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?

Jacob L. Jaremko, Omar Azmat, Robert G.W. Lambert, Paul Bird, Ida K. Haugen, Lennart Jans, Ulrich Weber, Naomi Winn, Veronika Zubler, and Walter P. Maksymowych

ABSTRACT. ***Objective.*** To assess reliability and feasibility of using a Web-based interface and interactive online calibration tool for magnetic resonance imaging (MRI) scoring of bone marrow lesions (BML) in osteoarthritis (OA), applied to the Hip MR Inflammation Scoring System (HIMRISS).

Methods. Seven readers new to HIMRISS (3 radiologists, 4 rheumatologists) scored coronal short-tau inversion recovery MRI from a hip OA observational study obtained pre- and 8-week post-steroid injection (n = 40 × 2 scans × 2 hips = 160 hips). By crossover design, Group B (4 readers) scored 20 patients (40 hips) using conventional spreadsheet-based methods and then another 20 using a Web-based interface and an online real-time iterative calibration (RETIC) training module. Group A (3 readers) reversed the order, scoring the first 20 subjects by the new method and the final 20 conventionally. Outcomes included ICC and reader survey.

Results. Interobserver reliability for BML status was high by both spreadsheet and Web-based methods (0.84–0.90), regardless of the order in which scoring was performed. Reliability of change scores was moderate and improved with training. Improvement was greater in readers who began with the spreadsheet method and then used the Web-based method than in those who began with the Web-based method, especially at the acetabulum. Readers found Web-based/RETIC scoring more user-friendly and nearly 50% faster than traditional spreadsheet methods.

Conclusion. HIMRISS offers reliable BML scoring in OA, whether by conventional spreadsheet-based scoring or by a Web-based interface with interactive feedback. The new method allowed faster readings, provided a consistent training environment that helped inexperienced readers achieve reliability equivalent to that of conventional methods, and was preferred by the readers. (First Release July 1 2017; J Rheumatol 2017;44:1713–17; doi:10.3899/jrheum.161101)

Key Indexing Terms:
HIP JOINT OSTEOARTHRITIS MAGNETIC RESONANCE IMAGING
SCORING METHODS OMERACT BONE MARROW

As options for osteoarthritis (OA) therapy emerge, objective outcome measures are increasingly needed to quantify disease status and treatment response. Magnetic resonance imaging (MRI)–based semiquantitative scoring systems assess hip OA by whole-organ approach [Hip Osteoarthritis MRI Scoring System (HOAMS[1]), Scoring Hip Osteoarthritis with MRI (SHOMRI[2])] or focus on active pathology including bone marrow lesions [BML; Hip MR Inflammation

From the Department of Radiology and Diagnostic Imaging, University of Alberta; Division of Rheumatology, Faculty of Medicine and Dentistry, University of Alberta, Edmonton, Alberta, Canada; Division of Medicine, University of New South Wales, New South Wales, Australia; Department of Rheumatology, Diakonhjemmet Hospital, Oslo, Norway; Department of Radiology and Medical Imaging, Ghent University Hospital, Ghent, Belgium; King Christian 10th Hospital for Rheumatic Diseases, Gråsten; Institute of Regional Health Research, University of Southern Denmark, Odense, Denmark; Department of Radiology, Robert Jones and Agnes Hunt Orthopaedic Hospital, Oswestry, UK; Department of Radiology, Balgrist University Hospital, Zurich, Switzerland.

Jacob Jaremko is the Capital Health Chair in Diagnostic Imaging at the University of Alberta; funds for this position derive from an endowment directed toward musculoskeletal imaging research.

J.L. Jaremko, MD, PhD, FRCPC, Department of Radiology and Diagnostic Imaging, University of Alberta; O. Azmat, MB, FRCP, Department of Radiology and Diagnostic Imaging, University of Alberta; R.G. Lambert, MB, FRCPC, Department of Radiology and Diagnostic Imaging, University of Alberta; P. Bird, MD, Division of Medicine, University of New South Wales; I.K. Haugen, MD, PhD, Department of Rheumatology, Diakonhjemmet Hospital; L. Jans, MD, PhD, Department of Radiology and Medical Imaging, Ghent University Hospital; U. Weber, MD, King Christian 10th Hospital for Rheumatic Diseases, and Institute of Regional Health Research, University of Southern Denmark; N. Winn, MBBS, FRCR, Department of Radiology, Robert Jones and Agnes Hunt Orthopaedic Hospital; V. Zubler, MD, Department of Radiology, Balgrist University Hospital; W.P. Maksymowych, MB ChB, FRCP(C), FACP, Division of Rheumatology, Faculty of Medicine and Dentistry, University of Alberta.

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

Accepted for publication May 9, 2017.

Scoring System (HIMRISS)[3,4]]. HIMRISS BML scoring differs from scoring in HOAMS and SHOMRI in that it is closer to quantitative scoring, involving binary decisions (BML present/absent, 1/0) in numerous small periarticular bone regions. In HOAMS or SHOMRI, scoring decisions assign one of multiple grades to features of arthropathy in fewer, larger 3-D regions, including estimates of percentages involved by BML. By whichever system, MRI-based BML scoring may be difficult in anatomically complex regions. Calibration tools are limited to published descriptions of these systems, and it is unclear to what degree acceptable reliability can be attained beyond the readers who developed these systems.

Experience from studies in rheumatoid arthritis has shown that reliability of semiquantitative MRI scoring is improved by systematic user training[5,6,7,8,9]. However, in-person training by experts is time-consuming and logistically difficult. Real-time iterative calibration with real-time feedback (RETIC) is a new concept that aims to enhance reader-expert calibration using a Web-based digital overlay superimposing outlines of scoring regions on MRI. Overlay color-coding gives immediate learning feedback comparing reader versus expert scores for each region.

Validation of this novel calibration technology was performed by a subgroup of the Outcome Measures in Rheumatology (OMERACT) MRI in Arthritis Working Group from January–April 2016 presented at OMERACT 13 (Whistler, British Columbia, Canada, May 2016). In accordance with the OMERACT handbook[10], no previous calibration tools were found in a literature review by a fellow in this group, which in agreement with the OMERACT executive committee included clinical professionals, methodologists, and healthcare professionals. We tested feasibility and interreader reliability using the relevant aspects of the OMERACT Filter 2.0[11,12] for inexperienced readers using this new tool versus traditional spreadsheet-based scoring.

MATERIALS AND METHODS

Interactive online interface. In OMERACT 12, HIMRISS readers preferred the use of a digital image overlay[3], with a total of 100 regions to score in 15 slices. We extended this concept to make the overlay touch- or click-sensitive within a Web-based interface. Readers upload or open an appropriate coronal MRI sequence in a Web browser at www.carearthritis.com (under "Osteoarthritis Imaging;" accounts free to registered users). The reader moves/resizes a transparent overlay to fit the femoral head on a reference scoring slice. Overlay gridlines may be adjusted from clearly visible to invisible using an onscreen opacity slider control so that actual image findings are not obscured. The reader scrolls through all slices, touching or mouse-clicking each overlay region containing BML. This causes shading to appear and the Web tool records a score of 1 to indicate it has been selected. A default score of 0 (no BML) is assumed; the reader clicks only on regions with BML. Upon scoring completion, the Web tool outputs a spreadsheet file containing per-region, per-slice scores (0/1) and summary statistics.

Use of the RETIC tool. For OMERACT 12, new reader training consisted of viewing instructional slides including a scoring atlas giving examples of true BML versus confounders including hematopoietic marrow. To improve on substantial limitations identified in that exercise[13], we added a scoring demonstration video (youtu.be/p2Mrfj2R9WM) and the new Web-based RETIC tool. In RETIC training mode, the reader scores cases previously scored by experts. When the reader has finished selecting positive regions, the overlay changes color in each region indicating whether reader versus expert scores are concordant/discordant. ICC between reader and experts are instantly updated (Figure 1). This allows real-time calibration with experienced readers to attain a prespecified acceptable target for reliability and rapid progressive learning with each case. For RETIC training, 8 cases (16 hips at 2 timepoints) from a previous study of hip steroid injection efficacy[14] were scored by 2 experienced HIMRISS developers, with discrepancies resolved by consensus.

Data. With University of Alberta Health Research Ethics Board approval and written informed consent, 97 adults with hip OA scheduled for fluoroscopically guided intraarticular steroid injection underwent MRI immediately pre-injection and 8 weeks post-injection. We used the first 40 consecutive subjects for whom complete data were available; 25/40 were men, mean age was 60 years (range 43–87), and mean body mass index was 29.5 kg/m^2 (range 18.8–44.3). We scored coronal short-tau inversion recovery (STIR) images (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). Left and right hips for each patient were scored separately at each timepoint, i.e., n = 160 hips. Anonymized STIR images for each subject were uploaded to www.carearthritis.com where each reader logged in for scoring. Once the reader selected the range of MRI slices containing the femoral head, the digital overlay template was applied automatically to images for readings. Readers were blinded to timepoint.

Readers. We had 7 readers: 3 musculoskeletal radiologists and 4 rheumatologists. Only 1 reader had previously used HIMRISS.

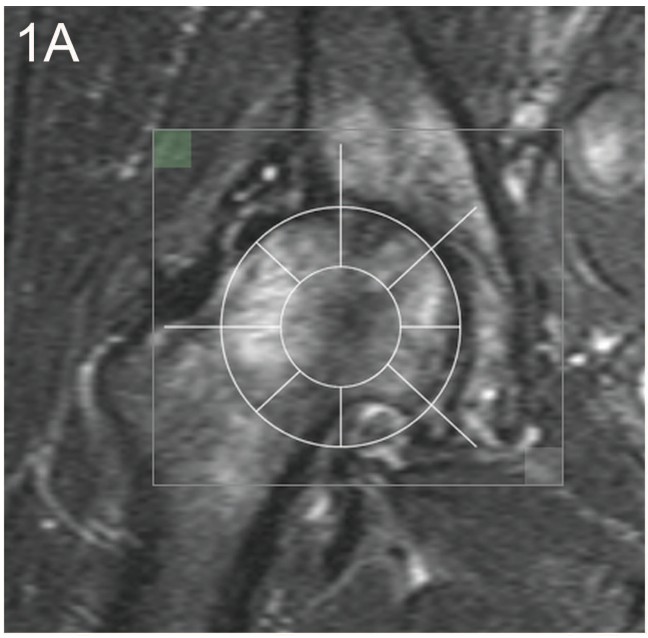
Exercise design. We wished to compare feasibility and reliability of scoring HIMRISS by conventional method (trained by reading a manuscript and slide presentation, manual spreadsheet score entry aided by physical printout of the hip grid overlay) versus the RETIC method (Web-based, touch-sensitive overlay, interactive calibration tool). This required each reader to score some cases by each method. To avoid learning bias, which could exist depending on which method was used first by each reader, after consultation within OMERACT we used a crossover design: readers were randomized into Group A (3 readers), who first scored cases 1–20 by the new method and then scored cases 21–40 by the conventional method, and Group B (4 readers), who first scored cases 1–20 by the conventional method and then scored cases 21–40 by the new method (Figure 2). Because our reader group included rheumatologists and radiologists with a wide range of training backgrounds, the crossover design helped control for variation in initial reader knowledge and experience.

Statistical analysis. Given the scoring range 0–100, interobserver ICC for BML status and change were calculated per reader pair and per reader group. We computed BML scores for the whole joint and for acetabular and femoral regions. We computed the smallest detectable change (SDC) as 1.96 × chi-square × standard error of mean.

RESULTS

Interobserver reliability was high for whole-joint BML status score by conventional or new methods, regardless of the order in which scoring was performed (ICC range among all readers 0.84–0.90; Table 1A and Table 1B). Reliability was lower in the acetabulum than femur (ICC range 0.76–0.86 vs 0.84–0.94).

Change [mean (SD, range)] in femoral, acetabular, and total BML score 8 weeks after steroid injection was 1.4 (7.7, –14 to 35), 0.4 (2.5, –5 to 11), and 1.8 (8.3, –12 to 36), respectively. Reliability of change scores was moderate but



1B

Timepoint	ICC
Timepoint A	.703
Timepoint B	.2

Compared to consensus expert read:

Agreement

False negative

False positive

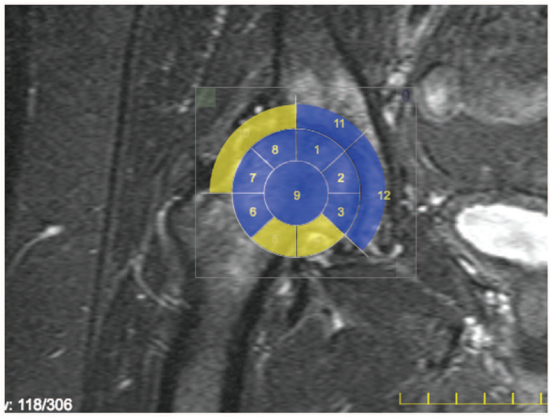


Figure 1. Real-time iterative calibration training in progress, shown in screen captures. (A) What the reader sees during scoring, i.e., the hip magnetic resonance imaging with a grid overlay consisting of thin lines where transparency can be varied. In this particular hip image, most sectors contain BML so the reader must click on most regions to score appropriately. (B) What the reader sees after scoring a training case. At the current slice, the regions the reader clicked on or touched to identify as containing BML agreed with the expert consensus in most regions (blue), but failed to identify the BML in 3 false-negative regions (yellow). The ICC between the reader and expert consensus for this case is displayed. There was also a timepoint B in this training case, in which the reader performed poorly (ICC 0.703 in timepoint A vs 0.200 in timepoint B); the reader will also review timepoint B before scoring the next case. This interactive training with rapid feedback is intended to make learning experiential. BML: bone marrow lesions.

improved with reader training in both groups (Table 1A and Table 1B). For change score in the acetabulum, a region that was difficult to score in our OMERACT 12 exercise[13], reliability was lower but improved more in reader Group B, who scored by conventional method (ICC 0.42) and later by new method (ICC 0.59). Reliability at the acetabulum was also more consistent between reader pairs when using the new method (ICC 0.38–0.67 vs 0.09–0.73). SDC in total BML was 5.0–7.0 depending on reader and scoring method. Only 16/40 hips showed change greater than the SDC.

RETIC training times averaged 10 min/hip × 8 hips, with wide user variation. In a postexercise survey, readers reported shorter scoring times for HIMRISS using the new method (3–12 min vs 5–20 min per hip, with hips containing little or no BML closer to 3–5 min and severe OA with extensive BML taking more time because of more mouse clicks). Six

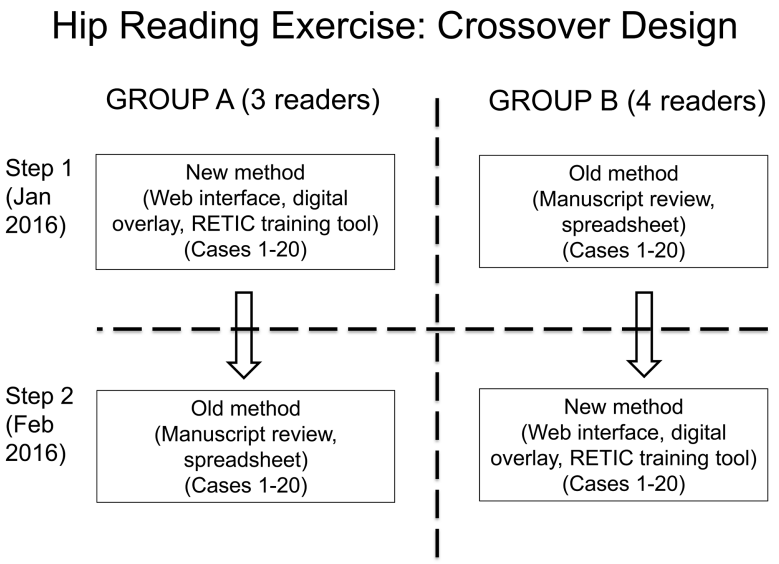


Figure 2. Crossover design for hip magnetic resonance imaging reading exercise, with the readers randomized into 2 groups based on which method was used to score cases first. RETIC: REal Time Iterative Calibration.

Table 1A. ICC scores for the hip scoring exercise for total joint, femur, and acetabulum. Group B (4 readers).

BML Site	First Exercise Conventional Method			Second Exercise RETIC Method		
	Pooled ICC, All Readers	95% CI for Pooled ICC	Range of ICC for Reader Pairs	Pooled ICC, All Readers	95% CI for Pooled ICC	Range of ICC for Reader Pairs
Acetabulum						
Status	0.79	0.68–0.87	0.66–0.95	0.76	0.64–0.85	0.71–0.83
Change	0.42	0.25–0.59	0.09–0.73	0.59	0.44–0.74	0.38–0.67
Femur						
Status	0.84	0.75–0.91	0.76–0.97	0.94	0.90–0.96	0.93–0.96
Change	0.79	0.69–0.87	0.68–0.92	0.78	0.67–0.86	0.73–0.87
Total						
Status	0.84	0.75–0.91	0.76–0.96	0.90	0.85–0.94	0.88–0.94
Change	0.69	0.56–0.80	0.49–0.85	0.76	0.64–0.85	0.66–0.87

BML: bone marrow lesion; RETIC: REal Time Iterative Calibration.

Table 1B. ICC scores for the hip scoring exercise for total joint, femur, and acetabulum. Group A (3 readers).

BML Site	First Exercise RETIC Method			Second Exercise Conventional Method		
	Pooled ICC, All Readers	95% CI for Pooled ICC	Range of ICC for Reader Pairs	Pooled ICC, All Readers	95% CI for Pooled ICC	Range of ICC for Reader Pairs
Acetabulum						
Status	0.83	0.67–0.91	0.77–0.87	0.86	0.77–0.92	0.81–0.92
Change	0.47	0.27–0.66	0.40–0.53	0.55	0.37–0.71	0.30–0.68
Femur						
Status	0.88	0.75–0.94	0.81–0.95	0.85	0.75–0.91	0.75–0.91
Change	0.51	0.31–0.69	0.34–0.69	0.68	0.53–0.81	0.64–0.74
Total						
Status	0.89	0.75–0.95	0.84–0.96	0.87	0.77–0.92	0.79–0.93
Change	0.54	0.33–0.71	0.39–0.70	0.67	0.51–0.79	0.60–0.73

BML: bone marrow lesion; RETIC: REal Time Iterative Calibration.

out of 9 readers found the new method "very user friendly" versus just 2/9 for the conventional method.

DISCUSSION

In our study, we compared feasibility and reliability of a new Web-based scoring platform and calibration tool with a conventional scoring approach to assess BML by HIMRISS method in hip OA. Interreader reliability for BML status was high and broadly similar whether readers learned and scored by conventional spreadsheet-based technique or by the new Web-based approach with RETIC interactive calibration. While the Web/RETIC method offered reliability similar to that of the conventional method, a key advantage was its feasibility; Web/RETIC scoring was substantially faster and was preferred by readers.

Our study had limitations. Whether by conventional or RETIC/online methods, HIMRISS scoring focuses on active lesions only and does not consider structural damage. Assessment for enhanced reliability postcalibration was compromised by high reliability on the first exercise, even for inexperienced readers. The crossover design may have resulted in a learning effect for both reader groups. Finally, to more completely assess the reliability of the method, further study is required in datasets showing substantial interval variation.

Overall, the use of a Web-based scoring interface with RETIC interactive calibration improved feasibility of HIMRISS scoring in terms of time, reader confidence, and satisfaction, while suggesting a possible advantage in anatomically challenging areas. The Web/RETIC approach could also apply to other image-based scoring systems. Further validation is warranted.

ACKNOWLEDGMENT

Thanks to Joanne McGoey for her invaluable assistance with study subjects, Stephanie Belton for her statistical expertise and long hours of analysis, and Joel Paschke for his computer programming skills, which were essential to development of the Web interface.

REFERENCES

1. Roemer FW, Hunter DJ, Winterstein A, Li L, Kim YJ, Cibere J, et al. Hip Osteoarthritis MRI Scoring System (HOAMS): reliability and associations with radiographic and clinical findings. Osteoarthritis Cartilage 2011;19:946-62.
2. Lee S, Nardo L, Kumar D, Wyatt CR, Souza RB, Lynch J, et al. Scoring hip osteoarthritis with MRI (SHOMRI): A whole joint osteoarthritis evaluation system. J Magn Reson Imaging 2015;41:1549-57.
3. 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.
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. van der Heijde D, Boers M, Lassere M. Methodological issues in radiographic scoring methods in rheumatoid arthritis. J Rheumatol 1999;26:726-30.
6. Bird P, Joshua F, Lassere M, Shnier R, Edmonds J. Training and calibration improve inter-reader reliability of joint damage assessment using magnetic resonance image scoring and computerized erosion volume measurement. J Rheumatol 2005;32:1452-8.
7. Østergaard M1, Edmonds J, McQueen F, Peterfy C, Lassere M, Ejbjerg B, et al. An introduction to the EULAR-OMERACT rheumatoid arthritis MRI reference image atlas. Ann Rheum Dis 2005;64 Suppl 1:i3-7.
8. Bird P, Conaghan P, Ejbjerg B, McQueen F, Lassere M, Peterfy C, et al. The development of the EULAR-OMERACT rheumatoid arthritis MRI reference image atlas. Ann Rheum Dis 2005;64 Suppl 1:i8-10.
9. Ejbjerg B, McQueen F, Lassere M, Haavardsholm E, Conaghan P, O'Connor P, et al. The EULAR-OMERACT rheumatoid arthritis MRI reference image atlas: the wrist joint. Ann Rheum Dis 2005;64 Suppl 1:i23-47.
10. Boers M, Kirwan JR, Tugwell P, Beaton D, Bingham CO III, Conaghan PG, et al. The OMERACT handbook. [Internet. Accessed May 9, 2017.] Available from: www.omeract.org/pdf/OMERACT_Handbook.pdf
11. Bellamy N. Clinimetric concepts in outcome assessment: the OMERACT filter. J Rheumatol 1999;26:948-50.
12. 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.
13. Jaremko JL, Pitts M, Maksymowych WP, Lambert RG. Development of image overlay and knowledge transfer module technologies aimed at enhancing feasibility and external validation of magnetic resonance imaging-based scoring systems. J Rheumatol 2016;43:223-31.
14. 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.