Clinical Imaging 54 (2019) 103–107

Contents lists available at ScienceDirect

Clinical Imaging

journal homepage: www.elsevier.com/locate/clinimag

Musculoskeletal and Emergency Imaging

Sensitivity and specificity of post-operative interference gap assessment on plain radiographs after cementless primary THA

Maartje Belt[a,*], Bjørn Gliese[a], Omar Muharemovic[b], Henrik Malchau[c,d], Henrik Husted[a], Anders Troelsen[a], Kirill Gromov[a]

[a] Dept. of Orthopaedic Surgery, Copenhagen University Hospital Hvidovre, Kettegård Alle 30, 2650 Hvidovre, Copenhagen, Denmark
[b] Department of Radiology, Centre for Functional and Diagnostic Imaging and Research, Copenhagen University Hospital Hvidovre, Denmark
[c] Dept. of Orthopaedic Surgery, Sahlgrenska University Hospital, Mölndal, Gothenburg, Sweden
[d] The Harris Orthopaedic Laboratory, Orthopedic Department, Massachusetts General Hospital, Boston, USA

ARTICLE INFO

Keywords:
X rays
Radiolucency
Total hip replacement
Computed tomography

ABSTRACT

Introduction: Implant performance of cementless THA is often evaluated by radiolucency on plain radiographs, often classified as interference gaps on direct post-operative radiographs. However, the diagnostic performance is unknown. The aim was to evaluate the diagnostic performance of radiographic assessment of post-operative gaps after primary THA by comparing it with CT confirmed gaps, and secondary to define optimal cut-off criteria for assessing gaps on plain radiographs compared with CT.

Material and methods: Patients (N = 40) with a primary cementless THA performed between July 2015 and March 2016 were enrolled in the study. Radiolucency was assessed on post-operative AP pelvic digital radiographs by two observers independently. Maximum width and percentage of coverage per zone were reported. Gap volume was measured by manual segmentation on CT images.

Results: When defining a gap as a radiolucency extending through > 50% of a zone, the interrater agreement Kappa was 0.241. Sensitivity was 65.8% for observer 1 (Kappa = 0.432), and 86.8% for observer 2 (Kappa = 0.383). When defining a gap as a radiolucency with a width > 1 mm, the interrater agreement Kappa was 0.302. Sensitivity was 55.3% and 50% for observer 1 and observer 2, respectively. The ROC-curve resulted in an optimal threshold of 0.65 mm (AUROC = 0.888) and 0.31 mm (AUROC = 0.961) for the two observers.

Conclusion: The diagnostic performance of observers detecting interference gaps on radiographs showed low sensitivity. Further on, the inter-rater agreement is too low to do a general recommendation about thresholds for defining gaps. Evaluating progression of radiolucency on radiographs should be performed in the light of these findings.

1. Introduction

The use of cementless fixation in primary THA is increasing worldwide, in part due to the proposed benefit of biological fixation following cementless THA [1–3]. However, acetabular component seating in cementless hip arthroplasty can be challenging as initial press fit fixation is required. Obtaining press fit may be facilitated by underreaming the acetabulum; however, due to underreaming, higher friction forces are generated during cup placement and can cause a gap between the reamed bone and the cup. Some studies tested for an association between interference gaps and the need for revision, however, the number of revisions were too small (n < 5) to detect any significant results [4,5]. Thus, whether a gap caused by incomplete acetabular component seating results in increased acetabular loosening is not yet established.

Implant performance following THA is often evaluated by radiolucency on plain radiographs, as it has been suggested, that radiolucent lines on early post-operative radiographs around the acetabular component are associated with early failure and the need for revision surgery [6,7]. In radiolucency measurements, the presence of radiolucency on immediate post-operative radiographs - classified as interference gaps - is used as a baseline measurement to compare follow-up radiographs and evaluating radiolucency progression. However, radiolucency has a poor sensitivity in assessing loosening of the acetabular component [8], and the diagnostic performance of plain radiographs in detecting post-operative interference gaps has not been

* Corresponding author at: Hvidovre Hospital, Kettegård Alle 30, 2650 Hvidovre, Denmark.
E-mail address: maartjebelt@gmail.com (M. Belt).

https://doi.org/10.1016/j.clinimag.2018.12.009
Received 14 May 2018; Received in revised form 13 December 2018; Accepted 27 December 2018

validated yet. Hence, it would be useful to know the value of evaluating interference gaps on postoperative radiographs when using it as a baseline measurement. Further on, different criteria are used to define a gap on radiographs. Commonly, gap location is assigned using the DeLee and Charnley [9] acetabular zone classification [4,10–12]. However, to provide more detailed information, sometimes the acetabular bone is divided into five zones [13,14]. Some studies define a gap as continuous radiolucency covering > 50% of a zone [4,10], while others include gaps with a maximum width over 1 mm [11,12]. Additionally, in two studies, all radiolucencies were included as a gap, regardless of gap width and gap coverage [15,16]. Thus the optimal definition of radiolucency on post-operative radiographs remains unknown.

To gain more insight in the prevalence and diagnosis of interference gaps, a valid diagnostic tool should be used. Although CT scans have been proven to be accurate in detecting gaps, it is not the best option for clinical practice [17]. The primary aim of this study was to evaluate the diagnostic performance of radiographic assessment of post-operative interference gaps after primary THA by comparing it with CT confirmed gaps. The secondary aim was to define optimal cut-off criteria for assessing interference gaps on plain radiographs.

2. Material and methods

2.1. Study design

Patients that were enrolled in a randomized controlled trial including a postoperative CT-scan were included in the present study (Clinical Trials reg. no.: NCT02518269, Ethics Committee approval no. H-1-2014-120). Patients received a primary cementless THA between July 2015 and March 2016. All patients were operated on in one university hospital. A posterolateral approach was used in all patients. Patients allocation were concealed using envelopes that were opened during surgery. All patients received a cementless, porous plasma spray (PPS) G7™ acetabular cup implanted (ZimmerBiomet, Warsaw, Indiana, USA) with either an ArcomXL™ liner and a metal femoral head, an E1™liner and a metal femoral head, or a ceramic liner and a ceramic femoral head. According to the manufacturer's instructions reaming was one size less than the cup size implanted. No screws were utilized for fixation of the cups. An EchoBimetric cementless femoral component (ZimmerBiomet, Warsaw, Indiana, USA) was used in all patients.

2.2. Radiographic measurements

Acetabular radiolucency was assessed on post-operative AP pelvic digital radiographic images using mDesk software (RSA Biomedical Inc. Umea, Sweden). Acetabular radiolucency was defined as linear decrease in periprosthetic bone density, with or without sclerotic lines. Radiolucency was separately measured in the three DeLee and Charnley zones (Fig. 1) [9], with maximum width assessed in mm and the percentage of coverage in the corresponding zone. The coverage was divided into four categories: 0–25%, 26–50%, 51–75%, and 76–100%. The measurements were performed by two observers independently, and twice per patient in randomized order, with minimum 1 week between measurements. Both observers were blinded to each other's measurements and own previous measurements.

2.3. CT image measurements

Post-operative CT scans of the pelvis were performed in the supine position (Fig. 2). The subjects were scanned in the axial plane. Coronal and sagittal images were reconstructed from the axial images. A Toshiba ONE Aquilion Vision Edition 320 slice with metal artifact reduction was used to make the CT scans. CT images were typically taken at 120 kV and 225 mA s (slice thickness: 0.5 mm). These CT images were used to assess the 3D gap volume around the acetabular cup. A gap was defined as a continuous gap between the surface of an acetabular component and the bone of the acetabular bed. One observer assessed the gap volume of every patient on CT using VitreaCore 6.7.3 (Vital Images, Inc., Minnetonka, MN, USA). On all slices of the CT, the boundaries of the gap were manually traced and the software calculated the 3D volume.

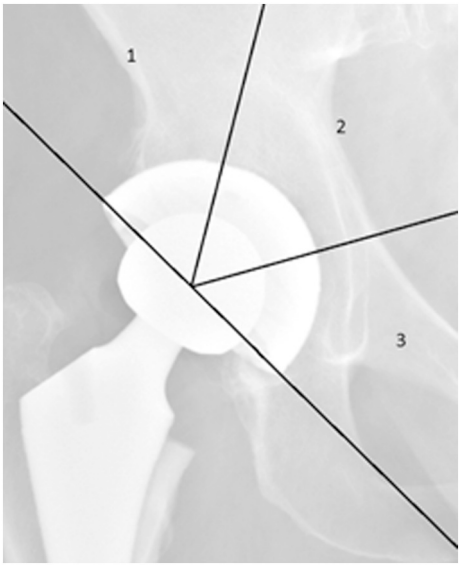

Fig. 1. Three DeLee and Charnley zones.

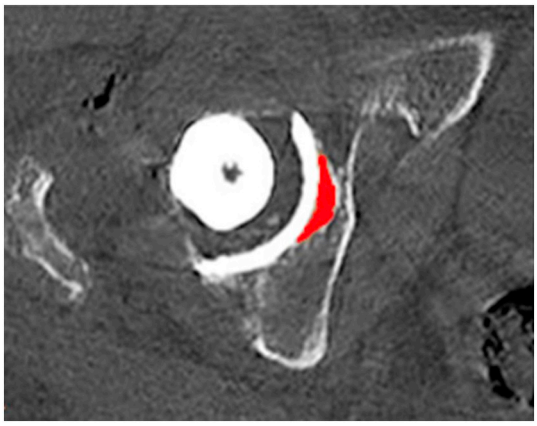

Fig. 2. A CT image with the gap segmented (red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.4. Statistical analyses

The performance of the observer when analyzing gaps on plain radiographs was evaluated by calculating the sensitivity, specificity and predictive value(PV), whereby the presence or absence of a gap on the CT images was considered as the true condition [18,19]. The level of inter- and intra-observer agreement on the presence of a radiolucent line on the radiograph was calculated using the kappa statistic, without any additional grading. The performance was assessed for the two most used definitions for interference gaps, namely a coverage of the lucent line of > 50% of one zone without a minimal requirement for gap width, and second a lucent line with a width > 1 mm at maximum, independent of the gap coverage.

To assess the performance of the independent variables that are currently used, we ran a linear predictor from a logistic model into a receiver operating characteristic (ROC) analysis. It was performed to

calculate the most optimal sensitivity and specificity in predicting the presence of a gap when using maximum gap width as continuous variable and gap coverage as categorical variable. Also, the area under the ROC (AUROC) curve was calculated to measure the overall ability of plain radiographs to discriminate between subjects with and without a gap. Furthermore, the effect of different cut-off points for coverage of the gap was assessed by reducing the logistic model. The corresponding optimal cut-off point for gap width was determined by finding the optimal threshold using the ROC curve. All statistical tests were performed using R (version 3.3.2). The irr package was used for calculating Cohen's Kappa, and the Epi package was used for calculating the AUROC curve [20–22].

3. Results

3.1. Baseline characteristics and CT assessment

Forty patients were included in the study. The median age of the patients was 68.5 years old (range: 45–75). 50% of the patients were male. Thirty-eight of the 40 (95%) patients had an interference gap present on the post-operative CT images. The median (IQR) volume of the gap was 2.63 ml (0.3675–3.84). The median (IQR) maximum gap width measured by observer 1 was 1.0 mm (0.6–1.3) and 0.9 mm (0.4–1.8) for observer 2.

3.2. Assessment on plain radiographs

When defining a gap as a continuous radiolucency of any width, that covers > 50% of a zone on the radiographs, observer 1 correctly identified 25 of the 38 interference gaps. In 1 case an interference gap was falsely identified by observer 1. Observer 2 correctly identified 32 of the 38 interference gaps, and 2 interference gaps were falsely identified by observer 2. The Kappa for inter-rater agreement is 0.241 (p = .0777). The diagnostic performance of the two observers using this definition is displayed in Table 1.

When defining a gap as a continuous radiolucency with a width > 1 mm, observer 1 correctly identified 21 of the 38 interference gaps, and no patients were falsely identified. Observer 2 correctly identified 18 of the 38 interference gaps, and no patients were falsely identified. The Kappa for inter-rater agreement is 0.302 (p = .0551). The diagnostic performance of the two observers using this definition is displayed in Table 2.

3.3. Logistic model

Model 1 was the full logistic model based on maximum gap width as a continuous variable and gap coverage as a categorical variable. The results of the model represent the most optimal result that can be obtained with these two variables. Running the linear predictor of model 1 into the ROC, results in an AUROC of 0.908. At the optimal cut-off value for the linear predictor, the sensitivity is 86.8% and the specificity is 100%. The positive PV is 71.4% and the negative PV is 0%.

The model is reduced by decreasing the number of categories for gap coverage to two categories. This resulted in model 3, 4, and 5 where the two categories were < 25% and ≥25%, < 50% and ≥50%, < 75% and ≥75%, respectively. Model 2 was based on only the maximum gap width. We ran a linear predictor of all four models into an ROC to test the performance. Besides, the ROC curve was used to calculate the optimal cut-off point for gap width to discriminate whether a gap is present or not. Table 3 displays the performance of the different models and the calculated cut-off points.

When repeating the logistic models with the data of the second observer, the effect of varying thresholds for gap coverage gives similar variation in optimal cut-off point for gap width. However, the cut-off point is ranging from 0.31–0.35 mm. The sensitivity is higher (around 95%) and has less uncertainty (95% CI: 86.8–100%).

Table 1
Diagnostic performance of the two observers of gap assessment on radiographs. A gap is defined as a continuous radiolucency of any width, that covers > 50% of a zone

	Observer 1	Observer 2
Sensitivity	65.8%	86.8%
Specificity	50%	0%
Positive PV	96.2%	94.1%
Negative PV	7.1%	0.0%
Intra-rater agreement	Kappa = 0.432 (p < .01)	Kappa = 0.383 (p < .01)

Table 2
Diagnostic performance of the two observers of gap assessment on radiographs. A gap is defined as a continuous radiolucency with a maximum width > 1 mm

	Observer 1	Observer 2
Sensitivity	55.3%	50%
Specificity	100%	100%
Positive PV	100%	100%
Negative PV	10.5%	9.5%
Intra-rater agreement	Kappa = 0.452 (p < .01)	Kappa = 0.95 (p < .0005)

Table 3
Performance of different thresholds based on logistic models

	Model 2	Model 3	Model 4	Model 5
Gap coverage ≥	0%	25%	50%	75%
Optimal gap width (mm) >	0.65	0.65	0.63	0.67
Sensitivity (95% CI)	76.3%	76.3%	84.2%	86.8%
	63.2–89.5	63.2–89.5	71.1–94.7	76.3–97.4
Specificity (95% CI)	100%	100%	100%	100%
	100–100	100–100	100–100	100–100
AUC	0.888	0.895	0.901	0.882

4. Discussion

The primary aim of the study was to determine the diagnostic performance of plain post-operative radiographs when analyzing interference gaps following primary THA. A gap present on a post-operative CT image was considered the diagnostic gold standard. The diagnostic performance was calculated for the two most used definitions of an interference gap separately. The first definition considers a gap present when it covers over 50% of one of the Delee and Charnley zones. The sensitivity was moderate but the specificity was poor. Also, the intra- and inter-rater agreement was poor. The second definition reports a gap as present when the maximum width of the radiolucency is > 1 mm, independent of the gap coverage. The sensitivity was poor, but the specificity was good. The intra-rater agreement was mostly fair to moderate, dependent on the observer and gap definition. The inter-rater agreement was poor.

We found that the use of both definitions for gaps on radiographs result in a poor to moderate sensitivity, meaning the proportion of true positives that are correctly identified by the test is not very high. The percentage of patients with a gap present that will not be identified ranges from 13.2% to 50%. The performance of observers when analyzing gaps on radiographs using the second definition (> 1 mm) is particularly poor. The sensitivity is around 50%, so it does not differ much from finding positive results by chance. It is important to emphasize that the specificity of the test can't be determined in this study population. Only 2 patients were true negatives, meaning the proportion of true negatives that are correctly identified by the test can only be 0%, 50% or 100% in this case. Thus, a larger sample size should be

tested to obtain specificity results that are more reliable.

The intra-rater agreement is mostly fair to moderate. This implies that the test results differ quite between two measurements. This indicates that the percentage of agreement is partly due to chance. Further, it also influences the reliability of the results when comparing a follow-up radiograph with the results of a post-operative radiograph. The inter-rater agreement is poor for both definitions, so the test results are even more susceptible to chance when two different raters are used to evaluate the gaps on radiographs.

The secondary aim was to determine optimal cut-off values for the gap width and extent of coverage per zone for interference gap on post-operative plain radiographs. Based on logistic regression model of the first rater, the most optimal cut-off value for gap width seems to be around 0.65 mm, independent of the percentage of gap coverage. The model more or less explains the same amount of variation without the gap coverage variable. The increase in precision of the model is not significant enough for the increase in complexity of the model. Further, the change in optimal threshold by adding the variable is clinically not relevant. However, it should be noted that the optimal cut-off changes when running the model on the data of another observer. For the second rater, the optimal cut-off range was around 0.33 mm. This can be explained by the low inter-rater agreement between the measurements. As a result, the cut-off value is specific for the rater. Therefore, we would not advocate using the cut-off criteria. A larger sample size with more different observers is needed for a threshold less specific to the rater.

Multiple studies have reported the prevalence of post-operative interference gaps after cementless THA. The prevalence varies widely from 18% to 79% [4,10–12,14–16]. The variation can partly be explained by the different descriptions that are used in defining a gap. In addition, the radiographic evaluation is dependent on the positioning of the patient, and the position of the gap with respect to the acetabular cup. Besides, as this study has shown, there are differences within and between observers. Therefore, CT images are a more reliable solution to establish the prevalence of interference gaps.

Abrahams et al. [8] had studied the diagnostic performance of radiographic criteria to detect acetabular aseptic loosening after revision THA. Aseptic loosening was measured by radiolucent lines, and compared with intra-operative classification of loosening. They found a sensitivity of 41% and a specificity of 100% of the radiolucent criteria. Even though their assessment of true disease status may be considered to be subjective, the sensitivity is still considered as poor. They do not recommend using the radiolucency criteria alone to exclude component loosening because of the sensitivity. These results might be explained by the low intra- and inter-rater agreement this study found. If the rater has a large influence on the measurements, comparing two measurements to find the progression of radiolucency is also affected by it.

In addition, Leung et al. (2005) and Puri et al. (2002) also studied the agreement between radiographs and CT images. The first selected hemipelvis specimens with CT confirmed bone defects, the latter patients with a primary THA, asymptomatic for osteolysis but thought to be at increased risk. Even though they focused on detecting osteolytic regions, the results were similar to this study. Leung et al. found the sensitivity of radiographs was 52% and Puri et al. observed a sensitivity of 62% [17,23]. Despite the larger volume of the osteolytic regions (mean 12.7 cm3 and 49 cm3), radiographs are still not really sensitive in identifying the regions.

This study has several limitations. First of all, because of the high prevalence of gaps in this population, the sample size was not large enough to give valid estimates for specificity. Although the limited sample size also resulted in a large confidence interval for sensitivity, it was sufficient for an indication of the sensitivity. Next, the agreement between raters was very low, which caused a high variation in optimal threshold between the raters. Therefore, the calculated threshold is not generally applicable. Including more raters in the study could result in a better insight in the difference in optimal threshold between observers, in order to be able to extrapolate the results. Lastly, the experience of the raters might also affect the results. The second observer was more experienced with the radiographic analysis, which might explain part of the difference in intra-rater agreement.

In conclusion, the diagnostic performance of plain radiographs to detect post-operative interference gaps has low sensitivity. Further on, the inter-rater agreement is not good, and therefore it was not possible to do a general recommendation about thresholds for defining gaps. Thus, analysis of interference gaps on plain radiographs should be performed in light of these findings. The diagnostic performance should also be taken into account when evaluating the progression of radiolucency during follow-up. For gap measurements in research settings, we would advise to use CT images for more reliable measurements. For instance, in future studies that look into the correlation between gaps, caused by incomplete acetabular component seating, and increased acetabular loosening. In clinical settings, the radiographic criteria can be used to support the clinical symptoms in making a diagnosis, while taking diagnostic limitations into consideration.

Declaration of conflicting interests

AT is consultant for ZimmerBiomet outside of this present study.

Funding

The 3 arm study is receiving financial support from ZimmerBiomet, but this study was initiated by the investigators and did not receive any additional support.

Acknowledgments

Nothing to acknowledge.

References

[1] Hailer NP, Garellick G, Karrholm J. Uncemented and cemented primary total hip arthroplasty in the Swedish hip arthroplasty register. Acta Orthop 2010;81:34–41.
[2] Makela KT, Matilainen M, Pulkkinen P, et al. Failure rate of cemented and uncemented total hip replacements: register study of combined Nordic database of four nations. BMJ 2014;348:f7592.
[3] Troelsen A, Malchau E, Sillesen N, et al. A review of current fixation use and registry outcomes in total hip arthroplasty: the uncemented paradox. Clin Orthop Relat Res 2013;471:2052–9.
[4] Gomes B, Olsen M, Donnelly M, et al. Should we worry about periacetabular interference gaps in hip resurfacing? Clin Orthop Relat Res 2013;471:422–9.
[5] Hulst JB, Ball ST, Wu G, et al. Survivorship of conserve(R) plus monoblock metal-on-metal hip resurfacing sockets: radiographic midterm results of 580 patients. Orthop Clin North Am 2011;42:153–9.
[6] Manley MT, Capello WN, D'Antonio JA, et al. Fixation of acetabular cups without cement in total hip arthroplasty. A comparison of three different implant surfaces at a minimum duration of follow-up of five years. J Bone Joint Surg Am 1998;80:1175–85.
[7] Yahiro MA, Gantenberg JB, Nelson R, et al. Comparison of the results of cemented, porous-ingrowth, and threaded acetabular cup fixation. A meta-analysis of the orthopaedic literature. J Arthroplasty 1995;10:339–50.
[8] Abrahams JM, Kim YS, Callary SA, et al. The diagnostic performance of radiographic criteria to detect aseptic acetabular component loosening after revision total hip arthroplasty. Bone Joint J 2017;99:458–64.
[9] DeLee JG, Charnley J. Radiological demarcation of cemented sockets in total hip replacement. Clin Orthop Relat Res 1976:20–32.
[10] Gruen TA, Poggie RA, Lewallen DG, et al. Radiographic evaluation of a monoblock acetabular component: a multicenter study with 2- to 5-year results. J Arthroplasty 2005;20:369–78.
[11] Nakashima Y, Mashima N, Imai H, et al. Clinical and radiographic evaluation of total hip arthroplasties using porous tantalum modular acetabular components: 5-year follow-up of clinical trial. Mod Rheumatol 2013;23:112–8.
[12] Springer BD, Griffin WL, Fehring TK, et al. Incomplete seating of press-fit porous-coated acetabular components: the fate of zone 2 lucencies. J Arthroplasty 2008;23:121–6.
[13] Bobyn JD, Toh KK, Hacking SA, et al. Tissue response to porous tantalum acetabular cups: a canine model. J Arthroplasty 1999;14:347–54.
[14] Macheras GA, Papagelopoulos PJ, Kateros K, et al. Radiological evaluation of the metal-bone interface of a porous tantalum monoblock acetabular component. J Bone Joint Surg Br 2006;88:304–9.
[15] Merican AM, Randle R. Early clinical and radiographic analysis of the Fitmore cup.

J Arthroplasty 2006;21:846–51.
[16] Morscher E, Berli B, Jockers W, et al. Rationale of a flexible press fit cup in total hip replacement. 5-year follow up in 280 procedures. Clin Orthop Relat Res 1997:42–50.
[17] Puri L, Wixson RL, Stern SH, et al. Use of helical computed tomography for the assessment of acetabular osteolysis after total hip arthroplasty. J Bone Joint Surg Am 2002;84:609–14.
[18] Altman DG, Bland JM. Diagnostic tests 1: sensitivity and specificity. BMJ 1994;308:1552.
[19] Altman DG, Bland JM. Diagnostic tests 2: predictive values. BMJ 1994;308:1552.
[20] Team RC. R: A language and environment for statistical computing R foundation for statistical computing Available from https://www.R-project.org; 2016.
[21] Gamer ML, J Fellows Puspendra Singh, I. irr: Various Coefficients of Interrater Reliability and Agreement Available from https://CRAN.R-project.org/package=irr; 2012.
[22] Carstensen B, Laara E, Hills M. Epi: A Package for Statistical Analysis in Epidemiology. https://CRAN.R-project.org/package=Epi; 2017 Available from:.
[23] Leung S, Naudie D, Kitamura N, et al. Computed tomography in the assessment of periacetabular osteolysis. J Bone Joint Surg Am 2005;87:592–7.