Radiographic assessment of uncemented total hip arthroplasty: reliability of the Engh Grading Scale

Susan W. Muir, PT, PhD[*]
Aziz Al-Ahaideb, MD[†]
John Huckell, MD[‡]
Mary Ann Johnson, MD[§]
D. Bill C. Johnston, MD[‡¶]
Lauren A. Beaupre, PT, PhD[‡**]

From the *Division of Geriatric Medicine, Department of Medicine, Schulich School of Medicine and Dentistry, University of Western Ontario, London, Ont., the †Department of Orthopaedics, College of Medicine, King Saud University, Saudi Arabia, the ‡Division of Orthopedic Surgery, Department of Surgery, University of Alberta, the §Departments of Diagnostic Imaging, University of Alberta and Alberta Health Services—Edmonton Zone, the ¶Division of Orthopedic Surgery, Department of Surgery, Alberta Health Services—Edmonton Zone, and the **Department of Physical Therapy, University of Alberta, Edmonton, Alta.

Accepted for publication
Jan. 12, 2011

Correspondence to:
Dr. L.A. Beaupre
Orthopedic Research
Alberta Health Services—Edmonton Zone
1F1.52 WMC, 8440-112 St.
Edmonton AB T6G 2B7
lauren.beaupre@albertahealthservices.ca

DOI: 10.1503/cjs.002610

Background: Radiographic evaluation has a prominent place in the follow-up of long-term results of uncemented total hip arthroplasty (THA). The most prominent scale reported in studies is the Engh Grading Scale, but there is a lack of literature on the reliability of the scale.

Methods: We evaluated intra- and interrater reliability of the Engh Grading System for uncemented THA using 26 follow-up radiographs of patients who had primary uncemented THAs. Four evaluators with different skill levels and specialties participated: 2 arthroplasty surgeons, an orthopedic resident and a radiologist. Reliability was measured using a weighted κ coefficient for paired comparisons among the evaluators.

Results: Intrarater reliability was dependent on the skill and specialty of the evaluator, with the highest values achieved for the arthroplasty surgeons (κ = 0.52 and κ = 0.68) and the lowest values for the radiologist (κ = 0.14). Interrater reliability was comparable among participants, regardless of skill or specialty, and rated a moderate level of reliability (κ = 0.29–0.41) for all pairings.

Conclusion: The Engh Grading Scale appears to be reliable when used by a single, experienced arthroplasty surgeon. Caution must be exercised when multiple raters are used, regardless of experience, as the interrater reliability achieved lower ratings.

Contexte : L'évaluation radiographique joue un rôle de premier plan dans le suivi des résultats à long terme de l'arthroplastie totale de la hanche (ATH) non cimentée. L'échelle d'évaluation de Engh (score de Engh) est la plus importante signalée dans les études, mais peu de publications en ont évalué la fiabilité.

Méthodes : Nous avons évalué la fiabilité intra-évaluateur et inter-évaluateurs du score de Engh dans le cas de l'ATH non cimentée en utilisant 26 radiographies de suivi de patients qui avaient subi une ATH primitive non cimentée. Quatre évaluateurs de divers degrés de compétence et de spécialités différentes ont participé : 2 chirurgiens en arthroplastie, 1 médecin résident en orthopédie et 1 radiologiste. On a mesuré la fiabilité au moyen d'un coefficient κ pondéré pour des comparaisons appariées entre les évaluateurs.

Résultats : La fiabilité intra-évaluateur a varié en fonction des compétences et de la spécialité de l'évaluateur. Les chirurgiens en arthroplastie ont obtenu les valeurs les plus élevées (κ = 0,52 et κ = 0,68) et les radiologistes, les moins élevées (κ = 0,14). La fiabilité inter-évaluateurs était comparable entre les participants, sans égard au degré de compétence ou à la spécialité, et toutes les paires ont conclu à une fiabilité modérée (κ = 0,29–0,41).

Conclusion : Le score de Engh semble fiable lorsqu'il est utilisé par 1 seul chirurgien en arthroplastie chevronné. Il faut faire preuve de prudence dans le cas de multiples évaluateurs, sans égard à l'expérience, car la fiabilité inter-évaluateurs a été plus basse.

As the population ages, the absolute number of adults older than 65 years increases, and people are living longer. Furthermore, the annual numbers of total hip arthroplasties (THAs) have been increasing in Canada at a substantial rate, particularly in patients younger than 65 years.[1] All of these factors have implications for arthroplasty surgery. Long-term surveillance of patient outcomes is necessary, and radiographic evaluation is an important and routine part of that assessment.[2–4] In fact, routine postoperative surveillance, including radiography, is a topic of ongoing discussion to determine the best practice for long-term management of patients who have had THAs.[2,3,5] Radiographic evaluation also has a prominent place in longitudinal studies on long-term results of uncemented THA. Thus, it is important to be able to reliably and effectively evaluate the status of uncemented THAs using standardized radiographic evaluation.

The most prominent scale reported in the literature is the Engh Grading Scale, a measurement scale that was first published in 1989.[6] It contains 2 subscales, fixation and stability, which are summed for a total score (Table 1).[6] Based on the total score, the implant is classified into 1 of 4 categories: "unstable" (< –10), "suboptimum but stable" (–10 to < 0), "in-growth suspected" (0 to +10) and "bone ingrown" (> +10).

A review of citations for this paper yielded 135 references in PubMed; 42 papers were published between 2005 and 2009, indicating that this scale is still considered relevant to establishing long-term results of THA. It is not always clear in this literature how researchers use the scale to interpret radiographic findings. Investigators will often discuss portions of the Engh scale to reflect the proportion of implants that are considered well-fixated and stable or unstable but do not report overall scores. Further, to our knowledge, aside from the original article describing the development and initial testing of the validity of the score, there has not been further evaluation into the reliability and utility of this scale in the hands of medical professionals.

Table 1. Engh Grading Scale score schema for the radiologic evaluation of uncemented total hip arthroplasty

	Score		
Scale	High	Undetermined	Low
Fixation			
Appearance of porous interface*	Extensive (≥ 50%) –5.0	0	None +5.0
Spot welds	Absent –2.5	0	Present +5.0
Score			(x/10.0)
Stability			
Appearance of smooth interface*	Extensive (≥ 50%) –3.5	0	None +5.0
Pedestal when end is unfixed	Present –3.5	0	Absent +2.5
Calcar modelling	Hypertrophy –4.0	0	Atrophy +3.0
Interface deterioration*	Present –2.5	0	Absent +2.5
Migration	Present –5.0	0	Absent +3.0
Particle shedding	Present –5.0	0	None +1.0
Score			(x/17)
Total score (fixation + stability)			x/27

*Lines/lucencies.

Considering the scale's prominent use in research for establishing the long-term success of THA, there is a lack of information on the reliability of the Engh Grading Scale and whether the specialty of the assessor influences reliability. Therefore, the primary purpose of our study was to determine the interrater reliability of the Engh scale among 4 evaluators: 2 arthroplasty surgeons (J.H. and D.B.C.J.) with more than 20 years of arthroplasty surgical experience, an experienced radiologist (M.A.J.) and a senior orthopedic resident (A.A.). The secondary purpose of the study was to determine the intrarater reliability. Interrater reliability was the primary outcome because research studies evaluating THA frequently involve multiple surgeons and/or centres, so it is important to know if different evaluators can reliably report radiographic findings. We hypothesized that the experienced arthroplasty surgeons would have the highest levels of interrater and intrarater reliability compared with the experienced radiologist and senior orthopedic resident.

Methods

The 4 evaluators mentioned above represent the typical specialists/trainees who would perform radiographic evaluations to assess the status of THAs. Two arthroplasty surgeons were selected to see if reliability was better between experienced orthopedic surgeons relative to the trainee or radiologist. The 4 evaluators participated in a training session before the commencement of data collection. The evaluators were all present during the training session to operationalize the defined categories of the Engh Grading Scale. The health region and hospital ethics review board approved our study protocol.

Each of the 4 evaluators assessed 26 plain radiographs of patients who had primary uncemented THAs. All patients had received the anatomic medullary locking (AML) femoral prosthesis (DePuy). Patients were at least 2 years postoperative, and the evaluators read the latest available

Table 2. Interrater reliability for the Engh Grading Scale using a weighted κ coefficient

	Assessor; κ (95% CI)		
Assessor	Arthroplasty surgeon 1	Arthroplasty surgeon 2	Radiologist
Orthopedic surgery resident	0.41 (0.17–0.65)	0.34 (0.17–0.52)	0.34 (0.12–0.56)
Arthroplasty surgeon 1		0.41 (0.15–0.66)	0.35 (0.17–0.53)
Arthroplasty surgeon 2			0.29 (0.10–0.48)

CI = confidence interval.

film. Two years was arbitrarily chosen as the minimum postoperative evaluation period because it allowed time for radiographic changes (e.g., ingrowth) to occur. Of the selected films, 7 (27%) were less than 5 years postoperative, 17 (65%) were between 5 and 10 years postoperative and 2 (8%) were more than 10 years postoperative. Further, 3 (12%) were prerevision films, where the revision surgery had been performed owing to aseptic loosening. The evaluators received no clinical information about either postoperative time period or need for revision surgery.

Films were selected based on having a good dispersion of postoperative time periods to evaluate and having a high-quality image with the full prosthesis clearly seen in the anteroposterior and lateral views.

All evaluation sessions were held independently to limit the potential for bias or influence between examiners. The evaluators were blinded to patient identity and surgeon. For evaluation purposes, the radiographs were assigned a number known only to the senior author (L.B.). The full set of 26 anteroposterior and lateral radiographs was used for evaluating intrarater and interrater reliability. Each examiner performed the review twice at 2 separate sessions with the films reviewed in random order. To compare interrater reliability, we used the first set of radiograph readings for all evaluators.

Statistical analysis

Reliability was calculated using the κ coefficient, as data were categorical. We calculated point estimates and 2-sided 95% confidence intervals (CIs) for each pairing of assessors using a weighted κ calculation on the total Engh scale score. Reliability values for the fixation and stability subscores could not be calculated as there was no established framework in the literature to categorize these values.

Landis and Koch proposed that the κ value can be interpreted using qualitative descriptors such that intraclass correlation values greater than 0.80 are "almost perfect," 0.61–0.80 "substantial," 0.41–0.60 "moderate," 0.21–0.40 "fair," 0.00–0.20 "slight" and values less than 0.00 are "poor."[7] We used these terms to describe the reliability coefficient values. Statistical data analyses were performed using SPSS version 17 (SPSS Inc.).

Results

Interrater reliability was fair to moderate for all pairings of the evaluators for the total Engh score (Table 2). Interestingly, there was not a large difference among the evaluators, although for the most part, the arthroplasty surgeons and the senior orthopedic resident achieved higher values in comparison with the radiologist. The 95% CIs were not precise, with the lower ends of all CIs in the "fair" range and only 2 of the upper ends of the CIs reaching the "substantial" rating.

In the intrarater evaluation, 1 arthroplasty surgeon achieved a "substantial" rating whereas the other arthroplasty surgeon achieved a "moderate" rating (Table 3). Both the senior orthopedic resident and the radiologist achieved a "fair" rating, with the lower ends of their 95% CIs in the "slight" or "poor" range (Table 3).

Discussion

To the best of our knowledge, this is the first reliability study of the Engh Scale. Although it is commonly reported in studies on long-term outcomes of uncemented hip arthroplasty, there have been very few evaluations of the scale. The original paper by Engh and colleagues[6] did not evaluate reliability, but did determine construct validity of the measure. The current study has demonstrated that interrater reliability was comparable among participants, regardless of skill or specialty, and rated a moderate level of reliability (κ 0.29–0.41) for all pairings. Intrarater reliability was dependent on the skill and specialty of the evaluator, with maximal values achieved for the arthroplasty surgeons ($\kappa = 0.52$ and $\kappa = 0.68$) and lowest values for the radiologist ($\kappa = 0.14$). As the scale had good intrarater reliability, particularly for the arthroplasty surgeons, our results indicate that experience reviewing THA radiographs does appear to improve reliability.

These results have implications for the serial evaluation of radiographs in the clinical setting and for research in single or multisite trials with different evaluators. Radiographic findings are an integral component of outcome assessment after THA surgery, so the process of evaluation should be a consideration when determining study methodology. Our results suggest that using the standardized Engh Grading Scale in its entirety as well as using their assigned grading system and a single, experienced arthroplasty surgeon evaluator would greatly improve the quality of reporting for radiographic findings in THA clinical trials.

It is important to note that reliability did achieve fair to moderate values among evaluators from different backgrounds. The consistency of ratings could be owing to the training session and setting operational definitions before data collection. If a single evaluator cannot be used in multicentre studies, investigators should at least consider discussing how the scale should be applied across evaluators to improve consistency of reporting. Caution should, however,

Table 3. Intrarater reliability for the Engh Grading Scale using a weighted κ coefficient

Assessor	Weighted κ (95% CI)
Arthroplasty surgeon 1	0.68 (0.38–0.98)
Arthroplasty surgeon 2	0.52 (0.39–0.64)
Orthopedic surgery resident	0.39 (0.13–0.66)
Radiologist	0.14 (0.00–0.36)

CI = confidence interval.

be used when interpreting radiographic results when multiple evaluators are used, and reports should include the experience and qualifications of the evaluators.

The strengths of this study include the training session, blinded evaluations and multiple examiners with different training backgrounds and years of experience. Use of different evaluators provides information about who should be chosen to rate radiographic findings in both the clinical and research setting. It appears that a senior orthopedic resident is already more consistent in reviewing THA radiographs and than an experienced radiologist. Interestingly, in a recent survey of orthopedic surgeons, respondents indicated a strong preference to review their own postoperative radiographs,[2] so this finding is likely not surprising to clinicians. Consideration perhaps should be given to training radiologists to review THA findings in a way that allows meaningful communication across disciplines.

Limitations

A limitation of this study is that no formal sample size calculation was performed and thus, the study may have been underpowered. The limiting factor for sample size was that we only used available high-quality films to evaluate the scale's performance under optimal circumstances. Reported CIs were wide, suggesting that the analysis was either underpowered or that there may have been substantial inherent variation in the measurement. Further evaluation of the scale may be warranted using larger quantities of radiographs and in other settings or centres.

Conclusion

This study has direct practical implications and clinical relevance. If the Engh Grading Scale is used for research purposes, the use of 1 evaluator with arthroplasty experience would be best; however, it may be acceptable to use multiple evaluators, particularly if they are experienced in reading postoperative THA radiographs. The performance of the scale in everyday situations may result in lower values of reliability and should not be the sole outcome measure to evaluate the stability of uncemented THA.

Competing interests: None declared.

Contributors: Drs. Al-Ahaideb, Huckell, Johnston and Beaupre designed the study. Drs. Al-Ahaideb, Huckell, Johnson and Beaupre acquired the data, which Drs. Muir, Johnston and Beaupre analyzed. Drs. Muir and Beaupre wrote the article, which Drs. Al-Ahaideb, Huckell, Johnson and Johnston reviewed. All authors approved its publication.

References

1. Hip and knee replacements in Canada — 2008–2009 Annual Report. Canadian Joint Replacement Registry (CJRR). Ottawa (ON): Canadian Institute of Health Information; 2009. Available: http://secure.cihi.ca/cihiweb/products/2008_cjrr_annual_report_en.pdf (accessed 2009 Dec. 17).
2. Teeny SM, York SC, Mesko JW, et al. Long-term follow-up care recommendations after total hip and knee arthroplasty. Results of the American Association of Hip and Knee Surgeons' Member Survey. *J Arthroplasty* 2003;18:954-62.
3. Clohisy JC, Kamath GV, Byrd GD, et al. Patient compliance with clinical follow-up after total joint arthroplasty. *J Bone Joint Surg Am* 2008;90:1848-54.
4. Beals RK, Tower SS. Periprosthetic fractures of the femur. *Clin Orthop Relat Res* 1996;327:238-46.
5. Sethuraman V, McGuigan J, Hozack WJ, et al. Routine follow-up office visits after total joint replacement. Do asymptomatic patients wish to comply? *J Arthroplasty* 2000;15:183-6.
6. Engh CA, Massin P, Suthers KE. Roentgenographic assessment of the biologic fixation of porous-surfaced femoral components. *Clin Orthop Relat Res* 1990;257:107-28.
7. Koepsell TD, Weiss NS. *Epidemiologic methods: studying the occurrence of disease*. New York (NY): Oxford University Press; 2003.

Le prix MacLean–Mueller

À l'attention des résidents et des directeurs des départements de chirurgie

Le *Journal canadien de chirurgie* offre chaque année un prix de 1000 $ pour le meilleur manuscrit rédigé par un résident ou un fellow canadien d'un programme de spécialité qui n'a pas terminé sa formation ou n'a pas accepté de poste d'enseignant. Le manuscrit primé au cours d'une année civile sera publié dans un des premiers numéros de l'année suivante et les autres manuscrits jugés publiables pourront paraître dans un numéro ultérieur du Journal.

Le résident devrait être le principal auteur du manuscrit, qui ne doit pas avoir été présenté ou publié ailleurs. Il faut le soumettre au *Journal canadien de chirurgie* au plus tard le 1er octobre, à l'attention du Dr G.L. Warnock, corédacteur, *Journal canadien de chirurgie*, Department of Surgery, UBC, 910 West 10th Ave., Vancouver BC V5Z 4E3.