Archives of Orthopaedic and Trauma Surgery (2019) 139:645–650
https://doi.org/10.1007/s00402-019-03123-9

TRAUMA SURGERY

The intra- and interobserver reliability of the Tile AO, the Young and Burgess, and FFP classifications in pelvic trauma

Josephine Berger-Groch[1] · Darius M. Thiesen[1] · Lars G. Grossterlinden[2] · Jan Schaewel[1] · Florian Fensky[1] · Maximilian J. Hartel[1]

Received: 30 May 2018 / Published online: 4 February 2019

Abstract

Introduction Several different systems of classification have been developed to understand the complexity of pelvic ring fractures, to facilitate communication between physicians and to support the selection of appropriate therapeutic measures. The purpose of this study was to measure the inter- and intraobserver reliability of Tile AO, Young and Burgess, and FFP classification in pelvic ring fractures. The Rommens classification system (FFP) is analyzed for the first time.

Materials and methods Four reviewers (2 × senior pelvic trauma surgeon, 1 × resident, 1 × medical student) separately analyzed and classified 154 CT scans of patients with pelvic fracture. The Tile AO, the Young and Burgess, and the FFP classifications (subgroup with patients ≥ 60 years) were compared. Another blinded re-evaluation was carried out after 2 months to determine intraobserver reliability.

Results The overall interobserver agreement was fair for all classification systems (ICC: OTA 0.55, Young and Burgess 0.42, FFP 0.54). For specific categories, (e.g. type B or C fractures), there was a substantial agreement between the experienced surgeons (kappa: OTA 0.64, Young and Burgess 0.62, FFP 0.68). For inexperienced observers, there was a fair agreement in all systems (kappa: OTA 0.23, Young and Burgess 0.23, FFP 0.36).

Conclusions All three classifications reach their maximum reliability with advanced expertise in the surgery of pelvic fractures. The novel FFP classification has proved to be at least equivalent when directly compared to the established systems. The FFP classification system showed substantial reliability in patients older than 60 years.

Keywords Pelvic trauma · Classification · Interobserver reliability · FFP · OTA · Young and Burgees

Introduction

The incidence of pelvic ring fractures ranges in age depending from 23 to 446 per 100.000 patients per year and is increasing with the aging of the population [1–3]. Pelvic ring injuries caused by high- or low-energy trauma present many problems in diagnosis, treatment and rehabilitation. As mortality rates may be as high as 13.5%, it is crucial to fully understand the complex differences between individual fracture patterns of different severity [4–7]. Therefore, several different systems of classification have been developed, to facilitate communication between physicians and to support the selection of appropriate therapeutic measures.

The Tile OTA and Young and Burgess classifications are in common clinical use. Moreover, Rommens et al. have proposed that the FFP classification should be used in geriatric trauma cases [8–11].

Josephine Berger-Groch and Darius M. Thiesen contributed equally and, therefore, share first authorship.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00402-019-03123-9) contains supplementary material, which is available to authorized users.

✉ Josephine Berger-Groch
j.berger@uke.de

[1] Department of Trauma-, Hand-, and Reconstructive Surgery, University Medical Center Hamburg-Eppendorf, Martinistr. 52, 20246 Hamburg, Germany

[2] Department of Orthopedics, Trauma-and Spine Surgery, Asklepios Hospital Hamburg-Altona, Paul-Ehrlich-Straße 1, 22763 Hamburg, Germany

While a number of publications have studied the intra- and interobserver reliability of the Tile OTA, and Young and Burgess systems, there has been no investigation of the new Rommens classification [12–14].

The aim of this study was to examine differences in inter- and intraobserver reliability between the Tile OTA, Young and Burgess, and FFP classifications. In addition, we compared the reliability with observer of different levels of experience (expert pelvic surgeon vs. resident vs. medical student).

Materials and methods

Patient population

One hundred and fifty-four cases with the diagnosis of an os sacrum fracture were randomly selected from the database of a Level I trauma facility for the period between 2004 and 2014. To prevent selection bias, no attempt was made to build representative samples for the different classification groups.

Injury classification

All CT scans were reviewed by four different observers using the PACS System (Picture Archiving and Communication System) allowing a dynamic view of the whole data set (coronal, sagittal and axial). Plain radiographs were not provided to the raters. The team had different levels of expertise ranging from high to no experience at all: 2 × specialist pelvic trauma surgeons, 1 × resident (3rd year, specializing in orthopedic trauma surgery) and 1 × medical student. All raters were provided with the same information sheets that depicted the three different classification systems (Supplement 1).

The results were entered into a blinded sheet and subsequently processed by a fifth independent person.

Since the FFP classification was developed for the elderly a subgroup analysis in patients ≥ 60 years was performed. The subgroup consisted of 97 patients.

For the OTA 61 classification, the reviewers classified the fractures according to the specific subtype, which included 26 possibilities. For the Young and Burgess system, the classifications were performed according to the eight possible categories. For the FFP classification (Rommens), 11 possibilities for the different subtypes were given (see Table 1).

The reviewers were not allowed to change their initial classification after gaining more experience during testing.

The reviewers were blinded to the patient's name, treatment, and outcome. The reviewers worked independently of each other, and no time constraints were imposed. After a 2-month period, each rater reevaluated a sample subgroup of 35 CT scans, to measure intraobserver reliability. The subgroup was chosen randomly and was the same for each reviewer.

Statistics

Differences between the patients and the three classification systems were analyzed using contingency tables and chi-square statistics. Whenever the expected numbers in cell entries were smaller than 5, Fisher's exact test was applied to calculate p values. To compare the male and female groups, descriptive statistics were presented as proportions for categorical variables and means plus standard deviations for continuous variables. All statistical analyses were performed with the SPSS statistical program 22.0 (SPSS, Chicago, IL). p values < 0.05 (two-tailed) were considered statistically significant.

We calculated Cohen's kappa coefficients with 95% confidence interval for all three classification systems. This acts as a measure of the level of agreement among the raters (see Table 2). A value of 1.0 means that there is a perfect agreement among the observers. A value of 0 suggests that the agreement was no better than chance alone. A value of less than 0 suggests greater than random disagreement [15]. The method of Landis and Koch was used to interpret the kappa coefficients [16]. As Cohen's kappa is used to evaluate the agreement between two raters, we also used the SPSS program to calculate an intraclass correlation coefficient (ICC) for an overall value (ICC interpretation: < 0.40, poor; 0.40–0.59, fair; 0.60–0.74, good; > 0.74 excellent) [17, 18].

Results

A total of 154 patients were included in the study. The mean age was 64.45 ± 19.11 years. The gender distribution was 110 women (70.5%) and 46 men (29.5%). The subgroup with patients over 60 years consisted of 85 women and 12 men. The mean age in this group was 77.32 ± 8.18 years. The underlying raw data are provided in Supplement 2.

Tile/OTA

The overall interobserver agreement for the tile OTA system was fair (0.55 ICC value, 95% CI 0.47–0.63). The two senior orthopedic surgeons were in exact agreement in 98 of 154 cases, corresponding to moderate agreement, with a kappa value of 0.52. The agreement in choosing the same main category (B/C) was substantial—with a kappa value of 0.64.

Comparison of the student to the senior pelvic trauma surgeons gave fair to moderate agreement, with kappa values of 0.31 and 0.42, for the different raters (see Table 3).

Table 1 Tile OTA classification of pelvic disruption [8], Young and Burgess classification of pelvic disruption [10], Rommens classification of pelvic disruption [11]

Classification of pelvic disruptions by the Tile OTA system:
Type A: stable
A1: fractures of the pelvis not involving the ring (A1.1; A1.2; A1.3)
A2: stable, minimally displaced fractures of the ring (A2.1; A2.2; A2.3)
A3: distal transverse fracture of the os sacrum (A3.1; A3.2; A3.3)
Type B: rotationally unstable, vertically stable
B1: open book (B1.1; B1.2)
B2: lateral compression: ipsilateral (B2.1; B2.2; B2.3)
B3: lateral compression: contralateral (bucket handle) (B3.1; B3.2; B3.3)
Type C: rotationally and vertically unstable
C1: unilateral (C1.1; C1.2; C1.3)
C2: bilateral (C2.1; C2.2; C2.3)
C3: associated with an acetabular fracture (C3.1; C3.2; C3.3)
Classification of pelvic disruptions by the Young and Burgess system
Anterior posterior compression
Type I: symphyseal diastasis—slight widening ± sacroiliac joint. Intact anterior and posterior ligaments
Type II: symphyseal diastasis—widening of sacroiliac joint, anterior ligaments disrupted, posterior ligaments intact
Type II: complete hemipelvis separation without vertical displacement. Symphyseal disruption and complete disruption of sacroiliac joint, anterior and posterior ligaments
Lateral compression
Type I: anterior transverse fracture of pubic rami plus ipsilateral sacral compression
Type II: plus—crescent (iliac wing) fracture
Type III: plus—contralateral anterior posterior compression injury
Vertical shear: vertical displacement, anterior and posterior through sacroiliac joint
Combined mechanical injuries: lateral compression/vertical shear or lateral compression/anteroposterior compression
Classification of pelvic disruption by the Rommens system (FFP)
FFP I: anterior ring fracture only
Type A: unilateral anterior lesion
Type B: bilateral anterior lesion
FFP II: non-displaced posterior lesions
Type A: non-displaced posterior lesion only
Type B: sacral crush with anterior disruption
Type C: non-displaced sacral, sacroiliac or iliac fracture with anterior disruption
FFP III: displaced unilateral posterior injury combined with an anterior pelvic ring lesion
Type A: displaced unilateral ilium fracture
Type B: displaced unilateral sacroiliac disruption
Type C: displaced unilateral sacral fracture
FFP IV: displaced bilateral posterior injuries
Type A: bilateral iliac fractures or bilateral sacroiliac disruptions
Type B: spinopelvic dissociation containing a bilateral vertical fracture through the lateral mass of the sacrum with a horizontal component connecting them
Type C: combination of different posterior instabilities

Young and Burgess

With the Young and Burgess system, interobserver reliability also gave fair overall agreement (0.42 ICC value, 95% CI 0.33–0.50). Once again, the senior surgeons showed the best agreement, with substantial kappa value of 0.6 and agreement in 111 of 154 cases. Comparison of the student to senior surgeons 1 and 2 resulted in a fair agreement with a kappa value of 0.23 (both senior surgeons). Comparison of the resident with the two senior surgeons gave fair agreement, with kappa values of 0.37 and 0.25 (see Table 4).

Table 2 Landis and Koch method of interpreting kappa coefficients [16]

Kappa value	Level of agreement
< 0.00	Poor
0.00 to 0.20	Slight
0.21 to 0.40	Fair
0.41 to 0.60	Moderate
0.61 to 0.80	Substantial
0.81 to 1.00	Excellent

Table 3 Interobserver reliability of pelvic fracture as classified by OTA classification systems from CT scans

Level of Agreement between observers for OTA61 classification for all cases			
	Observers	Agreement (%)	Kappa
Complete classification	S1 vs. S2	62	0.52
	S1 vs. RS	39	0.25
	S1 vs. STUD	46	0.31
	S2 vs. STUD	54	0.42
	S2 vs. RS	44	0.30
	RS vs. STUD	42	0.29
Four main categories	S1 vs. S2	82	0.64
	S1 vs. RS	64	0.30
	S1 vs. STUD	59	0.19
	S2 vs. STUD	68	0.37
	S2 vs. RS	70	0.39
	RS vs. STUD	63	0.23

S1 senior surgeon 1, *S2* senior surgeon 2, *RS* resident, *STUD* student

Rommens FFP

The FFP system (in patients over 60 years) gave moderate agreement between the more experienced surgeons (kappa 0.47). All other interobserver kappa values ranged between 0.18 and 0.41, corresponding to slight to moderate agreement. For the four main categories, the agreement reached a substantial level (0.68) between senior surgeons 1 and 2, and a fair level of agreement for the medical student and the resident (0.36). The overall interobserver agreement for the FFP system was moderate (0.54 ICC value, 95% CI 0.44–0.64) (see Table 5).

Intraobserver

The intraobserver examination showed weak to moderate intra-individual agreements, as Table 6 depicts.

Table 4 Interobserver reliability of pelvic fracture classification as based on the Young and Burgess classification systems from CT scans

Level of agreement between observers for the Young and Burgess classification for all cases			
	Observers	Agreement (%)	Kappa
Complete classification	S1 vs. S2	72	0.60
	S1 vs. RS	55	0.37
	S1 vs. STUD	42	0.23
	S2 vs. STUD	42	0.23
	S2 vs. RS	46	0.25
	RS vs. STUD	42	0.24
Four main categories	S1 vs. S2	75	0.62
	S1 vs. RS	58	0.35
	S1 vs. STUD	45	0.22
	S2 vs. STUD	46	0.22
	S2 vs. RS	50	0.23
	RS vs. STUD	49	0.23

S1 senior surgeon 1, *S2* senior surgeon 2, *RS* resident, *STUD* student

Table 5 Interobserver reliability of pelvic fractures based on the Rommens (FFP) classification system from CT scans

Level of Agreement between observers for FFP (Rommens) classification ≥ 60 years			
	Observers	Agreement (%)	Kappa
Complete classification	S1 vs. S2	61	0.47
	S1 vs. RS	30	0.18
	S1 vs. STUD	57	0.41
	S2 vs. STUD	45	0.31
	S2 vs. RS	32	0.19
	RS vs. STUD	39	0.24
Four main categories	S1 vs. S2	81	0.68
	S1 vs. RS	47	0.29
	S1 vs. STUD	74	0.58
	S2 vs. STUD	76	0.62
	S2 vs. RS	49	0.28
	RS vs. STUD	55	0.36

S1 senior surgeon 1, *S2* senior surgeon 2, *RS* resident, *STUD* student

Table 6 Kappa values—results of intratester analysis OTA/Young and Burgess/FFP classifications

Intratester analysis	Senior 1	Senior 2	Resident	Student
Tile/OTA	0.50	0.51	0.21	0.23
Young and Burgess	0.55	0.33	0.28	0.43
FFP ≥ 60 years	0.49	0.51	0.29	0.55

Discussion

This study shows differences between the Tile OTA, Young and Burgess and Rommens (FFP) classifications with respect to inter- and intraobserver reliability and the level of the observer's experience. Previous studies have dealt with the Tile OTA, and Young and Burgess classification systems of pelvic ring fractures, but the results have been inconsistent [12–14]. There has been no study with the FFP system. The present study included 154 2D CT scans, which makes it the largest published study.

Our study found moderate to substantial rates of agreement if the Young and Burgess classification is used by surgeons with high levels of experience. Furthermore, the current study shows a positive correlation between expertise and the level of agreement. The two senior orthopedic trauma surgeons were experienced in pelvic trauma surgery and gave the highest levels of agreement, whereas the medical student and the resident attained only fair results. We also found similar results for the Tile OTA classification, with moderate agreement between the two experts and a poorer agreement with the less experienced surgeon and the medical student.

If the classifications were simplified into the main categories, the level of agreement rose, but not to values as high as found in previous studies [13, 14]. In the Young and Burgess classification, the level of agreement increased from moderate to substantial, but only from 0.60 to 0.62 for the more experienced observers. After simplifying the Tile OTA and Rommens systems, the improvement in agreement was more marked. The high number of different subgroups in those two could make assessment more difficult.

These results support the findings of Koo et al. [14], who found a considerably lower rate of agreement for the Tile OTA classification in less experienced surgeons and an improvement when the system was combined into the main categories. In their investigation, six surgeons were classified by level of experience (two pelvis/acetabular surgeons, two orthopedic traumatologists and two senior trainees) and 30 patients with pelvic fractures were assessed related to interobserver reliability. The two pelvis/acetabular surgeons, attained a kappa value of 0.75 for tile OTA, while the two orthopedic traumatologists achieved 0.3 and the two senior trainees 0.17. In contrast to our findings, the overall agreement differed for Tile OTA with a kappa of 0.33 and for Young and Burgess with a kappa of 0.63. Our calculations of kappa are presumably more precise, as we examined 154 patients and CT scans, in contrast to the 30 patients in Koo et al.

To evaluate the results of the Rommens system regarding older patients, we performed a subanalysis in patients over 60 years. Substantial agreement was attained for the main fracture types by the two senior surgeons. The Rommens system gave the highest level of agreement (kappa = 0.36) for the unexperienced raters. In case of intraobserver agreement, the Rommens system showed the highest results for the unexperienced raters, too. This might suggest an easier understanding of this classification.

Gabbe et al. [12] found no improvement in Tile/OTA, but a discreet improvement in Young and Burgess after simplification. In their study, experienced orthopedic surgeons from three different trauma centers reviewed imaging data of 187 adult patients with a pelvic fracture, of which 117 provided 2D or 3D CT scans. In general, their results are quite different from our findings, as they found only slight agreement between Tile OTA, and Young and Burgess. Moreover, Gabbe et al. [12] concluded that the Tile OTA, and Young and Burgess classifications are unsuitable for the critical trauma patient, as their reliability is poor.

Furey et al. [13] studied inter- and intraobserver reliability of pelvic ring fractures. Their findings are in agreement with our results, as they recorded moderate agreement for Tile OTA, and Young and Burgess for five orthopedic trauma surgeons. In contrast to our results, an intratester analysis found substantial agreement for the Young and Burgess and moderate agreement for Tile OTA.

It was surprising that the medical student attained moderate intraobserver agreement in the FFP, and Young and Burgess classifications. This supports our assumption of an easier understanding of the FFP for the beginner.

The following limitations need mentioning. The observers had no access to the history of the trauma mechanism. This could have distorted the results, particularly for the Young and Burgess classification. The comparison between different levels of experience would have been more reliable with more participants in each level. Furthermore, the limitations of the methodological approach in this study need mentioning. The purpose of this study was to find out on the general applicability of all three classification systems on a patient cohort typically found in a large metropolitan area in central Europe, with a significant proportion of geriatric patients. The focus was not to determine the exact validity of each single subitem in the classifications. Therefore, no statement can be made on the classifications' exact accuracy in injury patterns that are rarely found in such cohorts.

In conclusion, all three classifications reach their maximum reliability with advanced expertise in the surgery of pelvic fractures. The novel FFP classification has proved to be at least equivalent when directly compared to the established systems. The FFP classification system showed substantial reliability in patients older than 60 years.

Funding There is no funding source.

Compliance with ethical standards

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

Ethical approval All procedures involving human participants were in accordance with the ethical standards of the institutional and national research committee (reference number: WF-009/18) and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

References

1. Buller LT, Best MJ, Quinnan SM (2016) A nationwide analysis of pelvic ring fractures: incidence and trends in treatment, length of stay, and mortality. Geriatr Orthop Surg Rehabil 7:9–17. https://doi.org/10.1177/2151458515616250
2. Pereira GJC, Damasceno ER, Dinhane DI et al (2017) Epidemiology of pelvic ring fractures and injuries. Rev Bras Ortop 52:260–269. https://doi.org/10.1016/j.rboe.2017.05.012
3. Kannus P, Parkkari J, Niemi S, Sievänen H (2015) Low-trauma pelvic fractures in elderly finns in 1970–2013. Calcif Tissue Int 97:577–580. https://doi.org/10.1007/s00223-015-0056-8
4. Krappinger D, Kammerlander C, Hak DJ, Blauth M (2010) Low-energy osteoporotic pelvic fractures. Arch Orthop Trauma Surg 130:1167–1175. https://doi.org/10.1007/s00402-010-1108-1
5. Studer P, Suhm N, Zappe B et al (2013) Pubic rami fractures in the elderly–a neglected injury? Swiss Med Wkly 143:w13859. https://doi.org/10.4414/smw.2013.13859
6. Taillandier J, Langue F, Alemanni M, Taillandier-Heriche E (2003) Mortality and functional outcomes of pelvic insufficiency fractures in older patients. Jt Bone Spine 70:287–289
7. Yang N-P, Chan C-L, Chu D et al (2014) Epidemiology of hospitalized traumatic pelvic fractures and their combined injuries in Taiwan: 2000–2011 National Health Insurance data surveillance. Biomed Res Int 2014:878601. https://doi.org/10.1155/2014/878601
8. Tile M (1996) Acute pelvic fractures: I. causation and classification. J Am Acad Orthop Surg 4:143–151
9. Marsh JL, Slongo TF, Agel J et al Fracture and dislocation classification compendium—2007: orthopaedic trauma association classification, database and outcomes committee. J Orthop Trauma 21:S1-133
10. Young JW, Burgess AR, Brumback RJ, Poka A (1986) Pelvic fractures: value of plain radiography in early assessment and management. Radiology 160:445–451. https://doi.org/10.1148/radiology.160.2.3726125
11. Rommens PM, Hofmann A (2013) Comprehensive classification of fragility fractures of the pelvic ring: recommendations for surgical treatment. Injury 44:1733–1744. https://doi.org/10.1016/j.injury.2013.06.023
12. Gabbe BJ, Esser M, Bucknill A et al (2013) The imaging and classification of severe pelvic ring fractures: experiences from two level 1 trauma centres. Bone Jt J 95–B:1396–1401. https://doi.org/10.1302/0301-620X.95B10.32134
13. Furey AJ, O'Toole RV, Nascone JW et al (2009) Classification of pelvic fractures: analysis of inter- and intraobserver variability using the Young–Burgess and tile classification systems. Orthopedics 32:401. https://doi.org/10.3928/01477447-20090511-05
14. Koo H, Leveridge M, Thompson C et al (2008) Interobserver reliability of the young–burgess and tile classification systems for fractures of the pelvic ring. J Orthop Trauma 22:379–384. https://doi.org/10.1097/BOT.0b013e31817440cf
15. Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20:37–46. https://doi.org/10.1177/001316446002000104
16. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174
17. Shrout PE, Fleiss JL (1979) Intraclass correlations: uses in assessing rater reliability. Psychol Bull 86:420–428
18. Cicchetti DV, Sparrow SA (1981) Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. Am J Ment Defic 86:127–137