Clin Orthop Relat Res (2009) 467:666–675
DOI 10.1007/s11999-008-0626-4

SYMPOSIUM: FEMOROACETABULAR IMPINGEMENT: CURRENT STATUS OF DIAGNOSIS AND TREATMENT

Radiographic Evaluation of the Hip has Limited Reliability

John C. Clohisy MD, John C. Carlisle MD, Robert Trousdale MD, Young-Jo Kim MD, PhD, Paul E. Beaule MD, FRCSC, Patrick Morgan MD, Karen Steger-May MA, Perry L. Schoenecker MD, Michael Millis MD

Published online: 2 December 2008

Abstract Radiographic evaluation provides essential information regarding the diagnosis and treatment of musculoskeletal disorders. We evaluated the ability of hip specialists to reliably identify important radiographic features and to make a diagnosis based on plain radiographs alone. Five hip specialists and one fellow performed a blinded radiographic review of 25 control hips, 25 hips with developmental dysplasia (DDH), and 27 with femoroacetabular impingement (FAI). On two separate occasions, readers assessed acetabular version, inclination and depth, position of the femoral head center, head sphericity, head-neck offset, Tönnis grade, and joint congruency. Observers made a diagnosis categorizing each hip as normal, dysplastic, FAI, or combined DDH and FAI (features of both). Reliability was determined using Cohen's kappa coefficient. Intraobserver values were highest for acetabular inclination ($\kappa = 0.72$) and determination of femoral head center position ($\kappa = 0.77$). Interobserver reliability values were highest for acetabular inclination ($\kappa = 0.61$) and Tönnis osteoarthritis grade ($\kappa = 0.59$). All other measurements, including diagnosis, had kappa values less than 0.55. We concluded many of the standard radiographic parameters used to diagnose DDH and/or FAI are not reproducible. Accordingly, a more clear set of definitions and measurements must be developed to allow for more reliable diagnosis of early hip disease.
Level of Evidence: Level III, diagnostic study. See the guidelines for authors for a complete description of the levels of evidence.

This work was supported in part by Award Number UL1RR024992 from the National Center for Research Resources (JCC). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Center for Research Resources or the National Institutes of Health. This work was also supported in part by the Curing Hip Disease Fund (JCC).
Each author certifies that his or her institution has approved the human protocol for this investigation, that all investigations were conducted in conformity with ethical principles of research, and that informed consent for participation in the study was obtained.

J. C. Clohisy (✉), J. C. Carlisle
Department of Orthopaedic Surgery,
Washington University School of Medicine,
One Barnes-Hospital Plaza, Suite 11300 West Pavilion,
Campus Box 8233, St Louis, MO 63110, USA
e-mail: clohisyj@wudosis.wustl.edu

R. Trousdale
The Mayo Clinic, Rochester, MN, USA

Y.-J. Kim, M. Millis
Adolescent/Young Adult Hip Unit,
Department of Orthopaedic Surgery,
Children's Hospital Boston, Boston, MA, USA

P. E. Beaule
Ottawa General Hospital, Ottawa, ON, Canada

P. Morgan
Department of Orthopedic Surgery,
University of Minnesota Medical School,
Minneapolis, MN, USA

K. Steger-May
Division of Biostatistics,
Washington University School of Medicine,
St Louis, MO, USA

P. L. Schoenecker
Shriner's Hospital for Children, St Louis, MO, USA

Introduction

Noninflammatory hip disease in the adolescent and young adult patient population commonly results from one of two conditions: hip instability or femoroacetabular impingement (FAI) [6, 13]. Each of these conditions occurs with a spectrum of severity and may coexist in the same hip. Hip instability often results from acetabular dysplasia, which is characterized by insufficient anterolateral femoral head coverage and superolateral inclination of the acetabular articular surface [32, 33]. Anterolateral acetabular rim overload, instability, and excessive shear stresses lead to joint degeneration [21]. FAI is a condition of abnormal contact or abutment between the proximal femur and acetabular rim, often secondary to excessive femoral head coverage and/or insufficient femoral head-neck offset. Abnormal rim loading or shear stresses from either or both of these mechanisms mediates progressive articular cartilage and labral disease [2, 10, 17]. Clinical symptoms associated with structural instability and FAI can have common characteristics, the so-called "acetabular rim syndrome" [16]. Therefore, making an accurate diagnosis can be challenging.

A definitive diagnosis should be based on a careful synthesis of detailed history, physical examination, and appropriate imaging. Plain radiographs remain the cornerstone of initial diagnosis of structural hip disease, although MRI and computed tomography are often useful in confirming the precise diagnosis. The process of obtaining quality radiographic views and subsequently interpreting those radiographs in an accurate fashion is extremely important in establishing a correct diagnosis [7, 19, 20]. Not unlike fracture classification, this process must permit the physician to choose an appropriate method of treatment and to provide a reasonably precise estimation of the outcome of that treatment [1]. For this to be accomplished, techniques of image interpretation must be functional, reproducing the same desired results time after time in the hands of multiple users. Consequently, reliable radiographic parameters of hip structural anatomy are needed for effective patient evaluation, the development of treatment algorithms, and multicenter clinical research initiatives. While radiographic classification systems and numerous radiographic parameters have been presented in the pediatric literature [3, 4, 14, 15, 24, 27], the reliability of these parameters in evaluating the skeletally mature hip remains unclear. Several authors have examined the interobserver and intraobserver reliability of selected measurements as an isolated part of a larger study [7, 9, 11, 19, 22, 24], however, few comprehensive studies report the reliability of multiple measurements in the adult literature [29, 30], and none (to our knowledge) have linked those measurements to a radiographic diagnosis.

We therefore analyzed the intraobserver and interobserver reliability of hip surgeons in evaluating radiographic parameters of hip structural anatomy. Second, we tested the agreement of readers in establishing a radiograph-based diagnosis. We presumed a surgeon study group could establish a panel of radiographic observations that provides good to excellent reliability in evaluating prearthritic hip conditions.

Materials and Methods

We used the computerized patient database of the senior author (JCC) to select preoperative images for 27 patients with acetabular dysplasia, 25 patients with cam, pincer, or combined impingement, and a control group of 25 patients with asymptomatic hips. All patients in the dysplasia and impingement groups were randomly selected from an alphabetized list of consecutive cases starting in 2001. All patients in the dysplasia group underwent periacetabular osteotomy, and all patients in the impingement group underwent either surgical dislocation with osteochondroplasty or hip arthroscopy with combined limited open osteochondroplasty. The control group consisted of a patient cohort that has previously been reported [25]. These control patients were evaluated clinically and radiographically by the senior author and were found to have no clinical evidence of hip disease. None of the patients had groin pain, a positive impingement test, or hip irritability on examination. All had signs and symptoms consistent with a disorder not involving the hip (for example, a lumbar spine disorder). We excluded cases from the control group if there was radiographic evidence of previous hip surgery or if the radiographs were taken before 2001, as images prior to this date did not meet the Digital Imaging and Communications in Medicine (DICOM) standard. We had prior Institutional Review Board approval for the protocol.

The radiographs were retrospectively reviewed by six orthopaedic surgeons (RT, YJK, PEB, PM, PLS, MM) with a primary interest in prearthritic hip disease and joint preservation surgery. Five of the observers had extensive experience (more than 5 years of practice) with the diagnosis and treatment of hip dysplasia and FAI. The sixth reader (PM) was a fellow with specific interest in hip surgery. Before the initial radiographic review, the Academic Network for Conservational Hip Outcomes Research (ANCHOR) Study Group agreed on a standardized set of criteria to be used for evaluating radiographic anatomy of the hip. In open discussion, all the observers in this study defined and agreed on an approach to the assessment of acetabular depth, inclination of the weight-bearing dome of the acetabulum, position of the center of the femoral head

relative to the acetabulum, sphericity of the femoral head, appearance of the femoral head-neck junction, congruency of the femoral head and acetabulum, overall degree of hip osteoarthritic change based on the Tönnis classification, and quality of the anteroposterior (AP) pelvis radiograph. These definitions were summarized in written form and were available to the readers as a reference. Readers were instructed to record their radiographic observations but were not required to make specific, detailed measurements. Using all radiographic views, this represented 14 distinct observations (Table 1). The specific method for evaluating each of these structural features included the following:

(1) Acetabular depth: Using an AP pelvis radiograph, the relationship of the floor of the fossa acetabuli and the femoral head was evaluated relative to the ilioischial line. Hips were classified as "profunda" if the floor of the fossa acetabuli touched or was medial to the ilioischial line and "protrusio" if the medial edge of the femoral head was medial to the ilioischial line. All hips that did not meet these criteria were assigned to a catch-all group and classified as "not deep." Hips with findings of either profunda or protrusion were considered at risk for pincer impingement.

(2) Acetabular inclination: Using an AP pelvis radiograph, acetabular inclination was classified as normal, increased, or decreased based on the degree of the Tönnis angle. This measurement was defined as follows. Three lines were drawn on the AP radiograph: (1) a horizontal line connecting the base of the acetabular teardrops; (2) a horizontal line parallel to Line 1 running through the most inferior point of the sclerotic acetabular sourcil (Point I); and (3) a line extending from Point I to Point L at the lateral margin of the acetabular sourcil (the sclerotic weight-bearing portion of the acetabulum). The Tönnis angle is formed by the intersection of Lines 2 and 3. Acetabuli having a Tönnis angle of 0° to 10° were considered normal, whereas those having an angle greater than 10° or less than 0° were considered increased and decreased, respectively. Hips with an increased Tönnis angle were considered to be at risk for structural instability, whereas those having a decreased inclination were considered at risk for pincer impingement.

(3) Acetabular version: Using an AP pelvis radiograph, acetabuli were classified as retroverted or anteverted based on the presence or absence of a "crossover sign." [26] Hips were considered anteverted if the anterior wall did not cross the posterior wall of the acetabulum before reaching the lateral aspect of the sourcil and retroverted if the anterior wall did cross the posterior wall of the acetabulum before reaching the lateral edge of the sourcil. Observers were instructed to make this assessment exclusively based on the presence or absence of the crossover sign, ignoring the element of error potentially introduced by excessive pelvic tilt and/or malrotation. Retroverted hips were considered at risk for pincer impingement.

(4) Hip center: Using an AP pelvis radiograph, observers classified the position of the hip center as lateralized or not lateralized based on the position of the medial aspect of the femoral head relative to the ilioischial line. The hip center was considered lateralized if the medial aspect of the femoral head was greater than

Table 1. Intrarater and interrater reliability in the assessment of structural features of the young adult hip*

Structural feature	Combined intraobserver reliability	Interobserver reliability
Acetabular version	K = 0.46 (95% CI: 0.37–0.55)	K = 0.39
Acetabular inclination	K = 0.73 (95% CI: 0.68–0.79)	K = 0.64
Acetabular depth	K = 0.61 (95% CI: 0.54–0.68)	K = 0.39
Position of head center	K = 0.76 (95% CI: 0.69–0.83)	K = 0.52
Head sphericity (AP)	K = 0.56 (95% CI: 0.48–0.63)	K = 0.46
Head sphericity (frog-lateral)	K = 0.60 (95% CI: 0.52–0.68)	K = 0.44
Head sphericity (crosstable)	K = 0.55 (95% CI: 0.45–0.64)	K = 0.41
Head-neck offset (AP)	K = 0.43 (95% CI: 0.36–0.50)	K = 0.24
Head-neck offset (frog-lateral)	K = 0.55 (95% CI: 0.48–0.62)	K = 0.19
Head-neck offset (crosstable)	K = 0.30 (95% CI: 0.23–0.37)	K = 0.22
Joint congruency	K = 0.50 (95% CI: 0.41–0.59)	K = 0.29
Pelvic tilt	K = 0.55 (95% CI: 0.47–0.63)	K = 0.37
Pelvic rotation	K = 0.57 (95% CI: 0.50–0.65)	K = 0.21
Tönnis osteoarthritis grade	K = 0.60 (95% CI: 0.54–0.66)	K = 0.59

* K = kappa value; AP = anteroposterior; CI = confidence interval.

10 mm from the ilioischial line and not lateralized if the medial aspect of the femoral head was less than 10 mm from the ilioischial line. Lateralized femoral heads were considered to be a sign of structural instability or dysplasia.

(5) Head sphericity: Using AP pelvis, frog-lateral, and crosstable lateral radiographs, the femoral head was classified as either spherical or aspherical using Mose templates as a reference (if desired by the observer). As a rudimentary guideline for determining asphericity in more subtle cases, it was agreed on that if the femoral epiphysis extended beyond the margin of the reference circle by 2 mm or more, the femoral head was considered aspherical. If the femoral head epiphysis did not extend beyond the Mose template by more than 2 mm, it was considered spherical. Hips with an aspherical head were considered at risk for impingement.

(6) Head-neck offset/junction: Using AP pelvis, frog-lateral, and crosstable lateral radiographs, the anterior femoral head-neck junction was classified in relation to the posterior femoral head-neck junction based on the gross appearance of the radius of curvature at each location. If the anterior and posterior concavities were grossly symmetric, the head-neck junction was defined as having symmetric concavity. Conversely, if the concavity at the anterior head-neck junction had a radius of curvature that was greater (less head-neck offset) than that of the posterior head-neck junction, the hip was considered to have a moderate decrease in head-neck offset. Lastly, if the anterior head-neck junction had a convexity, as opposed to a concavity, the head-neck junction was considered to have a prominence. Hips with decreased offset or a prominence were considered to be at risk for cam impingement.

(7) Tönnis grade: Using all four radiographic views, observers determined the degree of osteoarthritis present in each hip using the Tönnis classification system. As defined by Tönnis [32], grades of osteoarthritis range from 0 to 3 defined as: Grade 0, no signs of osteoarthritis; Grade 1, increased sclerosis of the head and acetabulum, slight joint space narrowing, and slight lipping at the joint margins; Grade 2, small cysts in the head or acetabulum, moderate joint space narrowing, and moderate loss of sphericity of the head; or Grade 3, large cysts in the head or acetabulum, joint space obliteration or severe joint space narrowing, severe deformity of the femoral head, or evidence of necrosis.

(8) Congruency: Using an AP pelvis, observers classified each hip as congruous or incongruous based on a subjective assessment of the degree of conformity between the femoral head and acetabulum. Hips were considered congruous if the arc of the head matched the arc of the acetabulum and incongruous if the two surfaces did not grossly match. Incongruent hips could variably be suggestive of dysplasia or impingement.

(9) Pelvic tilt/rotation: Using the AP pelvis radiograph, observers determined the quality of the radiographic images in relation to the tilt and rotation of the pelvis. Rotation was considered "perfect" if the obturator foramina were symmetric and imperfect if the obturator foramina were asymmetric. Likewise, tilt was considered "perfect" if the distance from the tip of the coccyx to the superior aspect of the symphysis pubis was between 1 cm and 3 cm and imperfect if the distance was less than 1 cm or larger than 3 cm.

All study patients had a series of four radiographs: an AP pelvis, crosstable, and frog-laterals and a false profile view of the hip. Though differing radiology technicians obtained the image sets, all radiographs were taken in the same department, at the same hospital, using the same imaging protocol. The AP pelvis radiograph was performed with the patient supine on the x-ray table with both lower extremities oriented in 15° of internal rotation in order to maximize the length of the femoral neck. The crosstable lateral radiograph was performed with the patient supine on the x-ray table with the contralateral hip and knee flexed beyond 80° and the symptomatic leg internally rotated 15° to expose the anterolateral surface of the femoral head-neck junction [9]. The frog leg lateral radiograph of the hip was obtained with the patient positioned supine on the x-ray table with their affected leg flexed at the knee approximately 30° to 40°, the hip abducted 45°, and the heel of the affected leg resting against the medial aspect of the contralateral knee. The false profile view was taken with the patient standing, with the affected hip against the cassette, and the pelvis rotated 65° in relation to the wall bucky [18].

All identifying data were removed from the radiographs, and each patient was assigned a study number. Digital images were accessed from the Washington University hard drive for the initial viewing and, after shuffling the viewing order, were mailed to authors on CD for the second reading. After reviewing the entire radiographic series for each patient, observers were asked to make a diagnosis based on their findings. The possible diagnostic categories were predetermined and included the following four groups: normal (no radiographic evidence of deformity), pathoanatomy consistent with acetabular dysplasia, FAI, or combined (acetabular dysplasia or structural instability with features capable of FAI).

All observers in the study were blinded with regard to the patient's identity, history, physical examination,

underlying diagnosis, treatments, and outcomes. Each observer performed the radiograph review independently and was blinded to the other participants' diagnoses and assessments of structural anatomy. To assess intraobserver reliability, a second review of the image sets took place 6 weeks after the first. Although the image sets remained the same, the sequence of images was randomly altered and each image set was labeled with a new identification number. Observers were not allowed access to the images between sessions and were asked not to discuss their data with other study participants between readings. Additionally, observers were not allowed access to their previous results at the time of the second review.

Before initiating the study, we employed a biostatistician to determine the appropriate number of cases needed for each group to determine statistical significance in regard to intraobserver and interobserver reliability in determination of each structural feature and the radiographic diagnosis. As the kappa coefficients in this study were not assessed for statistical significance, no formal sample size calculation was performed. A minimum sample size of 25 per group was determined to provide adequate spread across the rating categories and thus provide adequately precise confidence intervals for the kappa estimates.

We calculated kappa values [8] for intraobserver reliability, measuring the agreement of first and second readings for a given observer and for all readers combined (combined intraobserver reliability). Kappa values were also calculated to determine interobserver reliability of the first reading across all observers. Because of the potential for bias and practice effects, we did not calculate interobserver reliability using second readings. The simple kappa was reported for unordered variables. The weighted kappa is a refinement of the kappa coefficient that takes into account the magnitude of the disagreement between ratings and was used for Tönnis grade as a result of the ordering of the variable. Kappa values of 1.0 are indicative of perfect agreement, whereas values less than 1.0 suggest progressively less agreement between readers (in the case of interobserver reliability) or between reads (in the case of intraobserver reliability).

We calculated reliability measurements for diagnosis in two ways: (1) an initial analysis using the four previously described diagnostic categories (normal, dysplastic, impingement, or combined); and (2) a secondary analysis, which grouped the "combined" cases with both the impingement cases and the dysplasia cases. This created the following three groups for determination of diagnostic reliability: normal, dysplasia or combined, and impingement or combined. The purpose of restructuring the categories in this way was to assess the ability of readers to reliably identify structural features of dysplasia or impingement regardless of whether these features coexisted in the same patient in a "combined" fashion.

Results

The majority of the radiographic features analyzed had subjectively poor combined intraobserver and interobserver reliability (Table 1). Of the 14 structural features reviewed, only determinations of "acetabular inclination" and "position of the femoral head" demonstrated combined intraobserver reliability having a kappa value of 0.7 or greater. Six of the six readers had intraobserver reliability with kappa values greater than 0.5 for these two parameters (Table 2). A combined intraobserver agreement having a kappa value greater than 0.64 was not achieved for any of the structural features. Interobserver testing identified acetabular inclination, position of head center, and Tönnis osteoarthritis grade as having kappa values greater than 0.5. The remaining 12 structural features demonstrated interobserver kappa values less than 0.5.

Overall, the combined intraobserver agreement in making a radiographic diagnosis demonstrated kappa values above 0.55, although interobserver agreement was subjectively poor (Table 3). Nevertheless, study participants demonstrated improved intraobserver and interobserver reliability ($\kappa = 0.82$ and 0.80, respectively) in making the diagnosis of dysplasia when the individual

Table 2. Number of readers with individual intrarater reliability in making a radiographic assessment of various structural features of the hip with a kappa value > 0.5

Structural feature	Number of readers with intrarater reliability $\kappa \geq 0.5$	Number of readers with intrarater reliability $\kappa < 0.5$
Acetabular version	3	3
Acetabular inclination	6	0
Acetabular depth	5	1
Position of head center	6	0
Head sphericity (AP)	4	2
Head sphericity (frog-lateral)	6	0
Head sphericity (crosstable)	3	3
Head-neck offset (AP)	2	4
Head-neck offset (frog-lateral)	3	3
Head-neck offset (crosstable)	0	6
Joint congruency	3	3
Film quality assessment (tilt)	3	3
Film quality assessment (rotation)	4	2
Tönnis osteoarthritis grade	3	3

AP = anteroposterior.

Table 3. Intrarater and interrater reliability in making a radiographic diagnosis

Structural feature	Combined intrarater reliability	Interrater reliability
Diagnosis	K = 0.61 (95% CI: 0.56–0.67)	K = 0.54
Diagnosis (dysplasia or combined)	K = 0.82 (95% CI: 0.77–0.88)	K = 0.80
Diagnosis (FAI or combined)	K = 0.56 (95% CI: 0.48–0.65)	K = 0.46

FAI = femoroacetabular impingement; CI = confidence interval.

Table 4. Number of readers with excellent, good, or poor reliability in making a radiographic diagnosis of structural deformity in the hip

Structural feature	Number of readers with poor intrarater reliability	Number of readers with good or excellent reliability
Diagnosis	5	1
Diagnosis (dysplasia or combined)	0	6
Diagnosis (FAI or combined)	6	0

FAI = femoroacetabular impingement.

diagnostic categories of "dysplasia" and "combined" were grouped together (Table 4). On the contrary, the reviewers did not demonstrate improved agreement in making the diagnosis of impingement when the individual categories of "impingement" and "combined" were grouped together. The 25 cases classified as "normal" were a major source of disagreement between the observers (Table 5). The number of cases correctly identified as normal ranged from 0% to 84% with an average of 36%. Thus, the majority (64%) of asymptomatic hips were assigned a radiographic, structural abnormality.

Discussion

Radiographs provide essential information to diagnose and treat musculoskeletal disorders. However, while radiographic classification systems and numerous radiographic parameters have been reported in the pediatric literature [3, 4, 14, 15, 24, 27], their reliability in the skeletally mature hip remains unclear. We therefore determined the intra- and interobserver reliability of a collection of radiographic parameters and structural features that have been commonly used to diagnose structural instability and impingement of the hip. In particular, we sought to assess the ability of both individual readers and a group of readers to reproducibly evaluate 14 elements on plain radiographs. In addition, we questioned whether or not those 14 elements could be synthesized to create a reproducible radiographic diagnosis.

We note several limitations to the study. First, the observers in this study had no clinical information regarding the cases, and this likely detracted from the reliability of determining clinically major structural abnormalities and deriving the diagnosis. Certainly, the radiographic findings could not be put into a clinical context making the diagnostic portion of the study difficult. In fact, despite the poor intraobserver results, the authors of

Table 5. Number of cases identified as normal

Reader	Number of cases identified as 'normal'	Cases identified as normal/actual number of normal cases (N = 25)	Cases identified as normal/total number of cases (N = 77)
Reader 1, Read 1	7	28%	9%
Reader 1, Read 2	21	84%	27%
Reader 2, Read 1	18	72%	23%
Reader 2, Read 2	16	64%	21%
Reader 3, Read 1	7	28%	9%
Reader 3, Read 2	4	16%	5%
Reader 4, Read 1	8	32%	10%
Reader 4, Read 2	5	20%	6%
Reader 5, Read 1	10	40%	13%
Reader 5, Read 2	11	44%	14%
Reader 6, Read 1	3	12%	4%
Reader 6, Read 2	0	0%	0%
Average	9	36%	12%

Table 6. Summary of previously published plain radiograph reliability studies in the adult population

Study	Number of readers	Structural feature	Interobserver reliability	Intraobserver reliability
Gosvig et al. [11]	2	Alpha angle	ICC = 0.83	ICC = 0.90–0.96
Clohisy et al. [7]	2	Alpha angle	K = 0.56–0.85	K = 0.5–0.73
		Head sphericity	K = 0.66–0.82	K = 0.98–0.99
		Head-neck offset	K = 0.52	K = 0.63–0.74
Meyer et al. [19]	2	Alpha angle	R = 0.88	R = 0.95
Nelitz et al. [22]	3	Acetabular index	ICC = 0.85	ICC = 0.86–0.89
		Lateral subluxation of the hip	ICC = 0.80	ICC = 0.85–0.90
		Acetabular index of depth to width	ICC = 0.61	ICC = 0.65–0.69
Tannast et al. [30]	2	Acetabular inclination	K = 0.61	K = 0.74–0.89
		Cross-over sign	K = 0.60	K = 0.73–0.77
Tannast et al. [31]	2	Pelvic tilt	ICC = 0.94	ICC = 0.96–0.97
		Pelvic rotation	ICC = 0.91	ICC = 0.96–0.97
Steppacher et al. [28]	2	Tönnis OA grade	K = 0.74	K = 0.73–0.76

ICC = inter/intraclass correlation coefficient; K = kappa value; R = value of unpaired two-tailed t-test; OA = osteoarthritis.

this study believe each of these elements can still provide important information when factored in with other aspects of the patient presentation. It is likely that, given specific clinical information regarding hip symptoms and given complete details on physical examination, the diagnostic reliability of the observers would be improved. Second, the observers agreed on a set of definitions to describe the structural anatomy, yet the method of measurement (or simple observation) was left to the individual observer. We believed this better simulated the clinical situation of radiographic evaluation. Perhaps better reliability would have been achieved if all observers used the exact methodology to measure the radiographic parameters. Third, the radiographic techniques had some variability as a result of the use of different technicians. This was most notable with the cross-table lateral views that had some inconsistency with extremity rotation. However, inconsistencies in radiographic technique and image quality are inevitable in an everyday clinical setting, and reader interpretation of images will always vary from individual to individual, particularly those not familiar to specific parameters or definitions. These would undoubtedly reduce the reliability we report. An ideal marker for instability or impingement should remain reliable despite these limitations. Specifically because treatment is in part based upon radiograph review, excessively poor reliability in using commonly described structural features of the hip to diagnose disease could, in theory, result in misdiagnosis and therefore mistreatment.

While there have been an abundance of reliability studies in the pediatric literature [3–5, 12, 15, 24, 27] focused on the evaluation of radiographic parameters for dysplastic hips, fewer studies exist in the literature that are centered on the assessment of plain radiograph reliability in the assessment of markers of dysplasia or impingement in the skeletally mature hip.

Utilizing 100 hard-copy plain radiographs, Gosvig et al. [11] evaluated the inter- and intraobserver reliability of plain radiographic measurement of the alpha-angle of Notzli et al. [23] (a marker for cam impingement), and reported an interclass correlation coefficient of 0.83 and an intraclass correlation coefficient between 0.90 and 0.96 (Table 6). Clohisy et al. [7] also evaluated the reliability of plain radiograph measurement of the alpha angle, in addition to other markers of impingement—head sphericity (measured on multiple views with a concentric circle template) and an objectively calculated measure of head-neck offset. Dependent on the radiographic view, interobserver kappa values from two readers ranged from 0.66 to 0.82 for sphericity, 0.52 for head-neck offset, and 0.56 to 0.85 for alpha angle. Intraobserver values ranged from 0.98 to 0.99 for sphericity, 0.63 to 0.74 for head-neck offset, and 0.5 to 0.73 for alpha angle. Meyer et al. [19] also evaluated the alpha angle on radiographic projections of cadaveric bone taken at six different angles and, using unpaired two-tailed t-tests, two readers found combined intraobserver and interobserver correlations of R = 0.95 and R = 0.88, respectively. Though the numbers do not correlate precisely with kappa values, the general trends translate to higher degrees of correlation than that found in our study.

While the results from these three studies might again suggest that objective rather than subjective markers provide a more reliable means of evaluating for impingement features, all of them included data from only two readers. It is certain that the addition of multiple readers will sequentially limit the ability to obtain near perfect reliability scores. Additionally, in the case of some studies, concerns exist about the use of hard-copy radiographs for

use in reliability measurements, as previous reader markings can often influence subsequent reader evaluation of the films. Nevertheless, the improved results in these objective studies highlight the need for a larger scale study with a higher volume of readers to evaluate a series of objective markers of impingement in adults.

Even fewer studies have evaluated the reliability of radiographic markers in the adult dysplastic hip. Nelitz et al. [22] had three observers analyze 100 radiographs of patients between 16 and 32 years of age. They assessed nine radiographic features, of which three had some overlap with our study—acetabular index, lateral subluxation of the hip, and acetabular index of depth to width. For the acetabular index they reported an intraclass correlation coefficient of 0.86 to 0.89 and a combined interclass correlation coefficient of 0.85. Though a direct comparison of correlation coefficients and kappa values cannot be done, their range for intraobserver reliability is similar to our kappa value of 0.73, while their combined interclass coefficient is markedly higher than the kappa value of 0.64 noted in our study. Similarly, for acetabular index of depth to width, they found an intraclass coefficient of 0.65 to 0.69 and a combined interclass coefficient of 0.61. Intraclass numbers again compare closely to the kappa values found in our study for assessment of acetabular depth ($\kappa = 0.61$), though again the interobserver values noted in our study remain low in comparison ($\kappa = 0.39$). This same pattern holds true for lateralization of the hip, with Nelitz et al. [22] intraclass values ranging from 0.85 to 0.90 and interclass values of 0.80 compared to kappa values of 0.76 and 0.52 noted in our study, respectively.

Using two readers and 100 hips, Tannast et al. [30] also evaluated multiple plain radiographic parameters of dysplasia and impingement in their study on validation of the Hip2Norm software. For parameters that overlapped with our study, they reported intraobserver reliability for acetabular inclination to range from a kappa value of 0.74 to 0.89, and for interobserver reliability, an interclass correlation coefficient of 0.61. These numbers are similar to the kappa values we found for intra- and interobserver reliability (0.73 and 0.64, respectively). For the crossover sign, they reported intraobserver kappa values of 0.73 to 0.77 and an interclass correlation coefficient of 0.60—numbers markedly improved in comparison to our intraobserver kappa of 0.46 and interobserver kappa of 0.39.

In a separate study, Tannast et al. [31] evaluated the reliability of two readers to accurately determine the vertical and horizontal distances between the sacrococcygeal junction and the pubic symphysis as indicators of tilt and rotation, respectively. They reported intraobserver values for both tilt and rotation of 0.96 to 0.97 (intraclass correlation coefficient) and interclass values of 0.94 for tilt and 0.91 for rotation. Those numbers are considerably improved compared to the kappa values of 0.55 and 0.39 noted for intraobserver and interobserver reliability in our study. This might be secondary to the subjective review of images performed in our study without routine calculation of distances in all cases. Alternatively, identification of the tip of the coccyx might be more difficult than identification of the sacrococcygeal junction, thereby limiting our results.

Steppacher et al. [28] evaluated the reliability of the Tönnis classification of osteoarthritis utilizing two readers and 50 image sets and reported intraobserver reliability of $\kappa = 0.73$ to 0.76. This is slightly higher than the mean combined intraobserver kappa value of 0.60 we found. Additionally, their interobserver kappa value of 0.74 was also higher than that in our study ($\kappa = 0.59$). The difference in values might be secondary to the introduction of a "normal" cohort of patients in our study (versus a dysplastic only cohort in their study), as the distinction between grade 0 and grade 1 osteoarthritis can sometimes be difficult and dependent on radiographic quality.

In general, it is unclear why expert readers demonstrated lower reliability in radiograph review compared to the medical student and resident reads as performed in the Nelitz study [22]. While the expert readers might have improved reliability with more objective measures, it still highlights the need for an improved means of interpreting plain radiographs. Future studies on this topic with readers of variable skill level and with variable clinical backgrounds (radiologists and orthopedic surgeons) would help to clarify this issue.

The results of this study demonstrate a clear need for either alternative measurement tools, incorporation of advanced imaging modalities, or a revised set of definitions to aid in the diagnosis of structural disorders of the young mature hip.

Steppacher et al. [29] performed a study evaluating the reliability of MRI to assess multiple parameters including head sphericity and the alpha angle and found high intraobserver reliability in both categories, with intraclass correlation coefficients of 0.81 to 0.82 and 0.79 to 0.86, respectively. Mean interobserver reliability was also improved in comparison to our study with interclass coefficients of 0.78 for sphericity and 0.81 for the alpha angle.

Conversely, for all of the radiographic parameters analyzed in our study, only three structural features (acetabular inclination, position of head center, Tönnis osteoarthritis grade) demonstrated intraobserver and/or interobserver reliability with kappa values above 0.55. Additionally, the intraobserver reliability for radiographic diagnoses (when all four diagnostic categories were included) resulted in a kappa of 0.61, although interobserver agreement resulted in a somewhat lower kappa of 0.54. Although the observers demonstrated acceptable agreement in identifying

acetabular dysplasia, there remained comparatively poor agreement between observers in regard to the diagnosis of impingement. Because acetabular inclination was the only structural feature with intraobserver and interobserver agreement kappa values greater than 0.6, the ability to reproducibly identify cases of dysplasia is potentially the result of the ability to reliably assess this specific radiographic parameter. Additionally, because this was the only quantitative feature included in this study, it is a possibility the qualitative nature of the other structural elements assessed limited their overall reliability. In either case, these results suggest if a reliable structural feature can be identified in cases of impingement, observers will ultimately be able to make the diagnosis in a more reliable fashion. To identify such a structural feature in the diagnosis of impingement, the logical place to start would be with those parameters that readers were able to consistently identify and reproduce between their first and second reads. Two data points that meet such criteria are the position of the head center and the assessment of the head-neck offset on the frog-lateral view. These are promising findings because position of the head center has a clear relationship to the diagnosis of pincer impingement just as the presence of decreased head-neck offset is directly related to the diagnosis of cam impingement. Perhaps, with an improved definition for each of these radiographic parameters, readers will demonstrate improved agreement in making these diagnoses. Alternatively, there are numerous other parameters that have been described in the literature (including, but not limited to the center-edge angle, the femoral head extrusion index, and the posterior wall sign) that might lead to more reliable measurements and, thus, a more reliable radiographic diagnosis. Nevertheless, our findings underscore the need for improved definitions of radiographic hip anatomy and refined methods for interpreting radiographic images.

Despite our attempt to define carefully and reasonably several standard diagnostic criteria commonly used in the diagnosis of structural hip abnormalities, our results demonstrate limited reliability in radiographic diagnosis. To some, this may serve to emphasize the importance of the history and physical examination in the workup of every patient. Patient symptoms and physical examination contribute considerably to the interpretation of the radiographs. Clinical findings ideally correlate with the imaging findings when reaching a diagnosis and when contemplating different surgical treatment options. Therefore, caution is urged when specific surgical treatments are recommended based on isolated radiographic findings, which may be prone to subjective interpretation. This also highlights the need to understand the pathomechanics of each individual case, especially in impingement surgery, rather than relying solely on static radiographic findings.

The diagnosis and treatment of prearthritic and early arthritic hip disease is an area of intense interest. As new and refined diagnostic and treatment modalities are introduced, it is essential to have reliable standards for patient evaluation and diagnosis. Our findings confirm that a need exists to establish a reliable and useful set of basic radiographic assessments to define the structural anatomy of the hip. Our immediate goal is to identify and test a precisely defined set of radiographic parameters that can provide reliable information regarding the pathoanatomy of the structurally abnormal prearthritic hip. This will facilitate improved dialogue regarding structural abnormalities of the hip and will also provide a foundation for future multicenter clinical trials.

Acknowledgments At the time of this study, the Academic Network for Conservational Hip Outcomes Research (ANCHOR) Study Group consisted of the following surgeons: Paul E. Beaule, MD, FRCSC; John C. Clohisy, MD; Young-Jo Kim, MD, PhD; Michael Millis, MD; Perry L. Schoenecker, MD; Rafael J. Sierra, MD; and Robert Trousdale, MD.

References

1. Beaule PE, Dorey FJ, Matta JM. Letournel classification for acetabular fractures. Assessment of interobserver and intraobserver reliability. *J Bone Joint Surg Am.* 2003;85:1704–1709.
2. Beck M, Kalhor M, Leunig M, Ganz R. Hip morphology influences the pattern of damage to the acetabular cartilage: femoroacetabular impingement as a cause of early osteoarthritis of the hip. *J Bone Joint Surg Br.* 2005;87:1012–1018.
3. Boniforti FG, Fujii G, Angliss RD, Benson MK. The reliability of measurements of pelvic radiographs in infants. *J Bone Joint Surg Br.* 1997;79:570–575.
4. Broughton NS, Brougham DI, Cole WG, Menelaus MB. Reliability of radiological measurements in the assessment of the child's hip. *J Bone Joint Surg Br.* 1989;71:6–8.
5. Carney BT, Rogers M, Minter CL. Reliability of acetabular measures in developmental dysplasia of the hip. *J Surg Orthop Adv.* 2005;14:73–76.
6. Clohisy JC, Keeney JA, Schoenecker PL. Preliminary assessment and treatment guidelines for hip disorders in young adults. *Clin Orthop Relat Res.* 2005;441:168–179.
7. Clohisy JC, Nunley RM, Otto RJ, Schoenecker PL. The frog-leg lateral radiograph accurately visualized hip cam impingement abnormalities. *Clin Orthop Relat Res.* 2007;462:115–121.
8. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas.* 1960;20:37–46.
9. Eijer H, Leunig M, Mahomed M, Ganz R. Cross table lateral radiograph for screening of anterior femoral head-neck offset in patients with femoroacetabular impingement. *Hip Int.* 2001;11:37–41.
10. Ganz R, Parvizi J, Beck M, Leunig M, Notzli H, Siebenrock KA. Femoroacetabular impingement: a cause for osteoarthritis of the hip. *Clin Orthop Relat Res.* 2003;417:112–120.
11. Gosvig KK, Jacobsen S, Palm H, Sonne-Holm S, Magnusson E. A new radiological index for assessing asphericity of the femoral head in cam impingement. *J Bone Joint Surg Br.* 2007;89:1309–1316.

12. Halanski MA, Noonan KJ, Hebert M, Nemeth BA, Mann DC, Leverson G. Manual versus digital radiographic measurements in acetabular dysplasia. *Orthopedics*. 2006;29:724–726.
13. Harris WH. Etiology of osteoarthritis of the hip. *Clin Orthop Relat Res.* 1986;213:20–33.
14. Heyman CH, Herndon CH. Legg-Perthes disease; a method for the measurement of the roentgenographic result. *J Bone Joint Surg Am.* 1950;32:767–778.
15. Kay RM, Watts HG, Dorey FJ. Variability in the assessment of acetabular index. *J Pediatr Orthop.* 1997;17:170–173.
16. Klaue K, Durnin CW, Ganz R. The acetabular rim syndrome. A clinical presentation of dysplasia of the hip. *J Bone Joint Surg Br*. 1991;73:423–429.
17. Lavigne M, Parvizi J, Beck M, Siebenrock KA, Ganz R, Leunig M. Anterior femoroacetabular impingement: part I. Techniques of joint preserving surgery. *Clin Orthop Relat Res*. 2004;418:61–66.
18. Lequesne M, de Seze. False profile of the pelvis. A new radiographic incidence for the study of the hip. Its use in dysplasias and different coxopathies [in French]. *Rev Rhum Mal Osteoartic.* 1961;28:643–652.
19. Meyer DC, Beck M, Ellis T, Ganz R, Leunig M. Comparison of six radiographic projections to assess femoral head/neck asphericity. *Clin Orthop Relat Res*. 2006;445:181–185.
20. Millis MB, Kim YJ. Rationale of osteotomy and related procedures for hip preservation: a review. *Clin Orthop Relat Res*. 2002;405:108–121.
21. Murphy SB, Ganz R, Muller ME. The prognosis in untreated dysplasia of the hip. A study of radiographic factors that predict the outcome. *J Bone Joint Surg Am.* 1995;77:985–989.
22. Nelitz M, Guenther KP, Gunkel S, Puhl W. Reliability of radiological measurements in the assessment of hip dysplasia in adults. *Br J Radiol.* 1999;72:331–334.
23. Notzli HP, Wyss TF, Stoecklin CH, Schmid MR, Treiber K, Hodler J. The contour of the femoral head-neck junction as a predictor for the risk of anterior impingement. *J Bone Joint Surg Br.* 2002;84:556–560.
24. Omeroglu H, Bicimoglu A, Agus H, Tumer Y. Measurement of center-edge angle in developmental dysplasia of the hip: a comparison of two methods in patients under 20 years of age. *Skeletal Radiol.* 2002;31:25–29.
25. Peelle MW, Della Rocca GJ, Maloney WJ, Curry MC, Clohisy JC. Acetabular and femoral radiographic abnormalities associated with labral tears. *Clin Orthop Relat Res.* 2005;441:327–333.
26. Reynolds D, Lucas J, Klaue K. Retroversion of the acetabulum. A cause of hip pain. *J Bone Joint Surg Br.* 1999;81:281–288.
27. Spatz DK, Reiger M, Klaumann M, Miller F, Stanton RP, Lipton GE. Measurement of acetabular index intraobserver and interobserver variation. *J Pediatr Orthop.* 1997;17:174–175.
28. Steppacher SD, Tannast M, Ganz R, Siebenrock KA. Mean 20-year followup of Bernese periacetabular osteotomy. *Clin Orthop Relat Res.* 2008;466:1633–1644.
29. Steppacher SD, Tannast M, Werlen S, Siebenrock KA. Femoral morphology differs between deficient and excessive acetabular coverage. *Clin Orthop Relat Res*. 2008;466:782–790.
30. Tannast M, Mistry S, Steppacher SD, Reichenbach S, Langlotz F, Siebenrock KA, Zheng G. Radiographic analysis of femoroacetabular impingement with Hip2Norm-reliable and validated. *J Orthop Res*. 2008;26:1199–1205.
31. Tannast M, Zheng G, Anderegg C, Burckhardt K, Langlotz F, Ganz R, Siebenrock KA. Tilt and rotation correction of acetabular version on pelvic radiographs. *Clin Orthop Relat Res*. 2005;438:182–190.
32. Tönnis D. *Congenital Dysplasia and Dislocation of the Hip in Children and Adults*. Berlin, Germany, New York, NY: Springer; 1987.
33. Wiberg G. Studies on dysplastic acetabula and congenital subluxation of the hip joint. With special reference to the complication of osteoarthritis. *Acta Chir Scand.* 1939;58:7–38.