

The OMERACT Ultrasound Task Force — Advances and Priorities

MARIA-ANTONIETTA D'AGOSTINO, PHILIP G. CONAGHAN, ESPERANZA NAREDO, PHILIPPE AEGERTER, ANNAMARIA IAGNOCCO, JANE E. FREESTON, EMILIO FILIPPUCCI, INGRID MOLLER, CARLOS PINEDA, FREDERICK JOSHUA, MARINA BACKHAUS, HELEN I. KEEN, GURJEET KAELEY, HANS-RUDOLF ZISWEILER, WOLFGANG A. SCHMIDT, PETER V. BALINT, GEORGE A.W. BRUYN, SANDRINE JOUSSE-JOULIN, DAVID KANE, INGRID MOLLER, MARCIN SZKUDLAREK, LENE TERSLEV, and RICHARD J. WAKEFIELD

ABSTRACT. This article reports the most recent work of the OMERACT Ultrasound Task Force (post OMERACT 8) and highlights of future research priorities discussed at the OMERACT 9 meeting, Kananaskis, Canada, May 2008. Results of 3 studies were presented: (1) assessing intermachine reliability; (2) applying the scoring system developed in the hand to other joints most commonly affected in rheumatoid arthritis (RA); and (3) assessing interobserver reliability on a deep target joint (shoulder). Results demonstrated good intermachine reliability between multiple examiners, and good applicability of the scoring system for the hand on other joints (including shoulder). Study conclusions were discussed and a future research agenda was generated, notably the further development of a Global OMERACT Sonography Scoring (GLOSS) system in RA, emphasizing the importance of testing feasibility and added value over standard clinical variables. Future disease areas of importance to develop include a scoring system for enthesitis and osteoarthritis. (*J Rheumatol* 2009;36:1829–32; doi:10.3899/jrheum.090354)

Key Indexing Terms:
ULTRASOUND
SCORING SYSTEM

INTERMACHINE RELIABILITY
RESEARCH AGENDA

Musculoskeletal ultrasound (US) is an increasingly utilized method for the assessment and quantification of joint inflammation and damage in rheumatoid arthritis (RA) and other inflammatory arthritides. Despite its widespread use in daily practice for helping clinicians make decisions about patient care (e.g., change of diagnosis, monitoring of therapy efficacy or successful guidance of needles for aspiration or injection)^{1,2}, its use in research and therapeutic trials has been hampered by a perception of observer dependence and lack of validity.

PAST ACTIVITIES

In 2004 an international collaborative group of ultrasound experts was formed to address the metric qualities of US according to criteria specified by the OMERACT filter. They met for the first time in Asilomar, CA, USA, at OMERACT 7. A preliminary systematic review in the field of inflammatory arthritis on a joint-by-joint basis was presented during that meeting³. This work underlined the lack of published data about validity of US, especially for scan acquisition and definitions. Based on these results, a future

From the Université Paris Ouest-Versailles-Saint Quentin en Yvelines, Hôpital Ambroise Paré APHP, Paris, France; Academic Section of Musculoskeletal Disease, Leeds Institute of Molecular Medicine, University of Leeds, Leeds, UK; Severo Ochoa Hospital, Madrid, Spain; University La Sapienza, Rome, Italy; University of Ancona, Ancona, Italy; Hospital de Bellvitge, Barcelona, Spain; Instituto Nacional Rehabilitación, Mexico City, Mexico; University of New South Wales, Sydney, Australia; Humboldt University, Berlin, Germany; School of Medicine and Pharmacology, University of Western Australia; University Hospital, Bern, Switzerland; National Institute of Rheumatology and Physiotherapy, Budapest, Hungary; Medical Centre Leeuwarden, Leeuwarden, The Netherlands; Rheumatology Department, CHU Brest, Brest, France; Adelaide and Meath Hospital, Dublin, Ireland; University of Copenhagen Hvidovre Hospital, Copenhagen, Denmark; and The Parker Institute, Frederiksberg Hospital, Frederiksberg, Denmark.

M-A. D'Agostino, MD, PhD, Université Paris Ouest-Versailles-Saint Quentin en Yvelines, Hôpital Ambroise Paré APHP; P.G. Conaghan, MD, PhD, FRCP, Professor, Academic Section of Musculoskeletal Disease, Leeds Institute of Molecular Medicine, University of Leeds; E. Naredo, MD, Severo Ochoa Hospital; P. Aegerter, MD, PhD, Université Paris Ouest-Versailles-Saint Quentin en Yvelines, Hôpital Ambroise Paré

APHP; A. Iagnocco, MD, University La Sapienza; J.E. Freeston, MD, Academic Section of Musculoskeletal Disease, Leeds Institute of Molecular Medicine, University of Leeds; E. Filippucci, MD, University of Ancona; I. Moller, MD, Hospital de Bellvitge; C. Pineda, MD, Instituto Nacional Rehabilitación; F. Joshua, MD, PhD, University of NSW; M. Backhaus, MD, Humboldt University; H.I. Keen, MD, PhD, School of Medicine and Pharmacology, University of Western Australia; G. Kaeley, MD, Lakewood, Washington, USA; H-R. Zisweiler, MD, University Hospital; W.A. Schmidt, MD, Berlin-Buch, Berlin, Germany; P.V. Balint, MD, PhD, National Institute of Rheumatology and Physiotherapy; G.A.W. Bruyn, MD, PhD, Medical Centre Leeuwarden; S. Jousse-Joulin, MD, Rheumatology Department, CHU Brest; D. Kane, MD, Adelaide and Meath Hospital; I. Moller, MD; M. Szkudlarek, MD, PhD, University of Copenhagen Hvidovre Hospital; L. Terslev, MD, PhD, Parker Institute, Frederiksberg Hospital; R.J. Wakefield, MD, Academic Section of Musculoskeletal Disease, Leeds Institute of Molecular Medicine, University of Leeds.

Address reprint requests to Dr. M.A. D'Agostino, Université Versailles Saint Quentin en Yvelines, Service de Rhumatologie, Hôpital Ambroise Paré, 9, avenue Charles de Gaulle, 92100 Boulogne-Billancourt, France. E-mail: maria-antoinetta.dagostino@apr.ap-hop-paris.fr

research agenda was agreed on. The first step was to obtain preliminary consensus definitions for US-defined inflammatory pathology⁴. A prerequisite for longitudinal multicenter evaluation of responsiveness and predictive value of US is that the measure is reproducible. Until 2004, US studies have used joint pathology definitions, scanning techniques, and scoring systems developed within individual centers. Therefore, the evaluation of the reliability of the technique was considered a priority of our group. Subsequently a number of mainly European-based projects were undertaken to assess the reliability of US in inflammatory arthritis, but with a particular focus on RA.

At OMERACT 8, held in Malta in 2006, the group presented work undertaken post OMERACT 7. This included an overview of preliminary reliability exercises undertaken between individuals and groups for detecting pathological lesions (synovitis, erosions, tenosynovitis, and enthesopathy)⁵⁻⁷. Results of those exercises highlighted specific problems of standardization of the acquisition and interpretation of US images. Additional exercises undertaken in Paris (December 2004 to December 2005), which focused on metacarpophalangeal (MCP) joint synovitis in RA using the established OMERACT definition, were also presented. These iterative projects tested and retested interobserver and intraobserver reliability for both interpretation and acquisition of images of the MCP joint, including newly introduced semiquantitative grades⁸. During this exercise the same type of machine was used. Intra- and interobserver reliability for detecting and acquiring synovitis was assessed according to standard κ -coefficient and weighted κ -coefficient with absolute weighting [$\kappa(w)$] for semiquantitative grades. κ -coefficients were interpreted according to Landis and Koch⁹. Results confirmed that the OMERACT US definitions and consensus scoring system of synovitis combined with a standardized acquisition protocol provided good intra- and interobserver reliability (D'Agostino, *et al*, manuscript in preparation).

It was also apparent from this meeting that the following areas needed further development. First, to test intermachine reliability in particular for Doppler; second, to test whether a minimal core set of joints could replace multiple joint assessment in clinical practice and trials; and third, to assess the sensitivity to change of US-detected synovitis, and the predictive validity of US-detected synovitis with respect to disease-centered (e.g., erosions) and patient-centered (e.g., function) outcomes.

OMERACT 9 Ultrasound Special Interest Group session. At this session the results of activities post OMERACT 8 were presented and future research priorities were discussed (see below), and endorsed by OMERACT 9 participants.

1. Intermachine reliability exercise (Leeds, October 2007): In order to conclude the reliability exercises, we presented the results of an intermachine reliability study involving multiple observers. Financial support for this

study was obtained by the EULAR Standing Committee for International Clinical Studies Including Therapeutics (ESCISIT) with a view to producing EULAR/OMERACT recommendations.

The exercise focused on MCP joint synovitis and was conducted using the same methodology and protocol of previous Paris exercises. At the consensus meeting held the day before, it was decided to use the global scoring of synovitis tested in Paris 2005 as well as a scoring for all single components [power Doppler (PD) signal, effusion, and synovial hypertrophy]. Four different machine types were utilized, with the same sonographers who participated in the Paris exercises. US examinations were performed using each machine according to a Latin square design. Two rounds of scanning occurred in order to assess the intraobserver and intramachine reliability. The results varied according to the machine used. According to Landis and Koch⁹, 2 of the 4 machines appeared to provide more reproducible results: here κ -coefficients ranged from moderate (0.6) to good (0.8) for both single lesions and global scoring (0.8). For the other 2 machines results were mitigated: they ranged from poor (0.2) to moderate (0.5) for almost all of the lesions and scoring. The results of this exercise confirmed the good reliability among experts concerning the consensus of definitions and scoring system obtained in the last Paris exercise, especially when a particular machine was used¹⁰.

2. Testing the definition and scoring system of synovitis on other joints: The group's initial studies were based on the MCP joint, but the group decided to use the same methodology to assess other joints commonly involved in RA. In particular, the focus would be on those joints used in other disease activity scores (DAS, i.e., DAS28), and among these are some potential index joints such as the wrist. Some joints seem to be predictive of evolution and severity (metatarsophalangeals) and some joints are very important to the prognosis of disease; others are difficult to evaluate and often paucisymptomatic (shoulder).

(a) Shoulder reliability exercise: A shoulder study was conducted in Barcelona in June 2007 to compare US with magnetic resonance imaging for the detection of inflammatory and mechanical pathologies and to assess the intra- and interobserver reliability of US for detecting those lesions. Results of this exercise showed that the detection and scoring of inflammatory lesions such as synovitis and erosions were good, but underlined the difficulties for detecting and scoring mechanical-related abnormalities such as rotator cuff lesions¹¹.

(b) Interjoint reliability exercise: During the Leeds exercise (October 2007), an intra- and interobserver reliability exercise on static images was also performed to test the definition of synovitis and scoring systems for other joints. Joints proposed and included in this step were as follows: proximal interphalangeal, wrist, metatarsophalangeal, knee. A scoring system was developed for the MCP joint for both

single lesions (i.e., PD signal, effusion, and synovial hypertrophy) and global synovitis (i.e., PD + effusion + synovial hypertrophy). Interobserver reliability results presented ranged from good to excellent for both single lesions and global synovitis, with a high agreement (κ -coefficient superior to 0.7) for proposed global scoring for the MCP joint (D'Agostino, *et al*¹²).

3. Development of an US joint count: The prospect of developing an US joint count is attractive in that it might potentially be able to more objectively reflect a global level of synovitis and hence be more representative of disease activity in RA patients than conventional clinical measures. The results from a recently published multicenter longitudinal study were presented that compared an US 44-joint count with a number of reduced counts. It concluded that a 12-joint count was feasible and sensitive to change¹³. However, the validity of this reduced joint count remains to be confirmed in other multicenter longitudinal studies and particularly in an early RA cohort.

CURRENT ACTIVITIES

Current group work. Based on all the iterative exercises performed since 2004, the US group is currently developing OMERACT/EULAR recommendations for detecting and scoring synovitis in RA, including the production of an atlas with representative images for different joints.

Several published studies underlined the sensitivity of US above clinical examination^{1,2}. However, no consensual or validated scoring systems were used, making evaluation of results and comparison of studies difficult. A multicenter international study has, therefore, been planned to start at the end of 2008 in order to test the sensitivity to change of the OMERACT synovitis scoring system in multicenter clinical trials.

FUTURE ACTIVITIES

Agenda. Considering the greater sensitivity of US over clinical examination for detecting synovitis through the group discussions and feedback sessions, the concept of developing an US-based global scoring system in RA was felt to be important. Such a scoring system would be known as the RA GLOSS system. This global US scoring system at patient level would permit physicians to objectively follow patients under treatment in clinical and research practice by using a feasible and economical tool. However, emphasis was placed on investigating its added value over existing clinical measurements such as DAS28.

Interest in using US for evaluating other inflammatory or degenerative pathologies has also grown recently. Although the US features of peripheral joint pathology in spondyloarthritis (SpA) including psoriatic arthritis, osteoarthritis (OA), and crystal-related pathologies have been described, there is limited validity work and universally accepted semi-quantitative scoring systems for outcome assessment. As

with RA, a number of scoring methods have been described, with limited data on their psychometric properties. The US group has, therefore, decided to focus on standardization of enthesitis in SpA and joint-related pathologies in OA (including osteophyte and cartilage). A similar approach to the RA synovitis model will be used to assess other inflammatory and degenerative lesions seen on US. Preliminary work on standardizing US enthesitis has already been performed by some members of the group.

CONCLUSION

In summary, the group aims for continued improvement in RA outcome measurement using US. To date it has been successful in improving reliability of the technique and has helped in dispelling the myth that US is too subjective. The priority now is to test the credibility and responsiveness of this technique in the context of clinical trials. In particular, data about its added value over standard clinical measurements is crucial. Based on results of the exercises and discussions at OMERACT 9, future research directions of the group in RA include further validation of a global US scoring system at the patient level with particular reference to sensitivity to change and discrimination. At the same time a proposed atlas of a RA synovitis scoring system for helping ultrasonographers in their clinical daily practice is in preparation.

The growth of ultrasound as an outcome measure in other inflammatory diseases, such as SpA, provides scope for applying the rigorous OMERACT methodology to these conditions. Within the fields of SpA and OA, developing definitions of important US pathological lesions, as well as a standardized and reliable scoring system, will be the priority in the preparation for OMERACT 10.

REFERENCES

1. Karim Z, Wakefield RJ, Conaghan PG, et al. The impact of ultrasonography on diagnosis and management of patients with musculoskeletal conditions. *Arthritis Rheum* 2001;44:2932-3.
2. D'Agostino MA, Ayril X, Baron G, Ravaud P, Breban M, Dougados M. Impact of ultrasound imaging on local corticosteroid injections of symptomatic ankle, hind- and mid-foot in chronic inflammatory diseases. *Arthritis Rheum* 2005;53:284-92.
3. Joshua F, Lassere M, Bruyn GA, et al. Summary findings of a systematic review of the ultrasound assessment of synovitis. *J Rheumatol* 2007;34:839-47.
4. Wakefield RJ, Balint PV, Szekulnarek M, et al. Musculoskeletal ultrasound including definitions for ultrasonographic pathology. *J Rheumatol* 2005;32:2485-7.
5. Wakefield RJ, D'Agostino MA, Iagnocco A, et al. The OMERACT Ultrasound Group: status of current activities and research directions. *J Rheumatol* 2007;34:848-51.
6. Scheel AK, Schmidt WA, Hermann KG, et al. Interobserver reliability of rheumatologists performing musculoskeletal ultrasonography: results from a EULAR "Train the trainers" course. *Ann Rheum Dis* 2005;64:1043-9.
7. Naredo E, Möller I, Moragues C, et al. EULAR Working Group for Musculoskeletal Ultrasound. Interobserver reliability in musculoskeletal ultrasonography: results from a "Teach the

- Teachers" rheumatologist course. *Ann Rheum Dis* 2006;65:14-9.
8. D'Agostino MA, Wakefield R, Filippucci E, et al. Intra- and inter-observer reliability of ultrasonography for detecting and scoring synovitis in rheumatoid arthritis: A report of a EULAR ESCISIT Task Force [abstract]. *Ann Rheum Dis* 2005;64:62.
 9. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159-74.
 10. Wakefield R, D'Agostino MA, Backhaus M, et al. Combined evaluation of influence of sonographer and machine type on the reliability of power Doppler ultrasonography (PDUS) for detecting, scoring and scanning synovitis in rheumatoid arthritis (RA) patients: results of an intermachine reliability exercise of the EULAR/OMERACT US group. *Ann Rheum Dis* 2008;67:421.
 11. Bruyn GA, Naredo E, Moller I, et al. Reliability of ultrasonography in detecting shoulder disease in patients with rheumatoid arthritis. *Ann Rheum Dis* 2009;68:357-61.
 12. D'Agostino MA, Wakefield R, Backhaus M, et al. Evaluation of the intra- and inter-reader reliability of power Doppler ultrasonography (PDUS) for detecting and scoring synovitis of several joints in rheumatoid arthritis (RA) by using a validated scoring system: a report of the EULAR/OMERACT US group [abstract]. *Ann Rheum Dis* 2008;67:422.
 13. Naredo E, Rodríguez M, Campos C, et al. Validity, reproducibility, and responsiveness of a twelve-joint simplified power doppler ultrasonographic assessment of joint inflammation in rheumatoid arthritis. *Arthritis Rheum* 2008;59:515-22.