



UNIVERSITY OF LEEDS

This is a repository copy of *The OMERACT Stepwise Approach to Select and Develop Imaging Outcome Measurement Instruments: The Musculoskeletal Ultrasound Example*.

White Rose Research Online URL for this paper:
<http://eprints.whiterose.ac.uk/142903/>

Version: Accepted Version

Article:

Terslev, L, Naredo, E, Keen, HI et al. (8 more authors) (2019) The OMERACT Stepwise Approach to Select and Develop Imaging Outcome Measurement Instruments: The Musculoskeletal Ultrasound Example. *Journal of Rheumatology*, 46 (10). pp. 1394-1400. ISSN 0315-162X

<https://doi.org/10.3899/jrheum.181158>

© 2019 The Journal of Rheumatology. This is a pre-copyediting, author-produced PDF of an article accepted for publication in *The Journal of Rheumatology* following peer review. The definitive publisher-authenticated version: Terslev, L, Naredo, E, Keen, HI et al. (8 more authors) (2019) *The OMERACT Stepwise Approach to Select and Develop Imaging Outcome Measurement Instruments: The Musculoskeletal Ultrasound Example*. *Journal of Rheumatology*, 46 (10). pp. 1394-1400. ISSN 0315-162X, is available online at: <https://doi.org/10.3899/jrheum.181158>.

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



eprints@whiterose.ac.uk
<https://eprints.whiterose.ac.uk/>

The OMERACT stepwise approach to select and develop imaging outcome measurement instruments: the musculoskeletal ultrasound example.

Authors: L Terslev¹, E Naredo², HI Keen³, GAW Bruyn⁴, A Iagnocco⁵, R Wakefield⁶, PG Conaghan⁶, LJ Maxwell⁷, DE Beaton⁸, M Boers⁹, and MA D'Agostino¹⁰

Affiliations:

1. Copenhagen Center for Arthritis Research, Center for Rheumatology and Spine Diseases, Rigshospitalet, Copenhagen, Denmark
2. Department of Rheumatology, Bone and Joint Research Unit. Hospital Universitario Fundación Jiménez Díaz, IIS Fundación Jiménez Díaz, and Universidad Autónoma de Madrid. Madrid, Spain
3. Department of Rheumatology, University of Perth, Perth, Australia
4. Department of Rheumatology, MC Groep hospitals, Lelystad, the Netherlands
5. DSCB Università degli Studi di Torino, MFRU Città della Salute e della Scienza, Turin - Italy
6. Leeds Institute of Rheumatic and Musculoskeletal Medicine, University of Leeds, and NIHR Leeds Biomedical Research Centre, Leeds UK
7. University of Ottawa and Centre for Practice-Changing Research, Ottawa Hospital Research Institute, Ottawa, Canada
8. Institute for Work & Health and Institute for Health Policy Management and Evaluation, University of Toronto, Toronto ON Canada
9. Department of Epidemiology and Biostatistics; and Amsterdam Rheumatology and immunology center; Amsterdam UMC, Vrije Universiteit, Amsterdam The Netherlands
10. Rheumatology Department, APHP, Hôpital Ambroise Paré, Boulogne-Billancourt, INSERM U1173, Labex Inflammex, Université Versailles St-Quentin en Yvelines, Montigny Les Bretonneux, France.

Corresponding author: Prof. Maria Antonietta D'Agostino, MD, PhD,
Rheumatology department, Ambroise Paré Hospital, 9 avenue Charles de Gaulle, 92100 Boulogne-Billancourt, FRANCE

Email: maria-antonietta.dagostino@apr.aphp.fr

Words count:

Abstract: 250

Manuscript: 2414

Keywords: Ultrasound, validation, OMERACT, OFISA, outcome measurement instrument

Abstract

Aim

To describe the OMERACT stepwise approach to select and develop an imaging instrument with musculoskeletal ultrasound (US) as example.

Method

The OMERACT US working group (WG) developed a 4-step process to select instruments based on imaging. Step 1 applies the OMERACT Framework Instrument Selection Algorithm (OFISA) to existing outcome measurement instruments that use US for a specific indication. This step requires a literature review focused on truth, discrimination and feasibility aspects of the instrument for the target pathology. When the evidence is completely unsatisfactory, Step 2 is a consensus process to define the US characteristics of the target pathology including one or more so-called 'elementary lesions'. Step 3 applies the agreed definitions to the image, evaluates their reliability, develops a severity grading of the lesion(s) at a given anatomical site and evaluates the impact of the acquisition technique on feasibility and lesion(s) detection. Step 4 applies and assesses the definition(s) and scoring system(s) in cross-sectional studies and multicenter trials. The imaging instrument is now ready to pass a final OFISA check.

Results

With this process in place, the US WG now has 18 subgroups developing US instruments in 10 different diseases. Half of them have passed step 3, and the groups for enthesitis (spondyloarthritis, psoriatic arthritis), synovitis and tenosynovitis (rheumatoid arthritis) have finished step 4.

Conclusion

The US WG approach to select and develop outcome measurement instruments based on imaging has been repeatedly and successfully applied in US but is generic for imaging and fits with OMERACT Filter 2.1.

Introduction

The Outcome Measures in Rheumatology (OMERACT) initiative works to develop core outcome sets for trials and observational studies in rheumatology and provides guidelines for the development and validation of outcome measurement instruments for use in clinical research. This ensures valid and comparable results between trials and benefits the clinical decision makers.

The development of core sets consists of decisions on *what* to measure, termed “core domains” and then decisions about *how* to measure each of the chosen domains, by selecting (or developing) at least one instrument for each domain. According to the OMERACT Filter 2.1, for a health condition the domains of interest should be selected within 4 specified “core” areas, now termed: manifestations/abnormalities; life impact, death/lifespan, societal/resource use. “How to measure” a specific domain implies selecting measurement instruments (1-3).

OMERACT has developed a methodology for selecting instruments, the OMERACT Instrument Selection Algorithm (OFISA)(4). Whatever the instrument (i.e. questionnaire, a score obtained through physical examination, a laboratory measurement, a score obtained through observation of an image etc.), the selection should follow the same rigorous process, including the assessment of its metric properties. OFISA uses 4 signaling questions to help evaluate the existing evidence. These questions are based on the 3 pillars of the original OMERACT filter: Truth, Discrimination and Feasibility (5). Therefore, an outcome measurement instrument must be truthful, discriminate between situations of interest and be feasible in the context of clinical trials (5,6). The OFISA is based primarily on a deep evaluation of the existing literature on the target instrument and a careful analysis of all validation studies. Responses to the OFISA evaluation questions are rated (and color-coded) and then combined into an overall rating for the validity of the instrument. “Red” always means ‘stop, do not continue’, “Amber” means ‘a caution is raised but you can continue’ (and a research agenda is needed), “Green” means ‘go, this question is definitely answered affirmatively’, and “White” indicates an absence of evidence, where the working group has to choose between discarding the instrument or creating the necessary evidence. This methodology works well for tools such as questionnaires, clinical composite scores, “linear” instruments (biological assays) etc., but needs elaboration for the selection of imaging instruments.

Imaging is a rapidly evolving field within medicine and imaging techniques usually enter clinical practice before a full evaluation of their measurement properties has been performed. Literature assessing the metric qualities is often scarce or mostly focused on evaluating the capability of the technique to show pathological findings (against other imaging techniques used as gold standards). These “validation studies” usually apply an “ad-hoc score” to the images obtained, and are often performed in one center only. Like other instruments, an imaging outcome measurement instrument comprises not only the technique, but

also the scoring system for the lesions, so the validity of the technique and the score should be tested in the intended setting.

One of the main challenges related to imaging is the complex relationship between the technical characteristics of the imaging device, the setting in which it is applied, and the interpretation of the acquired data. These interactions generate variability which needs to be accounted for before any scoring system based on the technique can be accepted as outcome measurement instrument. In addition, some imaging techniques, such as ultrasound (US) and magnetic resonance imaging (MRI), present additional sources of variability related to the concomitant image acquisition, including patient positioning and slice thickness for MRI or positioning of the probe for US, the level of training of the operator, agreed definition(s) of what should be measured and grading of severity of the studied lesion(s). To date, these key additional sources of variability have not been fully described in OFISA, and in the OMERACT Filter 2.1, (7) and have rarely been evaluated in existing imaging instruments. Thus, the OFISA appraisal of measurement properties often ends with “White” responses (i.e. complete absence of evidence or absence of studies addressing the technical validity in a degree that prevents making conclusions about the proposed instrument), which would lead to “Red” or, in a better case, to “Amber” for the whole instrument. To date, within OMERACT most instruments based on imaging have had to be developed “from scratch”, with little or no guidance on how to develop such instruments and how to build the evidence needed for an OMERACT endorsement.

The OMERACT US working group (WG) was established in 2004 with the aim to validate US-based outcome measurement instruments for rheumatic diseases. (8,9) This paper describes the original US WG stepwise approach to select and develop US instruments to pass OFISA, which is applicable across all imaging techniques.

Procedure

Under OMERACT filter 2.1, the domains of interest of US-based instruments belong to the “manifestations/abnormalities” core area, in particular “disease activity” and “structural damage” (2-4,7). The validation process follows four steps of appraising evidence, or, when necessary, developing and creating evidence (Figure 1). The movement from one step to the next is dependent on the level of success with that step.

The first step — **Step 1** — is to perform a systematic literature review following OFISA recommendations. The review serves several purposes to verify whether a US-based instrument for the topic of interest fulfills the OMERACT pillars of Truth, Discrimination and Feasibility. Truth covers face, content and construct

validity. Face validity is credibility – i.e. whether an instrument appears to measure what it is supposed to; whereas content validity is comprehensiveness - i.e. whether an instrument covers all aspects of the attribute to be measured. Face and content validities are essentially subjective (i.e. US provides good image quality and spatial resolution of a joint and its components). Construct validity is the consistency with theoretic concepts (for example that a US instrument of synovitis is related to other measures of synovitis). Discrimination requires that the instrument can detect clinically important degrees of change — or lack of change — including variation over time (longitudinal construct validity) with enough reproducibility, estimates of test-retest reliability, and differences in change between groups. Thresholds considered to be clinically meaningful (i.e. minimal degree of synovitis) are also defined under discrimination. Feasibility relates to the interpretability of the measurement result in terms of suitable time, monetary costs and patient acceptability. For an imaging technique the interpretability of the instrument is a key part of the instrument application. Observers possess different cognitive, visual, and perceptual abilities. To understand the performance of an imaging instrument, it is important to assess all critical components including the observers (10). Therefore, the first purpose of the literature review is to evaluate the presence of agreed definitions of pathology (i.e. “theoretical” or conceptual definition(s)) and related “elementary lesions” (11), taking into account both i) the impact of equipment used on feasibility and quality of visualization of the tissues under study, and ii) the interpretation made by the observer. The concept of “elementary lesion” refers to the individual imaging characteristics of the pathophysiological manifestation(s) under study (e.g. synovial hypertrophy and abnormal flow detected by Doppler mode are the “elementary lesions” that taken together constitute US-detected synovitis), where “theoretical” or conceptual definition indicates the US appearance of the pathology under study. The second purpose is to verify that the published US instruments can pass OFISA based on their application in randomized clinical trials or observational studies of sufficient quality. A standardized template has been specifically designed to extract and collect US data (8). However, as there is often a lack of agreement of US definitions applied in the literature for elementary lesions or disease pathologies, or a lack of good reliability studies, the second purpose of step 1 is almost never achieved and additional steps are needed to check technical evidence, define, and build clinical evidence needed for OFISA. Therefore, the instrument needs to go through additional steps (i.e. development steps).

In **Step 2** the group proceeds to develop a new US instrument by developing new or better definitions of elementary lesions for a defined pathology. The definitions are usually obtained through a Delphi process that combines data from the literature review with expert opinion. So-called ‘*theoretical or conceptual definitions*’ can be developed to describe the US aspect of the whole pathophysiological manifestation under study, e.g. US-detected synovitis, whereas ‘*operational definitions*’ are developed to describe the

single aspects, i.e. the “elementary lesions” measurable by US (i.e. the US aspect of a “synovial inflammation” which can be detected by the combined or isolated use of grey-scale and Doppler techniques, or, for analogy, in a MRI setting, the use of gadolinium enhanced T1 sequences instead of T2 weighted sequences for measuring inflammation). The proposed definitions are circulated among interested WG members, usually considered US experts in the chosen field, who then indicate their agreement with the proposals on a 0-5 scale, and can suggest modifications. Consensus is reached when the definition achieves >75% agreement of scores greater than 3 (where 3 means neutral or minimal agreement). Reaching consensus usually takes several rounds.

Step 3 is an iterative procedure aimed at:

- a) Testing the sonographers’ reliability to detect the pathology and their constituent elementary lesions when they apply the agreed definitions;
- b) Developing a grading of severity of the pathology at site level (i.e. site-level scoring system); and
- c) Evaluating the reliability of the scanning technique (e.g. acquisition of the information) independently of the US device used and the anatomical site to which the definition is applied.

Reliability is first assessed on static images with representative and clear pathology according to the definitions. Images collected among participants are used to create a web-based exercise. A set of the images is shown twice in random order to assess intra-observer reliability. The static image exercise may be followed by an additional test of the definitions on a video-clip exercise or directly followed by a patient-based exercise (i.e. patients with the disease entity in which US is being validated as an outcome measurement instrument and who potentially may have the lesion(s) of interest). The operational definition that moves forward is the one with high enough inter-observer reliability.

In step 3, the development of a scoring system - grading the severity of the lesion(s)- is developed at site level, with subsequent assessment of inter-and intra-observer reliability, and a sum scores for all sites, at patient level can be proposed. Finally, step 3 also assesses the inter-and intra-observer reliability of the definitions but now with the variation introduced by the acquisition technique . If (as usual) the reliability of the acquisition involves more sites and different US machines, the interaction of these 3 aspects (device, observer, site) on the reliability of the definition(s) of lesions and/or on scoring system(s) is also evaluated. Since most grading systems are semi-quantitative, reliability is preferably analyzed by kappa statistics (12-14). Additional statistical methods such as variance component analysis or generalizability theory permit a

multifaceted perspective on measurement error and its components (15). The procedure is usually iterative, with the possibility to improve definitions and standardize procedures.

Step 4. In this step, the body of evidence needed for a full Filter 2.1 endorsement is created and gathered. This include validity (cross-sectional construct) of the technique compared to other indicators of the same target lesion (i.e., histological findings, findings confirmed on other imaging techniques). Discriminatory validity of the imaging instrument (i.e. thresholds of meaning, test-retest reliability, responsiveness or longitudinal construct validity and the ability to discriminate between change in two groups, or between groups) is evaluated in a trial, as well as its feasibility in term of both sonographer acceptability (i.e. time needed for examining all selected sites), patient acceptability (i.e. time spent for the overall examination, number of sites examined, comfort) and interpretability of the scoring system(s).

The validated definitions and the developed scoring system(s) both at site and at patient level, are applied in cross-sectional and longitudinal randomized controlled trials, and compared to other instruments. Once the new instrument has gone through step 4 it is ready for a final OFISA check (return to step1).

How does the OMERACT US group work?

Three co-chairs and an overall group mentor lead the OMERACT US WG. The co-chairs have a term of 6 years (3 OMERACT meetings).

For each new target pathology (e.g. enthesitis, dactylitis, tenosynovitis) of a disease entity; or for better definition (or new development) of their constituent “elementary lesions” — a new subgroup is formed. A subgroup mentor (one of the US WG co-chairs) oversees the research agenda for the validation process and ensures a balanced participation of interested US members and member-experts (i.e. methodologists, statisticians, clinicians etc.). The subgroup has a core group to coordinate the work, which includes the organization of research meetings, securing solid financial funding and ensuring tight collaboration with a statistician.

The OMERACT US WG meets annually at both the European League Against Rheumatism (EULAR) and American College of Rheumatology (ACR) congresses and biennially at the OMERACT Conference. An update of work of all the subgroups is presented in these meetings and future research activities are developed in subgroup discussions. Information about the group activities, publications and meetings can be accessed at www.OMERACT-US.org.

Membership of a subgroup is open to every OMERACT participant. To minimize the variability among sonographers in the practical exercises, participants must be sufficiently proficient in US (i.e. EULAR competency level 1, or equivalent, as assessed by the subgroup mentor).

Currently the OMERACT US WG has 18 subgroups (table 1) working in 10 different disease entities: rheumatoid arthritis, spondyloarthritis, psoriatic arthritis, idiopathic juvenile arthritis, gout, calcium pyrophosphate deposits disease, large vessel vasculitis, Sjogren syndrome (salivary glands involvement), lupus (musculoskeletal manifestations) and osteoarthritis. The progress of work is shown in Figure 2 (16-40).

Discussion

To address specific challenges involved in selecting outcome measurement instruments based on imaging, the US WG has developed a 4-step adaptation and elaboration of the OMERACT Instrument Selection Algorithm (OFISA) to include the development and testing of new imaging outcomes. Most existing US measurement instruments (i.e. the technique plus the scoring system) fail the OFISA test in **step 1**, through absent or incomplete definition of the target lesions, or unsatisfactory validation of the scoring system. **Steps 2 and 3** comprise a standardized procedure to develop and perform basic validation of definitions and scoring systems for the disease manifestation at site level ("theoretical or conceptual definition(s)") and its elementary lesion(s) ("operational" definition(s)). In other words, new instrument development is more or less a standard procedure in OMERACT US (and other imaging) work, whereas it is often optional in the selection of instruments based on patient report outcomes or clinical assessments. The final **step 4** is the production of the evidence needed for the instrument to pass OFISA (step 1) so that it can be selected for inclusion in a core outcome measurement set. We feel the method is applicable across all imaging techniques and hope it will facilitate and improve future research in this area.

Acknowledgements

PGC is supported in part by the UK National Institute for Health Research (NIHR) Leeds Biomedical Research Centre. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

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. Maxwell LJ, Beaton DE, Shea BJ, Wells GA, Boers M, Grosskleg S, et al. Core Domain Set Selection according to OMERACT Filter 2.1: The 'OMERACT Way'. *J Rheumatol* 2018 (submitted).
3. Boers M, Beaton D, Shea BJ, Maxwell LJ, Bartlett SJ, Bingham III CO, et al. OMERACT Filter 2.1: elaboration of the conceptual framework for outcome measurement in health intervention studies. *J Rheumatol* 2018 (submitted).
4. Beaton DE, Shea BJ, Maxwell LJ, Wells GA, Boers M, Grosskleg S et al. Instrument selection using the OMERACT Filter 2.1: The OMERACT Way. *J Rheumatol* 2018 (submitted).
5. Boers M, Brooks P, Strand CV, Tugwell P. The OMERACT filter for Outcome Measures in Rheumatology. *J Rheumatol.* 1999;26:210-6.
6. Boers M, Kirwan JR, Tugwell P, Beaton D, Bingham CO III, Conaghan PG, et al. The OMERACT Handbook. [Internet. Accessed May 17, 2017.] Available from: <https://omeract.org/resources>.
7. D'Agostino MA, Boers M, Kirwan J, van der Heijde D, Østergaard M, Schett G, et al. Updating the OMERACT filter: implications for imaging and soluble biomarkers. *J Rheumatol* 2014; 5:1016-24.
8. Joshua F, Lassere M, Bruyn GA, Szkudlarek M, Naredo E, Schmidt W, et al. Summary findings of a systematic review of the ultrasound assessment of synovitis. *J Rheumatol* 2007; 34:839-47.
9. Wakefield RJ, Balint P, Szkudlarek M, Filippucci E, Backhaus M, D'Agostino MA, et al. Musculoskeletal Ultrasound Including Definitions for Ultrasonographic Pathology. *J Rheumatol* 2005; 32:2485-7.
10. Obuchowski NA. How Many Observers Are Needed in Clinical Studies of Medical Imaging? *AJR* 2004;182:867–869
11. Bruyn GA, Iagnocco A, Terslev L, Keen HI, Naredo E, Conaghan PG et al. OMERACT definitions for ultrasonographic pathologies and elementary lesions of rheumatic disorders fifteen years on. *J Rheumatol* 2018 (submitted).
12. Landis JR, Koch GG. The measurement of observer agreement. for categorical data. *Biometrics* 1977;33:159–74.
13. Conger A. Integration and generalisation of kappas for multiple raters. *Psychol Bull* 1980;88:322–8.
14. Crewson PE. Reader Agreement Studies; *AJR* 2005;184:1391–1397
15. Shavelson R. J, Webb N. (1991). *Generalizability Theory: A Primer*. Thousand Oaks, CA: Sage.
16. Alcalde M, D'Agostino MA, Bruyn GAW, Möller I, Iagnocco A, Wakefield RJ, et al. A systematic literature review of ultrasound definitions, scoring systems and validity according to the OMERACT filter for tendon lesion in rheumatoid arthritis and other inflammatory joint disease. *Rheumatology* 2012; 51:1246-60.
17. Szkudlarek M, Terslev L, Wakefield RJ, Backhaus M, Balint PV, Bruyn GA, et al. Summary Findings of a Systematic Literature Review of the Ultrasound Assessment of Bone Erosions in Rheumatoid Arthritis. *J Rheumatol* 2016;43:12-21.
18. Gandjbakhch F, Terslev L, Joshua F, Wakefield RJ, Naredo E, D'Agostino M. Ultrasound in the evaluation of enthesitis: status and perspectives. *Arthritis Res Ther* 2011; 13:R188.
19. Collado P, Jousse-Joulin S, Alcalde M, Naredo E, D'Agostino MA. Is ultrasound a validated imaging tool for the diagnosis and management of synovitis in juvenile idiopathic arthritis? A systematic literature review. *Arthritis Care Res (Hoboken)* 2012; 64:1011-9.

20. D'Agostino MA, Terslev L, Aegerter P, Backhaus M, Balint P, Bruyn GA, et al. Scoring ultrasound synovitis in Rheumatoid Arthritis: a EULAR-OMERACT Ultrasound Taskforce – Part 1: definition and development of a standardized, consensus-based scoring system. *RMD Open* 2017;3:e000428.
21. Terslev L, Naredo E, Aegerter P, Wakefield RJ, Backhaus M, Balint P, et al. OMERACT-EULAR Ultrasound Task Force - reliability and applicability of a standardized consensus-based combined synovitis scoring system in Rheumatoid Arthritis. *RMD Open* 2017;3:e000427.
22. D'Agostino MA, Wakefield RJ, Berner-Hammer H, Vittecoq O, Filippou G, Balint P, et al. OMERACT-EULAR-Ultrasound Task Force. Value of ultrasonography as a marker of early response to abatacept in patients with rheumatoid arthritis and an inadequate response to methotrexate: results from the APPRAISE study. *Ann Rheum Dis* 2016;75:1763-9.
23. Naredo E, D'Agostino MA, Wakefield RJ, Möller I, Balint PV, Filippucci et al. Reliability of a consensus-based ultrasound score for tenosynovitis in rheumatoid arthritis. *Ann Rheum Dis* 2013;72:1328-34.
24. Bruyn GAW, Hanova P, Iagnocco A, D'Agostino MA, Möller I, Terslev L et al. Ultrasound definition of tendon damage in patients with rheumatoid arthritis. Results of a OMERACT consensus-based ultrasound score focusing on the diagnostic reliability. *Ann Rheum Dis* 2014;73:1929-34.
25. Ammitzbøll-Danielsen M, Østergaard M, Fana V, Glinatsi D, Døhn UM, Ørnbjerg L, et al. Intramuscular versus ultrasound guided peritendinous betamethasone injection for tenosynovitis in patients with rheumatoid arthritis - A randomised, double-blind, controlled study. *Ann Rheum Dis* 2017;76(4):666-672
26. Ammitzbøll-Danielsen M, Østergaard M, Naredo E, Terslev L. Validity and sensitivity to change of the semi-quantitative OMERACT ultrasound scoring system for tenosynovitis in patients with rheumatoid arthritis. *Rheumatology* 2016, 55(12):2156-2166.
27. Iagnocco A, Conaghan PG, Aegerter P, Möller I, Bruyn GAW, Chary-Valckenaere I, et al. The reliability of musculoskeletal ultrasound in the detection of cartilage abnormalities at the metacarpo-phalangeal joints. *Osteoarthritis Cartilage* 2012; 20:1142-6.
28. Hammer HB, Iagnocco A, Mathiessen A, Filippucci E, Gandjbakhch F, Kortekaas MC, et al. Global ultrasound assessment of structural lesions in osteoarthritis: a reliability study by the OMERACT ultrasonography group on scoring cartilage and osteophytes in finger joints. *Ann Rheum Dis* 2016;75:402-7.
29. Bruyn GA, Naredo E, Damjanov N, Bachta A, Baudoin P, Hammer HB et al. An OMERACT reliability exercise of inflammatory and structural abnormalities in patients with knee osteoarthritis using ultrasound assessment. *Ann Rheum Dis* 2016;75:842-6.
30. Gutierrez M, Schmidt WA, Thiele R, Keen H, Kaeley G, Naredo E, et al. International Consensus for Ultrasound Lesions in Gout. Results of Delphi Process and Web-Reliability Exercise. *Rheumatology (Oxford)* 2015;54:1797-805.
31. Terslev L, Gutierrez M, Christensen R, Balint PV, Bruyn GA, Delle Sedie A, et al. Assessing Elementary Lesions in Gout by Ultrasound: Results of an OMERACT Patient-based Agreement and Reliability Exercise. *J Rheumatol* 2015;42:2149-54.
32. Filippou G, Scirè CA, Damjanov N, Adinolfi A, Carrara G, Picerno V, et al. Definition and Reliability Assessment of Elementary Ultrasonographic Findings in Calcium Pyrophosphate Deposition Disease: A Study by the OMERACT Calcium Pyrophosphate Deposition Disease Ultrasound Subtask Force. *Ann Rheum Dis.* 2018;77:1194-1199.

33. Terslev L, Naredo E, Iagnocco A, Balint PV, Wakefield RJ, Aegerter P et al. Defining enthesitis in spondyloarthritis by ultrasound: results of a Delphi process and of a reliability reading exercise. *Arthritis Care Res (Hoboken)*. 2014;66:741-8.
34. Balint PV, Terslev L, Aegerter P, Bruyn GAW, Chary-Valckenaere I, Gandjbakhch F et al. Reliability of a consensus-based ultrasound definition and scoring for enthesitis in spondyloarthritis and psoriatic arthritis: an OMERACT US initiative. *Ann Rheum Dis*. 2018 Aug 3. pii: annrheumdis-2018-213609.
35. Roth J, Jousse-Joulin S, Magni-Manzoni S, Rodriguez A, Tzaribachev N, Iagnocco A et al. Definitions for the sonographic features of joints in healthy children. *Arthritis Care Res (Hoboken)* 2015;67:136-42.
36. Collado P, Vojinovic J, Nieto JC, Windschall D, Magni-Manzoni S, Aw Bruyn G, et al. Omeract Ultrasound Pediatric Group. Toward standardized musculoskeletal ultrasound in pediatric rheumatology: Normal age related ultrasound findings. *Arthritis Care Res (Hoboken)* 2016;68:348-56.
37. Windschall D, Collado P, Vojinovic J, Magni-Manzoni S, Balint PV, Bruyn GAW et al. Age-related vascularization and ossification of joints in children: an international pilot study to test multi-observer ultrasound reliability. *Arthritis Care Res* 2017;Aug 4. Epub ahead of print
38. Roth J, Ravagnani V, Backhaus M, Balint P, Bruns A, Bruyn GA, et al. Preliminary definitions for the sonographic features of synovitis in children. *Arthritis Care Res (Hoboken)*. 2017;69:1217-1223
39. Chrysidis S, Duftner C, Dejaco C, Schäfer VS, Ramiro S, Carrara G et al. Definitions and reliability assessment of elementary ultrasound lesions in giant cell arteritis: a study from the OMERACT Large Vessel Vasculitis Ultrasound Working Group. *RMD Open*. 2018;4:e000598
40. Schäfer VS, Chrysidis S, Dejaco C, Duftner C, Iagnocco A, Bruyn GA et al. Assessing Vasculitis in Giant Cell Arteritis by Ultrasound: Results of OMERACT Patient-based Reliability Exercises. *J Rheumatol*. 2018;45:1289-1295.
41. Brulhart L, Ziswiler HR, Tamborrini G, Zufferey P. The importance of sonographer experience and machine quality with regards to the role of musculoskeletal ultrasound in routine care of rheumatoid arthritis patients. *Clin Exp Rheumatol*. 2015;33:98-101

Legends

Table 1

Subgroups working in the core area of pathophysiological manifestations listed by disease entity or lesions and in relation to domains. SpA = SPondyloArthropathy, PsA= Psoriatic Arthritis, RA = Rheumatoid Arthritis, SLE= Systemic Lupus Erythematosus, MSK = MusculoSkeletal, OA= OsteoArthritis, JIA= Juvenile Idiopathic Arthritis, CPPD= Crystal PyroPhosphates Deposition; FUSS-RA = foot UltraSound Synovitis in Rheumatoid Arthritis,

Figure 1. Development of outcome instruments based on imaging.

Shows the four steps of the selection and development process. The colors applied to the arrows refer to the OMERACT Instrument selection Algorithm (OFISA). When an instrument is found in the review, its evidence can be found to be positive (green, ready for use; or amber ,for use with caution, set a research agenda); negative (red, do not use); or absent/insufficient (white, discard or develop evidence). New evidence is created depending on what is available. To date, all ultrasound-based instruments have been newly developed, i.e. from Step 2 onwards.

Figure 2. Progress of ultrasound-based instrument development.

Shows the stage of development according to the stepwise process of each of the 18 subgroups. SSc = Systemic Sclerosis, SLE= Systemic Lupus Erythematosus, PsA= Psoriatic Arthritis, JIA= Juvenile Idiopathic Arthritis, RA = Rheumatoid Arthritis, CPPD = Calcium PyroPhosphate deposition Disease, SpA = SpondyloArthropathy, OA=OsteoArthritis.