OMERACT Hip Inflammation Magnetic Resonance Imaging Scoring System (HIMRISS) Assessment in Longitudinal Study

Jacob L. Jaremko, Robert G.W. Lambert, Susanne J. Pedersen, Ulrich Weber, Duncan Lindsay, Zeid Al-Ani, Kieran Steer, Marcus Pianta, Stephanie Wichuk, and Walter P. Maksymowych

ABSTRACT. ***Objective.*** To assess reliability, feasibility, and responsiveness of Hip Inflammation Magnetic resonance imaging Scoring System (HIMRISS) for bone marrow lesions (BML) in hip osteoarthritis (OA).

Methods. HIMRISS was scored by 8 readers in 360 hips of 90 patients imaged pre/post-hip steroid injection. Pre-scoring, new readers trained online to achieve intraclass correlation coefficient (ICC) > 0.80 versus experts.

Results. HIMRISS reliability was excellent for BML status (ICC 0.83–0.92). Despite small changes post-injection, reliability of BML change scores was high in femur (0.76–0.81) and moderate in acetabulum (0.42–0.56).

Conclusion. HIMRISS should be a priority for further assessment of hip BML in OA, and evaluated for use in other arthropathies. (First Release February 15 2019; J Rheumatol 2019;46:1239–42; doi:10.3899/jrheum.181043)

Key Indexing Terms:

HIP JOINT | OSTEOARTHRITIS | IMAGING
SCORING METHODS | OMERACT | BONE MARROW

Semiquantitative scoring of magnetic resonance imaging (MRI) features of arthritis offers an objective target for therapy. Per the Outcome Measures in Rheumatology (OMERACT) Filter 2.0[1,2,3], scoring systems should be carefully evaluated for reliability, feasibility, and discrimination. The Hip Inflammation MRI Scoring System (HIMRISS) evaluates markers of active hip inflammation including bone marrow lesion (BML)[4,5], recognizing that BML is increased T2/short-tau inversion recovery (STIR) signal intensity from inflammatory or noninflammatory processes[6]. In OMERACT 2016, a Web-based training module incorporating real-time iterative feedback calibration (RETIC) improved feasibility of HIMRISS scoring without sacrificing reliability[5]. Subsequent innovations in HIMRISS include Web-based interface and touch-sensitive electronic overlays to facilitate scoring. For OMERACT 2018, feasibility, reliability, and responsiveness of HIMRISS BML scoring were tested in a multireader scoring exercise on new prospectively obtained longitudinal data in patients receiving steroid injections for hip osteoarthritis (OA). The exercise was performed within the OMERACT MRI in Arthritis Working Group, from January to April 2018, and presented at OMERACT 14 (Terrigal, Australia, May 2018).

MATERIALS AND METHODS

HIMRISS. HIMRISS BML scoring has been described previously[4,6]. Within a Web-based interface (www.carearthritis.com; accounts free to registered users), a reader opens a coronal fluid-sensitive MRI sequence. The reader moves/resizes a semitransparent overlay (with adjustable opacity) to fit the femoral head on a reference slice, then scrolls through slices, identifying each region containing BML on each slice by touching or mouse-clicking on the overlay. This sets the score for that region to 1, while regions not clicked/touched by the user have a default score of 0 (no BML). A spread-

From the Department of Radiology and Diagnostic Imaging, University of Alberta; Department of Medicine, University of Alberta, Edmonton, Alberta, Canada; Rigshospitalet-Glostrup, Copenhagen Center for Arthritis Research, Center for Rheumatology and Spine Diseases, Copenhagen; King Christian 10th Hospital for Rheumatic Diseases, Gråsten; Institute of Regional Health Research, University of Southern Denmark, Odense, Denmark; St. Vincent's Hospital, Melbourne, Australia.

Supported by the Capital Health Chair in Diagnostic Imaging. Dr. Jaremko and Dr. Lambert are supported by Medical Imaging Consultants, Edmonton, Canada.

J.L. Jaremko, MD, PhD, FRCPC, Department of Radiology and Diagnostic Imaging, University of Alberta; R.G. Lambert, MB, FRCPC, Department of Radiology and Diagnostic Imaging, University of Alberta; S.J. Pedersen, MD, Rigshospitalet-Glostrup, Copenhagen Center for Arthritis Research, Center for Rheumatology and Spine Diseases; U. Weber, MD, Danish Hospital for Rheumatic Diseases, University Hospital of Southern Denmark, and Hospital of Southern Jutland, University Hospital of the Region of Southern Denmark, and Department of Regional Health Research, University of Southern Denmark; D. Lindsay, MD, Department of Radiology and Diagnostic Imaging, University of Alberta; Z. Al-Ani, MD, Department of Radiology and Diagnostic Imaging, University of Alberta; K. Steer, BSc, Department of Radiology and Diagnostic Imaging, University of Alberta; M. Pianta, MD, St. Vincent's Hospital; S. Wichuk, BSc, Department of Medicine, University of Alberta; W.P. Maksymowych, MB ChB, FRCP(C), FACP, Department of Medicine, University of Alberta.

Address correspondence to Dr. J.L. Jaremko, Radiologist and Associate Professor, Department of Radiology and Diagnostic Imaging, Faculty of Medicine, University of Alberta, 2A2.41 WMC, 8440–112 Street NW, Edmonton, Alberta T6G 2B7, Canada. E-mail: jjaremko@ualberta.ca

Accepted for publication December 5, 2018.

sheet file is automatically generated containing per-region, per-slice scores (0/1), and summary statistics. Fifteen 3-mm slices are scored, with 9 femoral head/3 acetabular regions for each of 5 middle slices, and 2 femoral head/2 acetabular regions for each of 5 anterior/5 posterior slices, for a total possible score of 65 (acetabulum) + 35 (femoral head) = 100.

Reader training. We trained readers using a previously described RETIC tool validated and reported for OMERACT[5] (www.carearthritis.com). New readers reviewed the HIMRISS instructional PowerPoint module, then scored up to 16 hips from 8 patients with hip OA scanned at 2 timepoints, for which consensus BML scores had been previously agreed by expert HIMRISS developers. RETIC feedback is provided immediately after scoring each case, by color-coding display of regions on the grid overlay superimposed on femur and acetabulum. This allows immediate review of regions concordant/discrepant with expert reader assessments. The intraclass correlation coefficient (ICC) value for status and change scores is provided after all cases have been assessed. This experiential calibration process with reader reliability targets (ICC for status/change score of > 0.8/> 0.6, respectively) enhances learning and reader satisfaction with the calibration exercise, ensuring training has been conducted in a standardized manner.

Reading exercise: data. In the University of Alberta Steroid Injection in Hip Osteoarthritis (STIHO) cohort, 97 adults with symptomatic hip OA presented to a radiology clinic for fluoroscopically guided steroid injections. With ethical approval (UofA HREB Pro00039139) and written informed consent, each subject underwent MRI of both hips, pre-injection and 8 weeks post-injection. Ninety complete datasets were available; 50/90 were male, age 59 ± 12.9 years (mean ± SD). Coronal STIR images were scored (repetition/echo/inversion times TR/TE/TI 4530/50/150 ms, matrix size 384 × 250, slice thickness 4 mm, field of view 350 × 350 mm). Scans from the 2 timepoints were read together for each subject, randomized and blinded as to which was the baseline scan.

Clinical data. Basic patient demographics (age, sex, symptom duration), baseline radiographic Kellgren-Lawrence (KL) OA grade (scored by a musculoskeletal radiologist with > 15 yrs of experience), and medication data were recorded (Supplementary Table 1, available with the online version of this article), along with baseline and 8-week posttreatment Western Ontario and McMaster Osteoarthritis Index (WOMAC) scores measuring pain, function, and stiffness related to the hip (Supplementary Table 2).

Readers. Eight readers included 3 musculoskeletal radiologists (of whom 2 were system developers with > 30 yrs and 15 yrs of experience), 1 rheumatologist developer, 2 musculoskeletal radiology fellows (6 yrs of experience each), and 2 rheumatologists.

Exercise design. Using the www.carearthritis.com Website, each reader scored the left and right hips for 90 subjects blinded-to-timepoint, i.e., n = 90 × 2 × 2 = 360 hips.

Statistical analysis. Descriptive statistics are expressed as mean ± SD. Interobserver intraclass correlation coefficients (ICC; single measure, absolute agreement, 2-way model) were assessed for all reader groups for BML status at baseline and interval change for the whole joint and for acetabular and femoral regions. Two-tailed Student t tests assessed whether change scores differed significantly from a mean of 0. Smallest detectable change (SDC) was calculated. For the statistical analysis, we used only the injected hip for each patient. When tests were repeated using both hips for each patient, considering injected and non-injected hips as separate data units, results were nearly identical and so these are not presented. Posthoc, we tested the effect on ICC of using only the central 5 slices of data to determine whether feasibility could be improved by reducing time for reading without substantial data loss.

RESULTS

Patients had a wide range of radiographic OA severity from KL grades 1–4. Pain, functional disability, and stiffness were substantial at baseline (mean WOMAC 44/100), as expected in subjects presenting for steroid injection, but there was only slight improvement at 8 weeks (mean WOMAC 37.7/100; Table 1).

At baseline, HIMRISS BML scoring reliability was excellent between the 2 expert readers (ICC 0.91 femoral head, 0.88 acetabulum), and although somewhat lower among the 6 other readers (0.73, 0.62), reliability was still excellent among all 8 readers (0.83, 0.83). Reliability declined slightly when only the central 5 slices were considered (all 8 readers: 0.82 femoral head, 0.76 acetabulum; Table 1).

Consistent with the small changes observed in WOMAC scores, the magnitude of BML change between baseline and 8 weeks postintraarticular steroid injection was small, averaging 2/65 at femoral head and < 1/35 in the acetabulum. Based on observed reliability, only 45% and 33% of femoral heads and acetabulae (respectively) showed change greater than the SDC (for femoral head SDC = 3.6/65, acetabulum SDC = 1.9/35; Table 2). Despite this, interobserver reliability for change remained high at the femoral head for experts (ICC 0.81) and all readers (ICC 0.76), and fair to moderate at the acetabulum (0.56, 0.42; Table 1). Reliability declined only slightly when only the central 5 slices were considered. These 5 slices contained most of the observed femoral head BML (mean BML score = 11.8 for central 5 slices vs 16.1 for all 15 slices) and half of the acetabular BML (5.3 vs 10.2).

The RETIC training (8 cases) required 2–6 hours for new users to complete. Then, scoring time in the exercise was 5–15 minutes per hip. Despite the lengthy reading task (360 hips), all readers completed the exercise and reader comments were highly positive regarding participation in future OMERACT scoring exercises.

DISCUSSION

In our study, we assessed feasibility, reliability, and discrimination of HIMRISS BML scoring in a multireader exercise on a large prospective longitudinal dataset. We observed excellent reliability, even among new readers, for BML status at baseline, and high reliability for detection of change in BML despite the small changes seen in this dataset, indicating high responsiveness. New reader ICC were only ~0.1–0.2 below that of experts on their first scoring exercise. These high levels of interobserver reliability are likely because of the insistence that new readers first achieve competence on training data using our interactive RETIC system.

In posthoc analysis, we found that simplifying scoring to only 5 central slices would have identified most of the total burden of BML (about 3/4 of femoral head BML, and about half of acetabular BML; Table 2), with only slightly decreased reliability for BML status and change, and a decrease in scoring time (removing 10/15 slices scored per hip), which may improve feasibility. However, it is not known whether restricting the number of slices scored would alter discrimination or effect-size of the tool because of the

Table 1. Interobserver reliability of HIMRISS BML scoring for baseline status and for interval change pre- versus 8 weeks post-injection. Values are mean (95% CI)

ICC for BML	All Readers, n = 8	Less Experienced Readers, n = 6	Experts, n = 2
Baseline			
Femoral head			
All 15 slices	0.83 (0.65–0.91)	0.73 (0.61–0.81)	0.91 (0.87–0.94)
Central 5 slices only	0.82 (0.65–0.90)	0.71 (0.59–0.80)	0.91 (0.87–0.94)
Acetabulum			
All 15 slices	0.83 (0.40–0.89)	0.62 (0.42–0.74)	0.88 (0.82–0.92)
Central 5 slices only	0.76 (0.66–0.85)	0.65 (0.51–0.75)	0.86 (0.79–0.90)
Total hip BML			
All 15 slices	0.83 (0.55–0.92)	0.71 (0.55–0.81)	0.92 (0.88–0.95)
Change (baseline to 8 weeks post-injection)			
Femoral head			
All 15 slices	0.76 (0.65–0.83)	0.72 (0.65–0.79)	0.81 (0.72–0.87)
Central 5 slices only	0.69 (0.56–0.78)	0.64 (0.55–0.72)	0.75 (0.64–0.83)
Acetabulum			
All 15 slices	0.42 (0.23–0.57)	0.36 (0.27–0.47)	0.56 (0.40–0.68)
Central 5 slices only	0.41 (0.22–0.57)	0.32 (0.23–0.43)	0.54 (0.37–0.67)
Total hip BML			
All slices	0.72 (0.60–0.81)	0.65 (0.57–0.73)	0.78 (0.68–0.85)

Data are presented for injected hips only; repeat analysis with both hips for each patient considered independent data units gave nearly identical results and is not presented here. HIMRISS: Hip Inflammation Magnetic resonance imaging Scoring System; BML: bone marrow lesion; ICC: intraclass correlation coefficient.

Table 2. Baseline values and observed changes in BML between baseline and 8 weeks postintraarticular steroid injection (n = 90 subjects) for all 8 readers.

Changes in BML	Slices Scored, n	Scoring Range	Baseline, Mean (SD)	Change, Mean (SD)	p	SDC	Proportion with Change > SDC
Femoral BML	15	0–65	18.9 (17.4)	2.2 (9.2)	0.02	3.6	44.6%
Femoral BML central only	5	0–45	14.2 (13.0)	1.3 (6.5)	0.06	3.0	41.1%
Acetabular BML	15	0–35	10.2 (7.0)	–0.08 (2.3)	0.73	1.9	33.3%
Acetabular BML central only	5	0–15	5.5 (3.9)	–0.05 (1.6)	0.75	1.4	25.6%
Total HIMRISS BML	15	0–100	29.1 (23.1)	2.1 (10.0)	0.05	4.6	35.6%

ML: bone marrow lesion; SDC: smallest detectable change; HIMRISS: Hip Inflammation Magnetic resonance imaging Scoring System.

limited temporal change in this dataset. We again noted lower reliability in the acetabulum than femur, as in our previous work[4,6], likely because of the complex acetabular shape and variation in red marrow distribution.

This study had limitations. Although HIMRISS was sufficiently reliable to detect small changes in BML (Table 2), discriminative capacity still cannot be assessed because of limited change in BML and self-reported clinical outcomes 8 weeks post-injection in this dataset. Intraarticular steroid injections have previously been shown to be significantly better than placebo for hip OA pain relief[7], and the magnitude of this effect is usually maximal between 2–8 weeks post-injection, declining rapidly after 3 months. It is unclear whether the timing of maximal temporal change on MRI corresponds to the timing of maximal therapeutic effect, thus the optimal timing of MRI is unknown and may vary according to the underlying condition and therapy.

Other limitations were that HIMRISS scoring considers only active lesions, not structural damage; and in a scoring exercise of this size (360 hips), it was not feasible to score the dataset twice, hence we could not test intraobserver reliability.

Overall, HIMRISS BML scoring is feasible and highly reliable when performed by readers trained using our RETIC interactive online calibration method. HIMRISS should be a priority for further assessment of hip BML in OA and evaluated for use in inflammatory hip arthropathies.

ACKNOWLEDGMENT

Thanks to Joanne McGoey for her assistance with study subjects, Joel Paschke for his computer programming skills, and Geoff Bostick and Linda Woodhouse for crucial support establishing the STIHO cohort.

ONLINE SUPPLEMENT

Supplementary material accompanies the online version of this article.

REFERENCES

1. Boers M, Kirwan JR, Wells G, Beaton D, Gossec L, d'Agostino MA, et al. Developing core outcome measurement sets for clinical trials: OMERACT filter 2.0. J Clin Epidemiol 2014;67:745-53.

2. Boers M, Kirwan JR, Gossec L, Conaghan PG, D'Agostino MA, Bingham CO 3rd, et al. How to choose core outcome measurement sets for clinical trials: OMERACT 11 approves filter 2.0. J Rheumatol 2014;41:1025-30.
3. Boers M, Kirwan JR, Tugwell P, Beaton D, Bingham CO III, Conaghan PG, et al. The OMERACT Handbook. [Internet. Accessed January 3, 2019.] Available from: https://omeract.org/resources
4. Maksymowych WP, Pitts M, Budak MJ, Gracey D, Lambert RG, McDougall D, et al. Development and preliminary validation of a digital overlay-based learning module for semiquantitative evaluation of magnetic resonance imaging lesions in osteoarthritis of the hip. J Rheumatol 2016;43:232-8.
5. Jaremko JL, Azmat O, Lambert RGW, Bird P, Haugen IK, Jans L, et al. Validation of a knowledge transfer tool according to the OMERACT filter: does Web-based real-time iterative calibration enhance the evaluation of bone marrow lesions in hip osteoarthritis? J Rheumatol 2017;44:1713-17.
6. Jaremko JL, Lambert RG, Zubler V, Weber U, Loeuille D, Roemer FW, et al. Methodologies for semiquantitative evaluation of hip osteoarthritis by magnetic resonance imaging: approaches based on the whole organ and focused on active lesions. J Rheumatol 2014;41:359-69.
7. Lambert RG, Hutchings EJ, Grace MG, Jhangri GS, Conner-Spady B, Maksymowych WP. Steroid injection for osteoarthritis of the hip: a randomized, double-blind, placebo-controlled trial. Arthritis Rheum 2007;56:2278-87.