/99) ($\chi^2=1.856$, $P=0.173$, chi-square test). The rates of overestimates and underestimates were 5 % (8/160) and 11.25 % (18/160) ($\chi^2=4.186$, $P=0.041$, chi-square test). The rates of overestimates for types A1, A2 and A3 were 5.7 % (2/35), 3.3 % (3/91) and 8.82 % (3/34), respectively, and the rates of underestimates were 14.3 %, 11.0 % (10/91) and 8.8 % (3/34), respectively. The rates of overestimates in each subgroup were A1.1 0 %, A1.2 0 %, A1.3 20.0 % (2/10), A2.1 0 %, A2.2 3.23 % (1/31), A2.3 5.9 % (2/34), A3.1 0 %, A3.2 7.7 % (1/13) and A3.3 18.2 % (2/11). The rates of underestimates in each subgroup were A1.1 16.7 % (2/12), A1.2 15.4 % (2/13), A1.3 10.0 % (1/10), A2.1 23.1 % (6/26), A2.2 9.7 % (3/31), A2.3 2.9 % (1/34), A3.1 20.0 % (2/10), A3.2 7.7 % (1/13) and A3.3 0 %.

The mean lengths of surgery for classifications A1, A2 and A3 were 51.2±12.7 min, 60.1±15.7 min and 73.0±15.4 min, respectively, in group 3 (Wilcoxon test, $P<0.001$) and 53.6±18.5 min, 64.2±17.7 min and 72.5±16.5 min in group 4, respectively (Wilcoxon test, $P<0.001$). Differences in the mean length of surgery between groups 3 and 4 in the same subgroup were from less than one minute to almost six minutes (Table 3). The differences between the groups were from 0.13 to 9.5 minutes in IM and from 0.32 to 2.5 minutes in EM. The differences between procedure length and subtracted rate of overestimates from underestimates are shown in Fig. 3 with a linear regression. The correlation between total and EM patients showed no significant difference (Pearson correlation=0.347 and -0.073, $P=0.361$ and 0.852). In IM patients, there was a positive correlation with a significant difference (Pearson correlation=0.953, $P<0.001$).

Discussion

The work presented here demonstrates a different way of considering the causes of prolonged surgery. We proposed a standard classification and attempted to reveal misclassification probabilities. Then, differences in lengths of surgery according to the method of examination were compared.

After testing for consistency and differences in procedure time, three surgeons were able to classify and compare their procedure times. Inter-surgeon agreement was higher when based on CT/3D CT images (Table 1, Fig. 2), as reported previously [13]. The κ coefficient was slightly higher in this study. This difference might be due to two factors. Lack of experience was one possible factor because the participants were senior orthopaedic residents and skeletal radiologists [16]. The fact that bone fragment condition was revealed in more detail by CT plus 3D reconstruction may be another factor. Also, the previous study reported that the quality of the imaging, the modality used and the skill of the observer were important for classification [13]. The value of using CT/3D CT images for classification of fracture types has also been reported [17].

Rates of correct discrimination of the classification and the misclassification probabilities were also revealed in this study (Table 2). Overall, the rate of correct discrimination was 83.75 %, and misclassification probabilities in stable fractures were not significantly higher than for unstable fractures (21.3 vs 13.1 %, $P=0.284$). The most common misclassifications were of types A1.3 and A2.1 (30.0 and 23.1 %, respectively). The rate of underestimates was significantly higher than that of overestimates (11.3 vs 5 %, $P=0.041$). The most common underestimated classifications were subgroups A2.1, A3.1 and A1.1. We found that it took

Table 3 Length of surgery between group 3 and group 4

Classification	Group 3 (min)			Group 4 (min)			Difference (min)		
	IM	EM	Overall	IM	EM	Overall	IM	EM	Overall
A1.1	40.00±7.07	46.00±7.42	43.33±7.50	48.33±16.63	47.00±11.51	45.90±14.11	8.30	1.00	2.57
A1.2	45.00±12.25	52.50±9.35	49.50±10.66	52.50±20.45	53.75±14.93	52.86±18.47	7.50	1.25	3.36
A1.3	72.5±17.68	62.50±5.00	65.83±10.21	71.00±26.32	65.00±13.23	65.63±19.54	-1.50	2.50	-0.20
A2.1	50.71±22.99	54.50±8.96	52.94±15.72	60.36±17.37	56.25±4.43	58.86±14.05	9.64	1.75	5.92
A2.2	61.54±19.19	56.82±9.82	59.38±15.49	67.05±21.73	58.18±10.79	64.46±18.27	5.52	1.36	5.08
A2.3	68.82±15.96	58.89±8.58	65.38±15.49	74.06±22.23	58.57±10.99	67.83±19.15	5.24	-0.32	2.45
A3.1	78.00±23.61	66.67±5.77	73.75±15.04	84.29±11.34	61.00±7.42	71.67±15.42	6.29	-5.67	-2.08
A3.2	77.78±16.03	63.75±7.50	73.46±15.19	77.91±20.94	63.33±7.53	71.67±18.31	0.13	-0.42	-1.79
A3.3	78.33±15.71	64.00±8.22	71.82±14.37	76.67±17.14	63.57±8.02	74.06±16.15	-1.67	-0.43	2.24
Total	65.22±20.42	57.37±9.59	61.61±16.77	68.28±21.68	58.49±10.30	64.40±18.64	3.06	1.12	2.79

much less time for the invited experienced surgeons to reach estimates using the X-ray images.

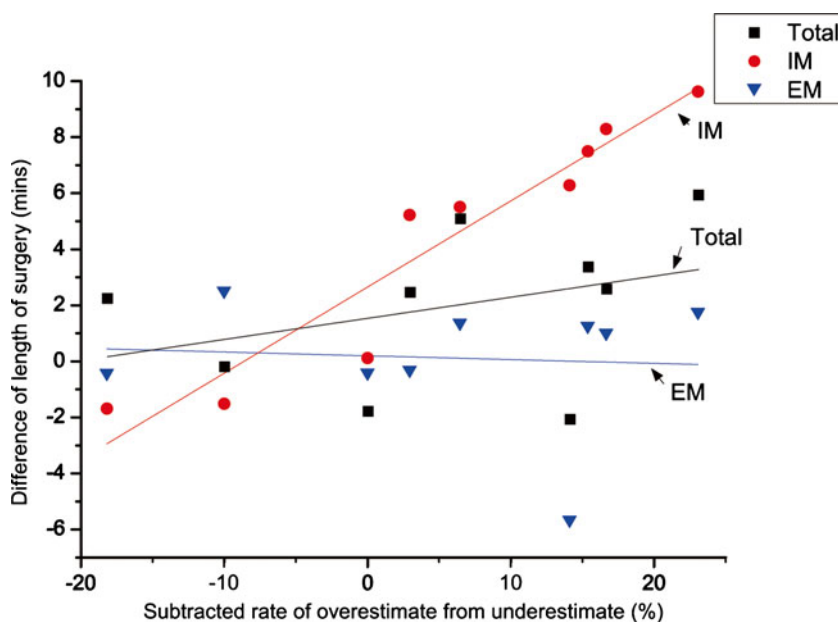
Overall, procedure length was slightly increased if no preoperative CT/3D CT examination had been conducted (Table 3). The difference was almost ten minutes for IM operations without a CT/3D CT examination, but was not significant in EM operations. This may have been because the longer incision and more exposure in EM could reduce the impact of an inadequate preoperative assessment. Also, the extra duration of 3.06 minutes in IM was due to the minimal incision and inadequate exposure, so that more time was needed for fluoroscopy to ensure that the fixation was in position. The higher the rate of overestimates subtracted from underestimates, the longer the procedure time

(Fig. 3). Particularly in the IM operation, this positive correlation was significant ($P < 0.001$).

The limitations of this study were as follows: selection bias might have been caused by the small number of surgeons who participated and the retrospective design. Secondly, the impact on procedure length was limited by the personal skills and degree of reduction by the precision surgeons.

Our data suggest that classifications based on the AO system were consistent. Preoperative assessments based on CT/3D CT examinations were more reliable and helpful. Intertrochanteric fractures were likely to be more severe than indicated by the X-ray images. The CT/3D CT examination facilitated a reduction in procedure length, especially for performance of an IM fixation.

Fig. 3 Correlation between differences of length of surgery and subtracted rate of overestimates from underestimates in the same classification. EM extramedullary fixation (blue triangles and linear fitting line). IM intramedullary fixation (red dots and linear fitting line). Total difference between group 3 and group 4 (black squares and linear fitting line)



Acknowledgments We thank Dr. Liang and Dr. Tao for participating in this study. We thank Prof. LiXing Zhu and Prof. WangLi Xu for statistical consultation and insightful suggestions on this work. We thank Xu Guo, Tao Wang and ZhengHui Feng for assisting in preparation of this manuscript.

Conflict of interest The authors declare that they have no conflict of interest.

References

- Dimai HP, Svedbom A, Fahrleitner-Pammer A, Pieber T, Resch H, Zwettler E, Chandran M, Borgström F (2011) Epidemiology of hip fractures in Austria: evidence for a change in the secular trend. *Osteoporos Int* 22:685–692. doi:10.1007/s00198-010-1271-9
- Hornby R, Evans JG, Vardon V (1989) Operative or conservative treatment for trochanteric fractures of the femur. A randomised epidemiological trial in elderly patients. *J Bone Joint Surg Br* 71:619–623
- Handoll HH, Parker MJ (2008) Conservative versus operative treatment for hip fractures in adults. *Cochrane Database Syst Rev* 3:CD000337. doi:10.1002/14651858.CD000337.pub2
- Anglen JO, Weinstein JN, American Board of Orthopaedic Surgery Research Committee (2008) Nail or plate fixation of intertrochanteric hip fractures: changing pattern of practice. A review of the American Board of Orthopaedic Surgery Database. *J Bone Joint Surg Am* 90:700–707. doi:10.2106/JBJS.G.00517
- Aros B, Tosteson AN, Gottlieb DJ, Koval KJ (2008) Is a sliding hip screw or im nail the preferred implant for intertrochanteric fracture fixation? *Clin Orthop Relat Res* 466:2827–2832. doi:10.1007/s11999-008-0285-5
- Parker MJ, Handoll HH (2005) Gamma and other cephalocondylic intramedullary nails versus extramedullary implants for extracapsular hip fractures in adults. *Cochrane Database Syst Rev* 4:CD000093. doi:10.1002/14651858.CD000093.pub3
- Kaplan K, Miyamoto R, Levine BR, Egol KA, Zuckerman JD (2008) Surgical management of hip fractures: an evidence-based review of the literature. II: intertrochanteric fractures. *J Am Acad Orthop Surg* 16:665–673
- Burstein AH (1993) Fracture classification systems: do they work and are they useful? *J Bone Joint Surg Am* 75:1743–1744
- Whitelaw GP, Segal D, Sanzone CF, Ober NS, Hadley N (1990) Unstable intertrochanteric/subtrochanteric fractures of the femur. *Clin Orthop Relat Res* 252:238–245
- Colton CL (1991) Telling the bones. *J Bone Joint Surg Br* 73:362–364
- Jensen JS (1980) Classification of trochanteric fractures. *Acta Orthop Scand* 51:803–810
- Gehrchen PM, Nielsen JO, Olesen B (1993) Poor reproducibility of Evans' classification of the trochanteric fracture. Assessment of 4 observers in 52 cases. *Acta Orthop Scand* 64:71–72
- Chapman CB, Herrera MF, Binenbaum G, Schweppe M, Staron RB, Feldman F, Rosenwasser MP (2003) Classification of intertrochanteric fractures with computed tomography: a study of intra-observer and interobserver variability and prognostic value. *Am J Orthop (Belle Mead NJ)* 32:443–449
- 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. doi:10.1016/j.injury.2009.10.007
- Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* 33:159–174
- Fung W, Jonsson A, Buhren V, Bhandari M (2007) Classifying intertrochanteric fractures of the proximal femur: does experience matter? *Med Princ Pract* 16:198–202. doi:10.1159/000100390
- Savolaine ER, Ebraheim NA, DeTroye R, Jackson WT (1992) Three-dimensional CT reconstruction for assessment of Pipkin fracture-dislocations of the hip. *Orthopedics* 15:49–51