Assessing the Reliability of a Semiautomated Segmentation Algorithm for Quantifying Erosions in the Metacarpophalangeal Joints of Patients with Rheumatoid Arthritis

Michael A. Tomizza, Matthew A. Jessome, Joshua Barbosa, Karen A. Beattie, William G. Bensen, Raja S. Bobba, Alfred A. Cividino, Patrick D. Emond, Chris Gordon, Lawrence Hart, George Ioannidis, Melissa X.P. Koh, Maggie Larché, Ruben Tavares, Stephen Tytus, and Jonathan D. Adachi

ABSTRACT. ***Objective.*** Assess the reliability of early erosions in rheumatoid arthritis (EERA) software for quantifying erosive damage to the metacarpophalangeal joints of patients with rheumatoid arthritis (RA).

Methods. One hundred magnetic resonance image sets from 68 patients with early referral RA were evaluated. Reliability was assessed using 95% limits of agreement and intraclass correlation coefficient (ICC) with 95% CI.

Results. Limits of agreement linearly depended on erosion volume: 0.44× between readers and 0.19× within readers. Interrater ICC was 0.976 (95% CI 0.965–0.984) and intrarater ICC was 0.996 (95% CI 0.994–0.997).

Conclusion. EERA is highly reproducible for quantifying erosions in patients with early RA. (First Release July 15 2015; J Rheumatol 2015;42:1582–6; doi:10.3899/jrheum.141139)

Key Indexing Terms:
RHEUMATOID ARTHRITIS MAGNETIC RESONANCE IMAGING
REPRODUCIBILITY OF RESULTS JOINT EROSIONS METACARPOPHALANGEAL JOINT

Given the emphasis on early detection and monitoring of bone erosions in rheumatoid arthritis (RA), computerized methods for evaluating erosive damage have been developed to complete this task reliably and efficiently. One of these programs is early erosions in rheumatoid arthritis (EERA), a semiautomated segmentation algorithm that provides a fully quantitative measure of metacarpophalangeal (MCP) joint erosion volume in mm^3 using magnetic resonance imaging (MRI)[1]. Preliminary work suggests a very strong correlation between EERA and the manual segmentation of MR image sets, which are considered the gold standard[2]. Reproducibility was also measured, with intraclass correlation coefficients (ICC) exceeding 0.90[2]. While these findings suggest that EERA is highly reliable, a detailed analysis of reader agreement, including the limits of agreement, was not performed.

The objective of our study was to provide a more robust and clinically relevant analysis of the reliability of EERA by investigating the limits of agreement and ICC, expanding the breadth of the image sets assessed, and focusing on total erosive damage of the hand.

MATERIALS AND METHODS

Participants and image selection criteria. Ethics approval was obtained from The St. Joseph's Healthcare Hamilton Research Ethics Board. A 2009–2012 database was accessed, containing MR image sets of the hands of patients 18 years or older determined to satisfy the Emery, *et al* criteria for early referral to a rheumatologist[3]: at least 3 swollen joints; or a positive compression test of either the MCP or metatarsophalangeal joints; or at least 30 min of morning stiffness, lasting for at least 6 weeks. Image sets were included in the study if 2 readers, MK and JB, agreed that at least 1 erosion was present in MCP joints 2–5. An erosion was defined using the RA MRI Score (RAMRIS) definition as a sharply marginated bone lesion with correct juxtaarticular localization and typical signal characteristics[4]; the erosion

From the Department of Medical Sciences, Department of Medicine, and Faculty of Health Sciences, McMaster University; Department of Diagnostic Imaging, St. Joseph's Healthcare Hamilton, Hamilton, Ontario, Canada.

M.A. Tomizza, BSc, Department of Medical Sciences, McMaster University; M.A. Jessome, BHSc; J. Barbosa, BHSc, Faculty of Health Sciences, McMaster University; K.A. Beattie, PhD; W.G. Bensen, MD; R.S. Bobba, MD; A.A. Cividino, MD, Department of Medicine, McMaster University, and St. Joseph's Healthcare Hamilton; P.D. Emond, PhD; C. Gordon, PhD, St. Joseph's Healthcare Hamilton; L. Hart, MD, Department of Medicine, McMaster University, and St. Joseph's Healthcare Hamilton; G. Ioannidis, PhD, Department of Medicine, McMaster University, and St. Joseph's Healthcare Hamilton; M.X. Koh, BHSc, Faculty of Health Sciences, McMaster University; M. Larché, MD, Department of Medicine, McMaster University, and St. Joseph's Healthcare Hamilton; R. Tavares, PhD; S. Tytus, MD, St. Joseph's Healthcare Hamilton; J.D. Adachi, MD, Department of Medicine, McMaster University, and St. Joseph's Healthcare Hamilton.

Address correspondence to Dr. J.D. Adachi, Charlton Medical Centre, 501-25 Charlton Ave. E., Hamilton L8N 1Y2, Ontario, Canada. E-mail: jd.adachi@sympatico.ca

Accepted for publication May 13, 2015.

must also be visible in 3 consecutive 1-mm slices, as previously recommended for EERA[2]. Consistent with RAMRIS criteria, erosions in the first MCP joint were excluded because of unique anatomy[5]. Patients with a history of wrist or hand surgery were excluded. From a total of 108 available image sets, 100 fulfilled the eligibility criteria and were used in our study. Thirty-two of these image sets were previously analyzed using different readers and methodology[2].

MRI variables. MRI was performed using a 1T magnet and a 100-mm diameter cylindrical transmit and receive coil. A 3-D–spoiled gradient echo sequence was used in favor of the more conventional spin echo technique for the advantage of reduced slice thickness. Measures were identical to those originally described by Emond, *et al*[2].

Erosion segmentation. Two non-radiologist readers, JB and MK, were trained by the EERA developer PE through a 1-h instructional session, followed by erosion segmentation practice on 10 test image sets. To operate EERA, readers placed a "seed" at the erosion's geometric center and separately applied 5 different algorithm variable sets to iteratively stabilize the seed (Appendix 1). The variable set that the reader judged to best identify erosion boundaries was selected, and EERA computed erosion volume[2]. Apart from this training and an understanding of the RAMRIS erosion definition, the readers were otherwise unfamiliar with imaging measures of bone erosion in RA.

JB and MK independently evaluated the total erosion volume of each image that included MCP joints 2–5. Seventy-two h elapsed before evaluation of all image sets was repeated by MK. Both readers were blinded to other segmentation measurements and patient information.

Statistical analysis. Modified Bland-Altman plots were used to determine 95% limits of agreement[6]. Because initial plots illustrated that differences were proportional to the mean, the data were log-transformed, as recommended by Bland and Altman[6]. To assess interrater agreement, the difference between JB's and MK's erosion volume assessments at baseline divided by the mean of their measurements was plotted against the mean on a logarithmic scale. Intrarater agreement was similarly assessed, instead using the difference between MK's baseline and 72-h measurements. Interpretability was enhanced by expressing limits in their original units rather than as a ratio[7].

Inter- and intrarater reliabilities between readers and between time periods were determined by ICC(2,1) with 95% CI[8]. Total erosion volume measures were log-transformed to make within-person variance independent of the mean level[9]. Readers were assumed to be selected at random from a population of similar readers, and a 2-way ANOVA was applied. Statistical analyses were performed using SPSS software (Version 21.0, SPSS Inc.).

RESULTS

Participants and image sets. Sixty-eight participants contributed a total of 100 image sets; 54 image sets were of the right hand. Patient demographics, disease activity measures, and medications are detailed in Table 1.

Scoring comparisons. JB measured a total of 124 erosions, whereas MK measured a total of 121 erosions, with both readers identifying the same 118 erosions. The median (interquartile range) total erosion volume per image was 38.22 mm^3 (20.48–91.43) for JB and 35.16 mm^3 (20.54–88.42) for MK at baseline, and 35.54 mm^3 (19.85–88.42) for MK after 72 h. The inter- and intrarater 95% limits of agreement for the differences of total erosion volume were 0.44× and 0.19×, respectively. Bland-Altman plots illustrate reliability in Figure 1. Absolute error for ranges of erosion sizes are provided in Table 2. ICC for log-transformed data were excellent, with values of 0.976 (95% CI 0.965–0.984) for interrater reliability and 0.996 (95% CI 0.994–0.997) for intrarater reliability.

Table 1. Patient demographics, disease activity at the time of image acquisition, and medications administered at the time of image acquisition. From 68 participants, 43 had 1 image evaluated, 18 had 2 images evaluated, and 7 had 3 images evaluated, for a total of 100 images. Values are mean (SD) unless otherwise specified.

Characteristics	Values	Total Patients, n = 68
Demographics		
Female, n (%)	48 (70.6)	68*
White, n (%)	56 (83.6)	67*
Age, yrs**	57.4 (10.3)	66*
Weight, kg	79.8 (17.6)	63*
Height, cm	167.5 (9.7)	61*
Disease activity at time of image acquisition		Total Image Sets, n = 100
Symptom duration, yrs	4.8 (4.5)	97*
TJC28	6.7 (6.8)	91*
SJC28	7.4 (6.0)	91*
ESR, mm/h	18.1 (14.7)	85*
DAS28-ESR_{3V}	4.0 (1.5)	83*
HAQ-DI	0.64 (0.59)	58*
Medications at time of image acquisition, n (%)		
Oral steroid	53 (53)	100
OTC medication	83 (83)	100
DMARD	87 (87)	100

* Values are missing because of incomplete recording of information, laboratory tests not ordered, or incomplete questionnaire responses. **Age when first image was taken. All patients imaged more than once had their last image taken no less than 6 months before and no later than 24 months after their first image. TJC28: tender joint count at 28 joints; SJC28: swollen joint count at 28 joints; ESR: erythrocyte sedimentation rate; DAS28-ESR_{3V}: 3 variable Disease Activity Score (TJC, SJC, ESR); DAS28: DAS at 28 joints; HAQ-DI: Health Assessment Questionnaire–Disability Index; OTC: over-the-counter; DMARD: disease-modifying antirheumatic drugs.

DISCUSSION

The purpose of our study was to provide a more clinically relevant investigation of the reliability of EERA. This was first accomplished by summing the erosive damage of MCP joints 2–5, rather than evaluating joints individually, providing an outcome that more accurately identifies overall damage in the hand. Second, image eligibility was not restricted by erosion sizes. In a previous study, only image sets with erosions less than half the size of the metacarpal head were included[2]. However, it is important to understand how EERA responds to a variety of erosions found in the clinical spectrum to establish proper usage guidelines.

We found that limits of agreement varied with the estimated size of the erosion. The absolute reliability is best for smaller erosions, suggesting that EERA is well suited to the early RA population, where smaller erosions are most clinically relevant. This finding is partially explained by the original design of EERA, which was calibrated to evaluate smaller erosions expected in early disease[1]. Larger erosions are also more challenging to segment because they often lack

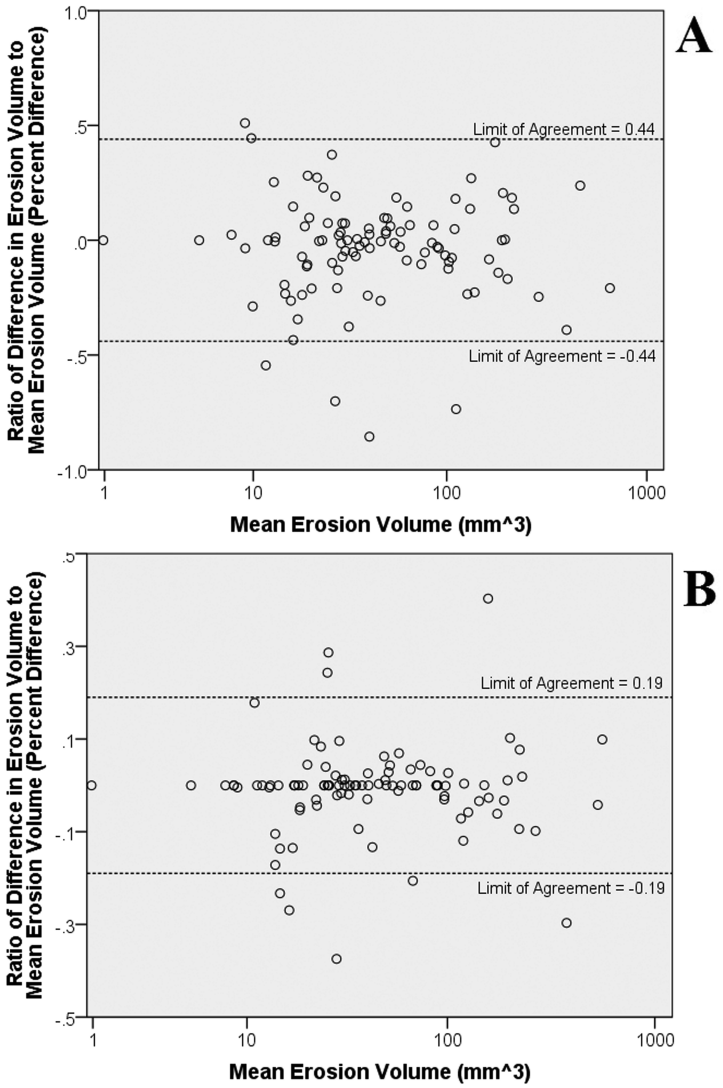

Figure 1. Modified Bland-Altman difference plots of total erosion volume with 95% limits of agreement calculated based on the method proposed by Euser, *et al*[7]. To illustrate the proportionality of the measurement differences to the mean measurement, the Y-axes represent the differences between readers JB and MK as a percent of their mean measurement (percent difference). The X-axes represent the mean measurement, on a logarithmic scale to condense the broad range of erosion sizes. A. For interrater agreement, percent differences in total erosion volume (JB subtracted by MK) are plotted against the logarithm of the mean of their measurements. The variability of the percent differences remains constant over all values of the mean, and the 95% limits of agreement are represented by the dashed lines. B. For intrarater agreement, the percent differences between MK's baseline and 72-h measurements (MK2 minus MK1) are plotted against the logarithm of the mean of the measurements. Intrarater limits of agreement are narrower, indicating better agreement in comparison to measurements between readers.

defined boundaries and may be composed of smaller, interconnected sub-erosions. Given the declining absolute agreement as erosive damage increases, the smallest detectable difference over time will likely be a function of baseline erosive damage.

Relative reliability was also assessed, with exceptional ICC exceeding 0.95 for both inter- and intrarater reliability, consistent with a previous report[2]. These results are comparable to figures reported by Poh, *et al*[10] using another computerized-assisted method for quantifying bone erosions, with EERA displaying higher interrater reliability (ICC 0.976 vs ICC 0.85). Correlations between EERA and RAMRIS

Table 2. Absolute limits of agreement for intra- and interrater reliability. Examples of the limits of agreement for small- to large-sized erosions are provided to illustrate how the limits of agreement, expressed as a volume, vary with the size of the erosion being evaluated.

Interval for Erosion Volume, mm^3	Interval for 95% Limits of Agreement, mm^3	
	Intrarater	Interrater
0–20	0–24	0–29
20–50	16–60	11–72
50–100	40–119	28–144
100–200	81–238	66–288
200–500	162–595	112–720

should be explored in the future, but are likely similar to the moderate correlation found in Poh, *et al*.

Collectively, these findings offer a novel contribution to the advancement of this software for clinical use in early RA. RAMRIS is currently the established method of assessing MRI bone erosions and also identifies synovitis and bone marrow edema, which EERA does not. However, for evaluating erosions, the semiquantitative nature of RAMRIS limits its precision for evaluating smaller erosions. Additionally, interrater ICC reported for RAMRIS range from 0.44–0.94[5,11,12,13,14,15,16,17] and RAMRIS must be used by a reader with considerable understanding of joint anatomy. EERA represents a practical alternative because it can easily be used by novice readers.

One limitation of our study is that the sample population was restricted to patients meeting early referral for RA criteria. However, EERA was designed for analysis of early-stage, small erosions because they hold the greatest implications for treatment initiation and prevention of subsequent damage. Second, only 1 reader completed the intrarater reliability phase of the study. Given the extremely high ICC found in our study and reported in previous assessments of EERA, the findings of the single reader are convincing, though examining EERA performance with more readers is warranted. Third, performing initial screening for erosions may have introduced bias in analysis; this effect is likely small, given the number of images evaluated. Finally, time constraints prevented interscan reliability assessment that helps estimate the error associated with scanning differences.

EERA is highly reliable for assessing erosive damage in the hands of patients with early RA. Its semiautomated, fully quantitative properties and suitability for novice readers make it attractive for use in the clinical setting. Further research assessing the validity, sensitivity to change, and responsiveness of EERA may allow for eventual implementation of the software into clinical practice.

ACKNOWLEDGMENT

Christine Fyfe and Caitlin Steven for their assistance with the study.

REFERENCES

1. Emond PD, Choi A, O'Neill J, Xie J, Adachi R, Gordon CL. The development of EERA: software for assessing rheumatic joint erosions. Can Assoc Radiol J 2009;60:63-8.
2. Emond PD, Inglis D, Choi A, Tricta J, Adachi JD, Gordon CL. Volume measurement of bone erosions in magnetic resonance images of patients with rheumatoid arthritis. Magn Reson Med 2012;67:814-23.
3. Emery P, Breedveld FC, Dougados M, Kalden JR, Schiff MH, Smolen JS. Early referral recommendation for newly diagnosed rheumatoid arthritis: evidence based development of a clinical guide. Ann Rheum Dis 2002;61:290-7.
4. Østergaard M, Peterfy C, Conaghan P, McQueen F, Bird P, Ejbjerg B, et al. OMERACT Rheumatoid Arthritis Magnetic Resonance Imaging Studies. Core set of MRI acquisitions, joint pathology definitions, and the OMERACT RA-MRI scoring system. J Rheumatol 2003;30:1385-6.
5. Lassere M, McQueen F, Østergaard M, Conaghan P, Shnier R, Peterfy C, et al. OMERACT Rheumatoid Arthritis Magnetic Resonance Imaging Studies. Exercise 3: an international multicenter reliability study using the RA-MRI Score. J Rheumatol 2003;30:1366-75.
6. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1:307-10.
7. Euser AM, Dekker FW, le Cessie S. A practical approach to Bland-Altman plots and variation coefficients for log transformed variables. J Clin Epidemiol 2008;61:978-82.
8. Weir J. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J Strength Cond Res 2005;19:231-40.
9. Bland JM, Altman DG. Transforming data. BMJ 1996;312:770.
10. Poh MQ, Lassere M, Bird P, Edmonds J. Reliability and longitudinal validity of computer-assisted methods for measuring joint damage progression in subjects with rheumatoid arthritis. J Rheumatol 2013;40:23-9.
11. Conaghan P, Lassere M, Østergaard M, Peterfy C, McQueen F, O'Connor P, et al. OMERACT Rheumatoid Arthritis Magnetic Resonance Imaging Studies. Exercise 4: an international multicenter longitudinal study using the RA-MRI Score. J Rheumatol 2003;30:1376-9.
12. Crowley AR, Dong J, McHaffie A, Clarke AW, Reeves Q, Williams M, et al. Measuring bone erosion and edema in rheumatoid arthritis: a comparison of manual segmentation and RAMRIS methods. J Magn Reson Imaging 2011;33:364-71.
13. Haavardsholm EA, Ostergaard M, Ejbjerg BJ, Kvan NP, Uhlig TA, Lilleås FG, et al. Reliability and sensitivity to change of the OMERACT rheumatoid arthritis magnetic resonance imaging score in a multireader, longitudinal setting. Arthritis Rheum 2005;52:3860-7.
14. McQueen F, Lassere M, Edmonds J, Conaghan P, Peterfy C, Bird P, et al. OMERACT Rheumatoid Arthritis Magnetic Resonance Imaging Studies. Summary of OMERACT 6 MR Imaging Module. J Rheumatol 2003;30:1387-92.
15. Ostergaard M, Klarlund M, Lassere M, Conaghan P, Peterfy C, McQueen F, et al. Interreader agreement in the assessment of magnetic resonance images of rheumatoid arthritis wrist and finger joints—an international multicenter study. J Rheumatol 2001;28:1143-50.
16. Ostergaard M, Conaghan PG, O'Connor P, Szkudlarek M, Klarlund M, Emery P, et al. Reducing invasiveness, duration, and cost of magnetic resonance imaging in rheumatoid arthritis by omitting intravenous contrast injection — Does it change the assessment of inflammatory and destructive joint changes by the OMERACT RAMRIS? J Rheumatol 2009;36:1806-10.
17. Bird P, Lassere M, Shnier R, Edmonds J. Computerized measurement of magnetic resonance imaging erosion volumes in patients with rheumatoid arthritis: a comparison with existing magnetic resonance imaging scoring systems and standard clinical outcome measures. Arthritis Rheum 2003;48:614-24.

APPENDIX 1. Erosion segmentation methodology.

To segment erosions using the Early Erosions in Rheumatoid Arthritis (EERA) software, a reader must first place a seed. The seed point serves to identify the erosion. Readers were instructed to place the seed point near the geometric center of the erosion, and then automatically re-run the seeding. Consecutively re-running the seed point allows the software to position the seed at the center of the preliminarily defined segmentation boundaries. The segmentation and re-running processes were repeated by the readers until successive seed positions were the same, indicating that a stable segmentation of erosion volume had been obtained. In the event that a seed would not stabilize, readers were instructed to place the seed point as close to the geometric center as possible and run the segmentation without re-seeding. In addition to the seed point, 15 scalar variables influencing the erosion mapping are defined in the quantification process. Allowing a reader to define each of the scalars maximizes the precision in erosion measures. However, how each variable changes the underlying hybridized region growing and level-set segmentation algorithm is not immediately apparent and requires a conceptual understanding of the mathematical construct behind the software. Thus, to simplify the quantification process, 5 sets of variables at fixed scalar values were made available to the readers. These were labeled A through E and are identical to the variable sets predetermined by Emond, *et al*[2]. In quantifying bone loss, a reader was to successively apply variable sets A through E to the eroded region, selecting the 1 that best visually identified the erosion boundary in all available images. Once a variable set is selected, EERA software determines a volume measure in all available image slices through a blocked construction method. In this segmentation technique, cross-sectional area identified in each 2-dimensional slice is multiplied by slice thickness. More detailed descriptions of EERA software are available from Emond, *et al*[1,2].