A. Toom · K. Fischer · A. Märtson · L. Rips · T. Haviko

Inter-observer reliability in the assessment of heterotopic ossification: proposal of a combined classification

Received: 2 September 2004 / Accepted: 20 September 2004 / Published online: 13 April 2005
© Springer-Verlag 2005

Abstract We applied four published classifications for assessment of heterotopic ossification after total hip arthroplasty (Arcq, Brooker, DeLee and Della Valle [1, 2, 5, 6]. The average incidence of heterotopic ossification varied from 19.8% to 27.7%. The inter-observer reliability of the various classifications was determined by Cohen's Kappa statistic. Kappa values ranged from 0.897 for Arcq's to 0.814 for Brooker's classification. In order to increase the reliability and consistency, we propose a new classification system combining Brooker's and Della Valle's classifications. This new classification preserves the high reliability of Della Valle's system and is comparable to previous publications since it includes Brooker's criteria.

Résumé Nous avons utilisé quatre classifications publiées pour l'estimation des ossifications hétérotopiques après arthroplastie total de la hanche (Arcq, Brooker, DeLee et Della Valle [1, 2, 5, 6]. La fréquence moyenne de ces ossifications variait de 19,8 à 27,7%. La précision inter-observateurs des classifications a été déterminée par le kappa statistique de Cohen. Les valeurs Kappa ont varié de 0.897 pour Arcq à 0.814 pour la classification de Brooker. Pour augmenter la précision et la cohérence, nous proposons un nouveau système qui combine les classifications de Brooker et de Della Valle. Cette nouvelle classification conserve la haute précision du système de Della Valle et est comparable aux publications antérieures puisque elle inclut les critères de Brooker.

Introduction

The incidence of heterotopic ossification (HO) after total hip arthroplasty (THA) varies widely between different reports (8–90%) [7, 10]. A relevant meta-analysis reports the overall incidence of HO to be 37% [8], which is in agreement with our earlier study [12].

The most common methods for quantitative and qualitative assessment of HO in clinical practice are based on comparing the antero-posterior X-rays of the hip during the postoperative period [1, 2, 5]. Currently, the most widely used classification for assessment of HO is Brooker's system [2]. It was recently critically analysed by Della Valle and co-authors who proposed a simplified rating system [6]. They recommended better differentiation between small ossifications that hardly cause any symptoms at all and clinically more severe ossification [6]. Classification from Arcq, which is widely used in German-speaking countries [1], and a classification system published by DeLee, which takes both severity and localisation of HO [5] into account, were also included in this study.

In order to achieve a higher agreement in evaluation of morphological parameters in pathology, Cohen's Kappa coefficient has been applied to improve the criteria for developing classification systems [3, 11]. Assessment of HO after THA is such a problem since relatively low reproducibility of Brooker's classification has been reported [9, 13]. A large multi-centre study reports the inter-observer reliability based on Cohen's Kappa value to be "fair" 0.50 (95% CI 0.30–0.70) [9]. After revision in 1994, Brooker's classification achieved a better inter-observer Kappa value from 0.57 to 0.68 [13].

The aims of our study were to record the incidence of HO; to estimate the inter-observer reliability of Brooker's [13], Della Valle's, Arcq's and DeLee's classifications for HO by using Cohen's Kappa coefficient and to improve the evaluation. A combined classification is proposed in this paper.

A. Toom (✉) · A. Märtson · L. Rips · T. Haviko
Clinic of Traumatology and Orthopaedics,
University of Tartu,
Tartu, Estonia
e-mail: alar.toom@ut.ee
Tel.: +372-53401143

K. Fischer
Department of Public Health,
University of Tartu,
Tartu, Estonia

Materials and methods

Patients

One hundred eleven patients who underwent THA in the Clinic of Traumatology and Orthopaedics at the University of Tartu during 2002 were included in this study. They comprised 65 female and 46 male patients with an average age of 67.6 years and 62.7 years, respectively. In 97 cases the indication for THA was osteoarthritis.

All patients received daily post-operative prophylaxis against HO with either 200 mg celecoxib or 100 mg diclofenac for 1 week. All patients received anti-thrombotic prophylaxis with low molecular weight heparin for at least 5 days. The prostheses used were either cemented: Link Lubinus Sp II in 97 cases or non-cemented Zweymüller model in 14 cases.

Antero-posterior X-rays were taken immediately after the operation and at 12–14 months. Four observers evaluated the X-rays: two were consultant surgeons and two trainees of orthopaedic surgery. The analysis was performed single blinded to other analyses, and observers were not allowed to discuss the evaluation. This study was approved by the ethics committee at the University of Tartu.

Classification principles of tested systems

Brooker's classification divides HO into four classes: class I: islands of bone within soft tissues of any size; class II: bone spurs from pelvis or femur, leaving at least 1 cm between opposing bone surfaces; class III: same as II but reducing the space between opposing bone surfaces to less than 1 cm; and class IV: ankylosis [2].

Della Valle's classification divides patients into three groups: A: HO is absent, or new bone islands less than 1 cm in diameter are present; B: isolated ossification with a diameter at least 1 cm and marginal ossification leaving at least 1 cm distance between pelvis and femur; and C: marginal ossification leaving less than 1 cm distance between pelvis and femur or ankylosis [6].

Arcq's classification also does not specify the minimum size of ossifications. Class I refers to isolated or marginal ossifications, which do "bridge" the opposing bone surfaces. Class II refers to the bony bridging on one side of the implant and class III on both sides of the implant [1].

Table 1 Incidence of heterotopic ossification (percentage) using different classifications

System	Brooker	Della Valle	Arcq	DeLee subclasses
Observer 1	27.9	21.6	27.9	24.3
Observer 2	24.3	12.6	24.3	19.8
Observer 3	28.8	20.7	28.8	26.1
Observer 4	29.7	24.3	29.7	27.9
Average	27.7 ^a	19.8 ^{a,b}	27.7 ^b	24.5

^{a,b}Indicate the statistical significances.

Table 2 Kappa values of different classification systems

	System	Brooker	Della Valle	Arcq	DeLee subclasses
<i>n</i> =111	Kappa value ^a	0.814	0.871	0.897	0.862
	95% CI	0.72–0.91	0.81–0.93	0.76–0.96	0.80–0.99

^aKappa values based on all measurements.

DeLee's classification is more complicated: Class I: isolated ossifications in the soft tissues exceeding 5 mm; classes II and III: presence of ossifications on the lateral or medial side of the implant, respectively. Additionally, letters A and B, which refer to a localisation of the ossification on the pelvis or femur, respectively, are added. Arabic numbers added mark the sub-class of severity where "1" refers to extension of HO less than 50% of the pelvic-femoral distance, "2" refers to more than 50% and "3" refers to complete bridging [5].

Statistical analyses The reliability of different classifications was assessed using the proportion of the agreeing observer pairs to the maximum of possible agreeing pairs using Cohen's Kappa coefficient, which is a measure of the chance-corrected agreement [3] and, therefore, reflects the reproducibility of different classifications. Statistical software package R, version 1.9.0 for Windows [4] and its contribution package "concord" were used for statistical calculations. Statistical significance was assessed at the conventional 5% level.

Results

The incidence of HO calculated using different classifications is presented in Table 1. The Kappa values of the different classifications are presented in Table 2. The fact that Brooker's classification has five classes instead of Della Valle's three and Arcq's four, may partly explain the lower rates of agreement and reproducibility. In order to eliminate this influence, we performed an additional analysis with the Brooker classes I and II together and III and IV together. This analysis returned the Kappa value for all groups as high as 0.881.

Discussion

Our results demonstrate that the incidence of HO varies considerably depending on the rating system used (Table 1). This has to be considered if studies are compared. Compared to the previous studies [6, 13] our Kappa values are relatively high. One possible reason for the apparent increase of inter-observer reliability is the fact that analyses with multiple classifications were performed sequentially on the same patients. Our data demonstrate that the most widely used Brooker's classification does result in lower inter-observer reliability than other systems assessed. The

Table 3 Comparison of the first and second levels of the proposed classification to other classifications studied

Proposed classification		Brooker's system	Della Valle's system	Arcq's system	DeLee's system
Level 1	Level 2				
A. Ossification absent or small	A0. Absence the ossifications	0	A	0	0
	A1. Isolated ossifications less than 1 cm in length	I	A	1	0 or 1
B. Ossifications leaving MORE than 1 cm distance between pelvis and femur	B1. Isolated ossifications at least 1 cm in length	I	B	1	1 or 2
	B2. Marginal ossifications	II	B	1	1 or 2
C. Ossifications leaving LESS than 1 cm distance between pelvis and femur or ankylosis	C1. Isolated ossifications at least 1 cm in length	I	B	1	1 or 2
	C2. Marginal ossifications	III	C	1	1 or 2
	C3. Ankylosis	IV	C	2 or 3	3

other classifications have very similar values of the Kappa statistic.

Although we were able to increase the reliability of Brooker's classification by limiting the number of classes, this action is arbitrary and may diminish the precision of this system. Furthermore, this action does not eliminate two problems mentioned by Wright [13]. The most common of these deals with isolated ossification having such a large size that it occupies most of the distance between the opposing bone surfaces. In our study, there was a similar problem in 5.9% cases (4–9 cases recorded by different observers). Such ossifications may potentially cause clinical symptoms, but Brooker's system classifies them as class I. Della Valle's system considers the size of such ossifications and places them in intermediate class (B).

Since Brooker's classification is currently the most widely used, we decided to create a new system, which is based on any of the other systems showing higher reliability than the Brooker's classification, so that the result will be comparable to the previous studies.

Arcq's classification showed the highest proportion of agreement, but lacked precision. Almost all cases of HO in our study group were graded "1", which partly accounts for the high Kappa value. DeLee's system grades the extension of HO based on a different principle. Therefore, we decided to combine Brooker's classification with Della Valle's system to provide adequate inter-observer reproducibility and improve consistency in classification of significant HO [6]. Della Valle's system is basically not comparable to previous studies because of its different criteria but is easily convertible to both of its parent systems.

This classification considers three levels. The first resembles the system of Della Valle using classes A, B and C. Reproducibility of this part of the classification is the same as in Della Valle's system.

The second level integrates Brooker's system, and we suggest using this level for clinical work and research. Numbers 0–3 are added to the letters so "0" marks the complete absence of ossification, "1" refers to ossifications located in soft tissues, "2" refers to marginal ossifications, and "3" indicates the presence of complete bridging, which clinically and radiologically indicates fibrous or bony ankylosis.

This level of classification solves the problem of extensive, isolated ossifications, allocating them between classes B1 and C1 on the same basis as marginal ossification. They are thus divided on the basis of the overall free distance left between the pelvis and femur. The first and second levels of the classification are presented and compared to the other antero-posterior rating systems in Table 3.

The third level of the new classification determines the localisation of the ossification. This makes our system comparable to the parts of DeLee's classification dealing with the localisation of HO. Briefly, capital letters "L" or "M" are attached to the numbers indicating ossifications on the lateral or medial side of the implant. In addition, the number "2" is supplemented by lowercase letters "p" or "f", referring to the pelvic or femoral localisation, respectively.

Our analysis revealed that the incidence of HO after THA is dependent to a large extent, on the classification system used. This has to be considered when comparing the incidence of HO in the literature.

We propose a combined classification system providing easier clinical assessment and higher inter-observer reliability. The classification is convertible to previous studies using Brooker's and Della Valle's classifications and partly convertible to previous studies using Arcq's and DeLee's systems.

Acknowledgements This work was supported by grants from the Estonian Science Foundation: no. 5210 and no. 5445, and the Swedish Institute (01829).

References

1. Arcq M (1973) Die paraartikulären Ossifikationen—eine Komplikation der Totalhüftendoprothese des Hüftgelenks. *Arch Orthop Unfall Chir* 77:108–131
2. Brooker AF, Bowermann JB, Robinson RA, Riley LH Jr (1973) Ectopic ossification following total hip replacement: incidence and method of classification. *J Bone Joint Surg Am* 55:1629–1632
3. Cohen J (1960) A coefficient of agreement for nominal scales. *Educ Psychol Meas* 20:37–46
4. Dalggaard P (2002) *Introductory statistics with package R*. Springer, Berlin Heidelberg New York

5. DeLee J, Ferrari A, Charnley J (1976) Ectopic bone formation following low friction arthroplasty of the hip. *Clin Orthop* 121: 53–59
6. Della Valle AG, Ruzo PS, Pavone V, Tolo E, Mintz DN, Salvati EA (2002) Heterotopic ossification after total hip arthroplasty. A critical analysis of the Brooker classification and proposal of a simplified rating system. *J Arthroplasty* 17:870–875
7. Lazansky MG (1973) Complications revisited: the debit side of total hip replacement. *Clin Orthop* 95:96–103
8. Neal BC, Rodgers A, Clark T, Gray H, Reid IR, Dunn L et al (2000) A systematic survey of 13 randomized trials of non-steroidal anti-inflammatory drugs for the prevention of heterotopic bone formation after major hip surgery. *Acta Orthop Scand* 71:122–128
9. Neal BC, Rodgers A, Gray H, Clark T, Beaumont DD, House T et al (2000) No effect of low-dose aspirin for the prevention of heterotopic bone formation after total hip arthroplasty. *Acta Orthop Scand* 71:129–134
10. Puzas JE, Miller MD, Rosier RN (1989) Pathologic bone formation. *Clin Orthop* 245:269–281
11. Svanholm H, Starklint H, Gundersen HJG, Fabricius J, Barlebo H, Olsen S (1989) Reproducibility of histomorphologic diagnoses with special reference to the kappa statistic. *Acta Pathol Microbiol Immunol Scand* 97:689–698
12. Toom A, Haviko T, Rips L (2001) Heterotopic ossification after hip arthroplasty. *Int Orthop* 24:323–326
13. Wright JG, Moran E, Bogoch E (1994) Reliability and validity of the grading of heterotopic ossification. *J Arthroplasty* 9:549–553