

UCLA

UCLA Previously Published Works

Title

Inter- and intraobserver variation in staging patients with proven avascular necrosis of the hip.

Permalink

<https://escholarship.org/uc/item/75h78462>

Journal

Clinical Orthopaedics and Related Research, issue 307

ISSN

0009-921X

Authors

Kay, RM

Lieberman, JR

Dorey, FJ

et al.

Publication Date

1994-10-01

Peer reviewed

Inter- and Intraobserver Variation in Staging Patients With Proven Avascular Necrosis of the Hip

Robert M. Kay, MD; Jay R. Lieberman, MD*;
Frederick J. Dorey, PhD*; and Leanne L. Seeger, MD***

Diagnosis and treatment of avascular necrosis of the hip has long been predicated on the evaluation of plain radiographs, though other modalities (such as magnetic resonance imaging) are being increasingly used to aid in this evaluation. The Ficat classification is commonly used to assess plain radiographs and to help determine what treatment is appropriate for a given patient. It also is used to help evaluate patient outcome after surgical or nonsurgical treatment. This study was designed to evaluate the adequacy of plain radiographs in the evaluation of avascular necrosis of the hip. The plain radiographs of 25 hips with avascular necrosis were analyzed on 3 occasions by 6 readers (450 total readings). A clinically significant difference in radiographic staging was defined as a Stage I or II reading of the radiographs of a given hip on 1 reading and a separate reading of the same radiographs as Stage III or IV. By this definition, interobserver variability for the first reading resulted in clinically significant

differences in 10 (40%) of the 25 hips. Intraobserver variability resulted in clinically significant differences for 10 (40%) of the 25 hips as well. The most experienced readers in the study (a total joint specialist and a musculoskeletal radiologist) were consistent internally in their readings in 90% of cases, and had a clinically significant difference in only 1 case (1%); however, these same 2 readers disagreed with one another on the staging of 9 (36%) of 25 hips, with 4 (16%) of 25 hips having a clinically significant difference in readings. The kappa statistic for intraobserver and interobserver variability was 0.82 ± 0.16 and 0.56 ± 0.01 , respectively. This high degree of variability suggests that plain radiographs alone are often inadequate for evaluating avascular necrosis of the hip.

Avascular necrosis of the femoral head remains a common, and problematic, disorder of the hip. Although avascular necrosis of the femoral head has been recognized for decades, the natural history of this disease is still not fully understood.^{3,5,8} Plain radiographs are used to stage avascular necrosis of the hip initially and throughout its course. Initially, disease stage helps determine treatment options, because early lesions are often treated differently than are late lesions. The treatment, in turn, affects the subse-

From the Departments of *Orthopaedic Surgery and **Radiological Sciences, University of California School of Medicine, Los Angeles, CA.

Reprint requests to Jay R. Lieberman, MD, Department of Orthopaedic Surgery, UCLA School of Medicine, 10833 LeConte Ave, Los Angeles, CA 90024-6902.

Received: December 28, 1993.

Revised: June 9, 1994.

Accepted: June 14, 1994.

quent clinical course of the patient. The difference between a radiographic staging of Ficat Stage II and III is, potentially, a clinically significant difference. This difference in radiographic staging will likely result in a different treatment plan, patient course, and patient outcome. Once a treatment plan is established, either surgical or nonsurgical, plain films are evaluated intermittently to assess the efficacy of treatment. The interpretation of radiographs is instrumental in determining disease progression and further treatment options.

This study was designed to evaluate the inter- and intraobserver variability in the evaluation of plain radiographs in avascular necrosis of the hip, as reflected in the Ficat classification scheme.

MATERIALS AND METHODS

Radiographs of 25 hips with a biopsy-proven clinical diagnosis of avascular necrosis were reviewed retrospectively by 2 junior orthopaedic residents (each had performed clinical research in the field of avascular necrosis and was in his 3rd postgraduate year of training), 1 chief orthopaedic resident, 1 fellowship-trained total joint arthroplasty surgeon, 1 fellowship-trained (nonjoint) orthopaedic surgeon, and 1 musculoskeletal radiologist. The films were reviewed independently and, based on radiographic criteria only, staged according to the Ficat classification. The modified Ficat classification was used, with transitional lesions being categorized as Stage II lesions.¹ Each reviewer had previously used the Ficat classification in clinical situations. The reviewers were provided with a table summarizing the classification system at the time of radiographic review. Because only hips with pathologic confirmation of the diagnosis of avascular necrosis were included in the study, Stage I readings were assigned to radiographically normal hips. The studies were read 3 times by each observer, with at least 1 week between readings. There were 450 total readings.

Intraobserver variability was assessed by comparing each reader's 2nd and 3rd readings of the films of a given hip to his 1st reading of

TABLE 1. Intraobserver Variability (N = 300 Subsequent Readings): Differences When Each Reader's 2nd and 3rd Readings Were Compared With the 1st Reading of the Same Radiographs

Any difference in staging	17.7% (53/300)
CSD difference in staging	7.7% (23/300)
Kappa statistic (mean $\pm$ SD)	0.82 $\pm$ 0.16

CSD = clinically significant difference.

the same hip films (6 readers $\times$ 2 subsequent readings per reader $\times$ 25 hips = 300 subsequent readings).

Interobserver variability was assessed by comparing the first reading of each set of hip films by each of the 6 readers.

The kappa statistic was calculated separately for interobserver and intraobserver variability. The kappa statistic was based on the reproducibility of staging a lesion as either early (Stage I or II) or late (Stage III or IV). A kappa value of 0 indicates that the amount of agreement between different observations is equal to that amount of agreement expected by chance alone, and a kappa value of +1 indicates perfect agreement.² For the purposes of kappa statistic assessment, readings of Stage I were not considered different from Stage II readings, and Stage III readings were not considered different from Stage IV readings. Statistics were performed on a personal computer using the statistical package Stata (Stata Corporation, College Station, TX).

RESULTS

A clinically significant difference in staging was defined as the assignment of an early (Stage I or II) and a late (Stage III or IV) stage to the same hip radiographs at different sittings, either by the same observer (intraobserver clinically significant difference) or different observers (interobserver clinically significant difference).

There were 53 (17.7%) intraobserver differences and 23 (7.7%) clinically significant intraobserver differences (Table 1). All the

clinically significant different readings occurred among 10 (40%) of the 25 hips evaluated: 5 (20%) had intraobserver clinically significant differences for multiple readers, and each of the remaining 5 (20%) had intraobserver clinically significant differences for 1 reader only. The kappa statistic for intraobserver variability was 0.82 ± 0.16 (mean $\pm$ standard deviation), and ranged from 0.54 to 1.00 for the 6 readers.

The intraobserver variability was least for the total joint surgeon and the musculoskeletal radiologist, whose subsequent readings differed from their initial readings in 6 cases (12%) and 4 cases (8%), respectively. They had 1 (2%) and 0 (0%) clinically significant differences, respectively, and kappa statistics of 1.00 and 0.93, respectively.

Interobserver variability was evaluated based on the initial reading by each reader. The mean score was 2.88 ± 1.01 . Six hips (24%) were staged the same by all 6 evaluators. (These hips were staged as 4 Stage IV and 1 each Stages I and III.) Two hips (8%) received the same stage from 5 of the 6 observers. Eleven (44%) of the hips were staged the same by 4 of the 6 observers. The remaining 6 hips (24%) received the same stage from only 3 of 6 observers. Overall, 10 hips (40%) had interobserver clinically significant differences on the first reading. Two of these hips were staged early (Stage I or II) by 3 observers and late (Stage III or IV) by the other 3. The kappa statistic for interobserver variability was 0.56 ± 0.01 (mean $\pm$ standard deviation), and ranged from 0.55 to 0.57 for the 3 readings.

The readings of the total joint surgeon and the musculoskeletal radiologist and those of 2 attending surgeons were compared in a pair-wise fashion (Table 2). The musculoskeletal radiologist and the fellowship-trained total joint surgeon, the 2 readers with the least intraobserver variability in the study, still differed in their interpretations of particular radiographs: staging was different in 9 (36%) of the 25 hips, including

4 hips (16%) with interobserver clinically significant difference readings. When the first reading of each film by the 2 attending surgeons (1 total joint surgeon and 1 nontotal joint surgeon) were compared, there was disagreement on the stage in 14 of 25 cases (56%), and 5 hips (20%) had interobserver clinically significant different readings.

When all readers are considered, 12 (48%) of the 25 hips had a clinically significant difference for interobserver variability, intraobserver variability, or both (Table 3). Eight of these 12 hips had a clinically significant difference for both inter- and intraobserver variability, and 2 each had either inter- or intraobserver clinically significant differences.

When the first reading of each film was evaluated, the reading of the orthopaedist was the same as that of the radiologist in 57% of cases (71 of 125 readings). In 98% of cases in which there was disagreement, the radiologist assigned a higher stage.

DISCUSSION

Plain radiographs have long been an important tool in the evaluation of avascular necrosis of the femoral head, both in the initial evaluation and in followup. The Ficat classification continues to be commonly used in this evaluation. Many studies have reported the prognostic importance of the Ficat classification in the progression of avascular necrosis of the femoral head with and without surgery.^{1,14,17} Since early lesions often respond better than late lesions to conservative surgical intervention such as core decompression or osteotomy, radiographic guidelines are often used to help determine patient treatment. Plain radiographs are often used to determine disease progression, even in studies that employ other modalities, such as magnetic resonance (MR) imaging, to assess the site and extent of femoral head involvement at the time of diagnosis.^{4,5,14-16} The implicit as-

TABLE 2. Interobserver CSD* Readings (N = 25 Hips): Differences in the Initial Staging of the 25 Radiographs by Pairs of Experienced Observers

Difference	2 Attending Orthopaedists	Joint Surgeon and Radiologist
Any difference in staging	56% (14/25)	36% (9/25)
CSD difference in staging	20% (5/25)	16% (4/25)

CSD = clinically significant difference.

sumption in such studies is that there is a true stage which can be assigned to each lesion, and that it has been done in each case.

The Ficat classification, despite its widespread use, does have limitations. It does not take into account the size or location of the osteonecrotic portion of the femoral head, both of which appear to be important prognostic factors.^{5,8,14-16} Its few stages may, at times, result in difficulty in precisely staging a hip with avascular necrosis. In addition, the Ficat classification does not benefit from the improved imaging provided by MR imaging and computed tomography (CT) scanning.

Other authors break the radiographic stages of avascular necrosis down more precisely to overcome these shortcomings.^{6,8,12,13} The system of Marcus et al⁶ is very precise in its radiographic descriptions, and therefore may result in less ambiguity

in the interpretation of particular plain radiographs. Steinberg and Steinberg¹² have developed a system that takes into account the extent of femoral head involvement in its precise descriptions of the various stages. Despite their precision and possible advantages, however, these schema are still not used as commonly as the Ficat classification. Their complexity may make their daily application challenging. In addition, some of these systems would only be applicable for an initial evaluation, because avascular necrosis is usually followed with serial plain radiographs and not serial MR imagings. To the authors' knowledge, the reproducibility of these systems has not previously been evaluated either.

The inter- and intraobserver variability demonstrated in this study shed doubt on the question of whether a Ficat stage can be accurately assigned to a hip based on plain radiographs. Because Stage I or II lesions are generally treated differently from Stage III or IV lesions, the authors defined a clinically significant difference in readings as that in which a given hip was staged as a I or II at one time and as a III or IV on another occasion. Certainly, there are extreme cases of avascular necrosis either very early or very late, which will be agreed on almost universally. But, as the current study demonstrates, a large proportion of plain radiographs that demonstrate the disease are read differently by different individuals, or by the same individual at different times.

TABLE 3. Inter- and Intraobserver CSD Readings (N = 25 Hips): All Clinically Significant Differences Found in All Readings of the Radiographs of the 25 Hips Evaluated

Interobserver CSD readings only	8% (2/25)
Intraobserver CSD readings only	8% (2/25)
Both inter- and intraobserver CSD readings	32% (8/25)
No CSD readings	52% (13/25)

CSD = clinically significant difference.

Twelve (48%) of the 25 sets of hip radiographs had an interobserver clinically significant difference, intraobserver clinically significant difference, or both. Eight (32%) of the 25 hips reviewed had inter- and intraobserver clinically significant differences. The kappa statistic, although clearly higher for intraobserver reliability than for interobserver reliability, still demonstrates that there is far from perfect reproducibility in the assessment of plain radiographs in avascular necrosis of the hip.

Even those readers with the most experience and least intraobserver variability (the total joint surgeon and the musculoskeletal radiologist) often differed in the stage assessed a particular hip. Considering the initial reading of each film, 9 (36%) of their 25 readings differed, and 4 (16%) of the 25 readings had an interobserver clinically significant difference.

The 2 attending orthopaedic surgeons differed even more frequently, with disagreement on the first reading of 14 (56%) of the 25 hips, and clinically significant difference readings in 5 hips (20%).

Recently, other imaging modalities (such as MR imaging, CT scans, and bone scans) have been increasingly used to assess avascular necrosis of the hip.^{3-5,7,9,16} The efficacy of these modalities to determine femoral head collapse has not been clearly defined. In contrast, plain radiographs are almost universally obtained and evaluated in the initial evaluation of patients with avascular necrosis, and in followup to determine patient outcome. The interpretation of such initial and followup radiographs also forms the basis for much of the current literature regarding the classification of avascular necrosis, treatment results, and treatment recommendations.

It is essential that any classification system be reliable and reproducible. Recently, the results of 2 studies evaluating the inter- and intraobserver variability of proximal humeral classification systems have been dis-

appointing.^{10,11} This study demonstrates the potential importance of inter- and intraobserver variability in the assessment of plain films in avascular necrosis of the hip. Because staging often determines treatment of the patient and the patient's subsequent clinical course, this variability takes on added significance. Plain films are generally followed to assess disease progression and to determine treatment efficacy. Caution must be exercised in the interpretation of studies comparing the results among patients with different stages of avascular necrosis, because this variability may affect the initial categorization of patients, their treatment, and evaluation of their results. CT scans and MR images may provide important additional information in the planning of treatment for these patients.

References

1. Ficat RP: Idiopathic bone necrosis of the femoral head: Early diagnosis and treatment. *J Bone Joint Surg* 67B:3-9, 1985.
2. Fisher LD, Van Belle G: Biostatistics: A Methodology for the Health Sciences. New York, John Wiley & Sons, Inc 256-259, 1993.
3. Fordyce MJF, Solomon L: Early detection of avascular necrosis of the femoral head by MRI. *J Bone Joint Surg*. 75B:365-367, 1993.
4. Kokubo T, Takatori Y, Ninomiya S, Nakamura T, Kamogawa M: Magnetic resonance imaging and scintigraphy of avascular necrosis of the femoral head: Prediction of subsequent segmental collapse. *Clin Orthop* 277:54-60, 1992.
5. Lafforgue P, Dahan E, Chagnaud C, et al: Early-stage avascular necrosis of the femoral head: MR imaging for prognosis in 31 cases with at least 2 years of follow-up. *Radiology* 187:199-204, 1993.
6. Marcus ND, Enneking WF, Massam RA: The silent hip in idiopathic aseptic necrosis: Treatment by bone grafting. *J Bone Joint Surg* 55A:1351-1366, 1973.
7. Mitchell DG, Steinberg ME, Dalinka MK, et al: Magnetic resonance imaging of the ischemic hip. *Clin Orthop* 244:60-77, 1989.
8. Ohzono K, Saito M, Sugano N, et al: Natural history of nontraumatic avascular necrosis of the femoral head. *J Bone Joint Surg* 73B:68-72, 1991.
9. Robinson H, Hartleben PD, Lund G, Schreiman J: Evaluation of magnetic resonance imaging in the diagnosis of osteonecrosis of the femoral head: Accuracy compared with radiographs, core biopsy

- and intraosseous pressure measurements. *J Bone Joint Surg* 71A:650-663, 1989.
10. Sidor ML, Zuckerman JD, Lyon T, et al: Neer classification system for proximal humeral fractures: An assessment of interobserver reliability and intraobserver reproducibility. *J Bone Joint Surg* 75A:1745-1750, 1993.
 11. Siebenbrock KA, Gerber C: The reproducibility of classification of fractures of the proximal end of the humerus. *J Bone Joint Surg* 75A:1751-1755, 1993.
 12. Steinberg ME, Steinberg DR: Avascular Necrosis of the Femoral Head. In Steinberg ME (ed). *The Hip and its Disorders*. Philadelphia, WB Saunders Company 623-647, 1991.
 13. Stulberg BN, Levine M, Bauer TW, et al: Multimodality approach to osteonecrosis of the femoral head. *Clin Orthop* 240:181-193, 1991.
 14. Sugano N, Takaoka K, Ohzono K, et al: Rotational osteotomy for non-traumatic avascular necrosis of the femoral head. *J Bone Joint Surg* 74B:734-739, 1992.
 15. Sugioka Y, Hotokebuchi T, Tsutsui H: Transtrochanteric anterior rotational osteotomy for idiopathic and steroid-induced necrosis of the femoral head: Indications and long-term results. *Clin Orthop* 277:111-120, 1992.
 16. Takatori Y, Kokubo T, Ninomiya S, et al: Avascular necrosis of the femoral head: Natural history and magnetic resonance imaging. *J Bone Joint Surg* 75B:217-221, 1993.
 17. Warner JJP, Philip JH, Brodsky GL, Thornhill TS: Studies of nontraumatic osteonecrosis: The role of core decompression in the treatment of nontraumatic osteonecrosis of the femoral head. *Clin Orthop* 225:104-126, 1987.