Proposal for Levels of Evidence Schema for Validation of a Soluble Biomarker Reflecting Damage Endpoints in Rheumatoid Arthritis, Psoriatic Arthritis, and Ankylosing Spondylitis, and Recommendations for Study Design

WALTER P. MAKSYMOWYCH, OLIVER FITZGERALD, GEORGE A. WELLS, DAFNA D. GLADMAN, ROBERT LANDEWÉ, MIKKEL ØSTERGAARD, WILLIAM J. TAYLOR, ROBIN CHRISTENSEN, PAUL-PETER TAK, MAARTEN BOERS, SILJE W. SYVERSEN, JOAN M. BATHON, CHRISTOPHER J. RITCHLIN, PHILIP J. MEASE, VIVIEN P. BYKERK, PATRICK GARNERO, PIET GEUSENS, HANI EL-GABALAWY, DANIEL ALETAHA, ROBERT D. INMAN, VIRGINIA BYERS KRAUS, TORE K. KVIEN, and DÉSIRÉE van der HEIJDE

ABSTRACT. ***Objective.*** At OMERACT 8 a framework for levels of evidence was proposed for the validation of biomarkers as surrogate outcome measures. We aimed to adapt this scheme in order to apply it in the setting of soluble biomarkers proposed to replace the measurement of damage endpoints in rheumatoid arthritis (RA), psoriatic arthritis (PsA), and ankylosing spondylitis (AS). We also aimed to generate consensus on minimum standards for the design of longitudinal studies aimed at validating biomarkers.

Methods. Before the meeting, the Soluble Biomarker Working Group prepared a preliminary framework and discussed various models for association and prediction related to the statistical strength domain. In addition, 3 Delphi exercises addressing longitudinal study design for RA, PsA, and AS were conducted within the working group and members of the Assessments in SpondyloArthritis International Society (ASAS) and the Group for Research and Assessment of Psoriasis and Psoriatic Arthritis (GRAPPA). This formed the basis for discussions among OMERACT 9 participants.

Results. The proposed framework was accepted by consensus. In the study design domain a requirement for both prospective observational studies and randomized controlled trials (RCT) in different drug classes was noted. A template for determining the level of statistical strength was proposed. The addition of a new domain on biomarker assay performance was considered essential, and participants suggested that for any biomarker this domain should be addressed first, i.e., before starting clinical validation studies. Participants agreed on most elements of a longitudinal study design template. Where consensus was lacking the working group has drafted solutions that constitute a basis for prospective validation studies.

Conclusion. The OMERACT 9 Soluble Biomarker Group has successfully formulated a levels of evidence scheme and a study design template that will provide guidance to conduct validation studies in the setting of soluble biomarkers proposed to replace the measurement of damage endpoints in RA, PsA, and AS. (J Rheumatol 2009;36:1792–9; doi.10.3899/jrheum090347)

Key Indexing Terms:
RHEUMATOID ARTHRITIS PSORIATIC ARTHRITIS ANKYLOSING SPONDYLITIS
BIOMARKERS STUDY DESIGN STRUCTURAL DAMAGE

From the Department of Medicine, University of Alberta, Edmonton, Alberta, Canada; St. Vincent's University Hospital, Dublin, Ireland; Department of Epidemiology and Community Medicine, University of Ottawa; Ottawa Health Research Institute, Ottawa, Canada; University of Toronto, Toronto Western Research Institute, University Health Network, Toronto, Canada; University Hospital Maastricht, Maastricht, The Netherlands; Department of Rheumatology, Copenhagen University Hospitals at Hvidovre and Herlev, Copenhagen, Denmark; Swedish Medical Center, Seattle, Washington, USA; Rehabilitation Teaching and Research Unit, University of Otago, Wellington, New Zealand; The Parker Institute, Musculoskeletal Statistics Unit, Frederiksberg Hospital, Frederiksberg, Denmark; Division of Clinical Immunology and Rheumatology, Academic Medical Center/University of Amsterdam, Amsterdam; Department of Clinical Epidemiology and Biostatistics, VU University Medical Center, Amsterdam, The Netherlands; Department of Rheumatology, Diakonhjemmet Hospital, University of Oslo, Oslo, Norway; Division of Rheumatology, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland; Clinical Immunology Research Center, University of Rochester Medical Center, Rochester, New York, USA; Rebecca MacDonald Center for Arthritis and Autoimmune Diseases, Mount Sinai Hospital, University Health Network, Toronto, Canada; INSERM Research Unit 664 and CCBR-SYNARC, Lyon, France; Internal Medicine/Rheumatology, Maastricht University Medical Centre, Maastricht, The Netherlands; Biomedical Research Centre, University Hasselt, Flanders, Belgium; University of Manitoba, Winnipeg, Canada; Division of Rheumatology, Department of Internal

Medicine III, Medical University of Vienna, Vienna, Austria; Toronto Western Research Institute, Toronto, Canada; Duke University Medical Center, Durham, North Carolina, USA; Diakonhjemmet Hospital, University of Oslo, Oslo, Norway; and Leiden University Medical Center, Leiden, The Netherlands.

W.P. Maksymowych is a Scientist of the Alberta Heritage Foundation for Medical Research.

W.P. Maksymowych, FRCPC, Professor of Medicine, Department of Medicine, University of Alberta; O. FitzGerald, MD, FRCPI, FRCP(UK), Newman Clinical Research Professor, St. Vincent's University Hospital; G.A. Wells, MSc, PhD, Professor, Department of Epidemiology and Community Medicine, University of Ottawa, Senior Scientist, Ottawa Health Research Institute; D.D. Gladman, MD, FRCPC, Professor of Medicine, University of Toronto, Senior Scientist, Toronto Western Research Institute, University Health Network; R. Landewé, MD, PhD, Professor of Medicine, University Hospital Maastricht; M. Østergaard, MD, PhD, DMSc, Professor in Rheumatology/Arthritis, Department of Rheumatology, Copenhagen University Hospitals at Hvidovre and Herlev; P.J. Mease, MD, Professor, Swedish Medical Center; W.J. Taylor, PhD, FRACP, FAFRM, Associate Professor, Rehabilitation Teaching and Research Unit, University of Otago; R. Christensen, MSc, Biostatistician, The Parker Institute, Musculoskeletal Statistics Unit, Frederiksberg Hospital; P-P. Tak, MD, PhD, Professor of Medicine, Division of Clinical Immunology and Rheumatology, Academic Medical Center, University of Amsterdam; M. Boers, MSc, MD, PhD, Professor of Clinical Epidemiology, Department of Clinical Epidemiology and Biostatistics, VU University Medical Center; S.W. Syversen, MD, Department of Rheumatology, Diakonhjemmet Hospital, University of Oslo; J.M. Bathon, Professor of Medicine, Division of Rheumatology, Department of Medicine, Johns Hopkins University School of Medicine; C.J. Ritchlin, MD, PhD, Director, Clinical Immunology Research Center, Professor of Medicine, University of Rochester Medical Center; V.P. Bykerk, MD, FRCPC, Assistant Professor of Medicine, University of Toronto, Rebecca MacDonald Center for Arthritis and Autoimmune Diseases, Mount Sinai Hospital, University Health Network; P. Garnero, PhD, DSc, INSERM Research Unit 664 and CCBR-SYNARC; P. Geusens, MD, PhD, Professor, Internal Medicine/Rheumatology, Maastricht University Medical Centre, and Biomedical Research Centre, University Hasselt; H. El-Gabalawy, MD, FRCPC, Professor of Medicine and Immunology, University of Manitoba; D. Aletaha, MD, PhD, Associate Professor, Division of Rheumatology, Department of Internal Medicine III, Medical University of Vienna; R.D. Inman, MD, FRCPC, Professor of Medicine and Immunology, University of Toronto, Senior Scientist, Toronto Western Research Institute; V.B. Kraus, MD, PhD, Associate Professor of Medicine, Duke University Medical Center; T.K. Kvien, MD, PhD, Professor of Rheumatology, Diakonhjemmet Hospital, University of Oslo; D. van der Heijde, MD, PhD, Professor of Rheumatology, Leiden University Medical Center.

Address reprint requests to Dr. W.P. Maksymowych, 562 Heritage Medical Research Building, University of Alberta, Edmonton, Alberta T6G 2S2, Canada. E-mail: walter.maksymowych@ualberta.ca

In healthcare all interventions should be aimed at improving patient outcome, defined as "how a patient feels, functions and survives"[1]. As longterm outcomes are often difficult to identify in the setting of a clinical trial, measurement of biomarkers that could serve as surrogate outcomes are an attractive possibility, but proper validation of a biomarker for use in this setting is difficult. At the OMERACT 8 conference a scheme was proposed that grades the level of evidence in support of a biomarker meeting the definition of a surrogate outcome (see Appendix)[1]. The soluble biomarker group felt that such a scheme could be adapted for a step earlier in the development process of a drug, i.e., to validate a biomarker that could replace the measurement of damage endpoints in early proof of concept studies. Development and validation of such biomarkers reflecting structural damage currently constitutes a high priority objective both for the drug discovery process and for the practising clinician, particularly for inflammatory disorders of joints and spine where damage progression is slow.

The scheme proposed at OMERACT 8 is based on 4 domains: target, study design, statistical strength, and penalties. For the domains target (that is, substituted by the marker), study design (of the best evidence), and statistical strength, the scores are additive. Penalties are then applied if there is serious counter-evidence. A total score (0 to 15) determines the 5 levels of evidence, with Level 1 the strongest and Level 5 the weakest. There was also agreement with the proposal that biomarkers that have been validated at only Levels 3 to 5 constituted disease-centered variables with no immediate or obvious meaning to patients or clinicians, while biomarkers that attained Levels 1 or 2 validation constituted patient-centered variables with obvious patient and clinical relevance. It was proposed that the term "surrogate" be restricted only to markers attaining Levels 1 or 2.

In discussions at that conference it was recommended for the study design domain that the rankings be more explicit in the minimum standards of design for both observational and randomized controlled studies. Work on the statistical strength domain was deferred to a statistics working group[2]. An important omission from the generic framework that is particularly relevant to soluble biomarkers is the absence of a performance domain that stipulates recommended standards for the handling and processing of soluble biomarker samples.

The Soluble Biomarker Working Group outlined 3 objectives in the program of work for OMERACT 9: (1) To adapt the generic biomarker levels of evidence framework for soluble biomarkers. (2) To set minimum standards for study design that validates biomarkers as reflecting structural damage in rheumatoid arthritis (RA), psoriatic arthritis (PsA), and ankylosing spondylitis (AS). (3) To propose a framework for quantifying the statistical strength of the association between the biomarker and the damage endpoint.

METHODS

Levels of evidence framework. The generic biomarker framework was presented at a specially convened meeting of the OMERACT 9 Soluble Biomarker Working Group that was held over 2 days in London, England, in November 2007. The primary objectives of biomarkers for RA, PsA, and AS were first discussed and agreed upon, and the generic scheme to assignment of levels of evidence developed at OMERACT 8 was reviewed. This was followed by discussion and critique of the framework with respect to its application to the validation of soluble biomarkers. A proposal for an adaptation of the framework for the validation of soluble biomarkers reflecting damage endpoints was then drafted. This new proposal was presented to participants at OMERACT 9. This included a document that highlighted the proposed modifications to the OMERACT 8 generic framework. The framework was discussed independently by 2 groups at the breakout sessions, and the statistical strength domain was discussed by a separate working group of methodologists and biostatisticians. Rapporteurs summarized the principal issues and concerns and the proposed modifica-

tions to the draft soluble biomarker framework at the report-back plenary session. After further discussion in the plenary session, modifications to the framework that generated consensus were incorporated into the new scheme, and participants were then asked to vote on the following question: "The working group has adopted the framework and domains outlined in the OMERACT 8 Surrogate Superworkshop for generating levels of evidence. Do you agree with the new framework for soluble biomarkers?"

Principal requirements for longitudinal study design. The principal aim of this initiative was to propose a minimum set of standards with respect to study design, principal outcomes, processing of biomarker samples, and documentation of potential confounders for the conduct of a longitudinal study aimed at the validation of a soluble biomarker reflecting damage endpoints. This was conducted using a Delphi approach. The principal design issues were identified at the London meeting using the framework for longitudinal studies generated at OMERACT 4 that highlighted core domains (health status, disease process, damage), potential covariates, demographic variables, and study design features that ought to be addressed when planning a longitudinal study[3].

The discussions in London constituted the first phase of the Delphi exercise, the solicitation of items, and addressed issues relevant to all 3 categories of arthritis. The subsequent steps in the Delphi were conducted separately for RA, PsA, and AS. Three steps in the Delphi exercise were organized electronically for each of the 3 different disease categories. The first electronic exercise solicited additional domains organized under categories of health status (symptoms, physical function, psychosocial function), disease process (joint tenderness/swelling, global disease, acute-phase reactants), and damage (imaging). Working group members were also asked to propose potentially confounding covariates and relevant demographic variables. Members were asked to propose items for core study methodology organized under the following headings: inclusion criteria, disease phenotype, study duration, approach to selection of patient cohort, treatment strategy, analysis of radiographic endpoint, frequency of clinical assessment, type of biomarker sample collected, frequency/time of biomarker sample collection, biomarker sample processing, and biomarker sample transport and storage. For RA, the convenor of the Delphi (WPM) provided a draft template of items based on discussions at the London meeting to the OMERACT 9 Biomarker Working Group members as a basis for further solicitation of items. For PsA and AS, solicitation of items was conducted electronically among the membership of the Group for Research and Assessment of Psoriasis and Psoriatic Arthritis (GRAPPA) and the Assessment in SpondyloArthritis International Society (ASAS), respectively, after a draft template was provided by convenors for the AS (WPM) and PsA (OF) Delphi.

Electronic voting was then conducted in the subsequent 2 rounds of the Delphi exercise among OMERACT 9 Biomarker Working Group members. In addition, voting was conducted among GRAPPA members for PsA, among ASAS members for AS, and among OMERACT 9 registrants for RA. Two types of voting questions were presented. One type requested selection of an item among a range of options. Consensus for selection of a particular item was defined on the basis of ≥ 70% of participants voting in favor of that item, while consensus for exclusion was defined as ≤ 30% of participants voting for that item in any round of voting. The second type of question was presented in a Likert format comprising 5 scoring categories ranging from 1 = definitely unacceptable and/or unnecessary, exclude from study design, to 5 = definitely acceptable, essential that it be included in the study design. An additional option was provided, namely, "don't know/not an expert." Consensus for selection or exclusion of a particular item was defined on the basis of ≥ 70% or ≤ 30%, respectively, of participants voting a score of 4 or 5 on the Likert scale in any round of voting. The results of the Delphi exercise were presented at OMERACT 9, and participants were presented with a summary handout at the soluble biomarker group plenary session. Principal areas of disagreement were highlighted in the handout and discussed at the plenary session.

Development of statistical strength domain. For the statistical strength domain, the strength of association and prediction models is the central theme. Various models for assessing the association between marker change and target change, and for assessing prediction of the effect of treatment on marker change and target change were presented and discussed at the OMERACT 9 Soluble Biomarker Working Group meeting in London. Relevant models presented were based on: change in the biomarker during therapy; change in the target outcome in the long term, including measuring the outcome repeatedly for greater insight into the progression of the target outcome; and change in pertinent covariates during therapy and/or repeatedly during the longterm period for greater insight into the nature of the "confounding" relationship. Depending on the nature of the data, various models were considered, including: regression analysis of target outcome on change in biomarker; multiple regression analysis of target outcome on change in biomarker and the covariates; longitudinal multiple regression analysis of target outcome on change in biomarker and the time-dependent covariates; and mixed model repeated measures.

The data used for demonstrating the various models were from the Combinatietherapie Bij Reumatoide Artritis (COBRA) trial dataset[4]. This trial showed that step-down combination therapy with prednisolone, methotrexate, and sulfasalazine (SSZ) was superior to SSZ monotherapy for suppressing disease activity and radiologic progression of RA. The analysis focused on investigating whether urinary C-terminal cross-linking telopeptide of type II (CTX-II) collagen, a specific biochemical marker of cartilage degradation, was associated with radiological damage and progression in patients with RA. Various regression-based models were presented and discussed, and the need for a template to determine the statistical strength for such models was identified. A research agenda was determined to review various schemas that could be used for categorizing and determining levels of statistical strength.

RESULTS AND DISCUSSION

Levels of evidence framework. The generic framework for a levels of evidence scheme was adapted for soluble biomarkers at the London meeting (see Appendix), and following modification at OMERACT 9 was accepted by 85% of workshop participants (Table 1). Agreement was reached on the following adaptations to the domains:

1. Target outcome domain. The grading of 0 (disease-centered, reversible) to 5 (death) was not considered relevant to the validation of a biomarker reflecting structural damage. A grading of 0 to 3 [patient centered, irreversible, minor organ/clinical morbidity (radiography)] was considered appropriate, with radiography being accepted as a patient-centered outcome. Some argued that the target outcome has already been defined as radiography in formulating the principal objectives of the validation process and that there is, therefore, no need to include this domain in the scheme. The counter-argument was that other measures of damage, e.g., magnetic resonance imaging (MRI), may be increasingly relevant as validation data increase. For example, it has been shown that bone marrow edema on an MRI has predictive validity for radiographic damage and can be reliably detected and quantified[5]. As clinicians increasingly target and require guidance in the treatment of pre-radiographic disease, MRI may increasingly constitute a relevant outcome for biomarker validation studies.
2. Study design domain. The grading of 0 (animal studies, case reports, cross-sectional, retrospective) to 5 (≥ 3 RCT each of different drug class, ≥ 3 randomized surrogate

Table 1. OMERACT 9 Levels of Evidence framework for validation of a soluble biomarker reflecting damage endpoints in rheumatoid arthritis, psoriatic arthritis, and ankylosing arthritis (adapted from the generic biomarker framework at OMERACT 8[1]).

Domain Components
1. Target outcome
0 Disease-centered, reversible
1 Disease-centered, irreversible
2 Patient-centered, reversible
3 Patient-centered, irreversible, minor organ/clinical morbidity (radiography)
2. Study design
1 Prospective, non-population, observational
2 Prospective, population observational or 1 RCT
3 ≥ 2 RCT and/or ≥ 2 prospective observational studies, same drug class (total of any 2)
4 ≥ 2 RCT and/or ≥ 2 prospective observational studies, each of different drug class (total of any 2)
5 ≥ 3 studies, ≥ 1 RCT and ≥ 1 prospective observational study (at least one of each study design), different drug class studies
3. Penalties*
–1 No evidence in ≥ 1 powered RCT
–1 Opposite assertion in epidemiological study
–1 No evidence in ≥ 1 epidemiological powered study
–1 ≥ 1 RCT demonstrating clinical heterogeneity
–2 ≥ 1 RCT supports opposite assertion
–3 ≥ 1 RCT use of marker confers patient harm
4. Performance criteria†
Reproducibility
Feasibility (readily accessible, availability of international standards, reasonable costs)
Confounders (assay related, non-assay related)
Stability

* Penalties are not additive. The highest penalty ranking is applied. Studies should meet minimum standards for longitudinal study design. † 1. Performance criterion domain meets criteria 1, 2, 4, and 5 of OMERACT 9 v2 criteria. 2. Performance criteria should be met in their entirety before clinical validation studies. RCT: randomized controlled trial.

objective trials) was modified to incorporate an equal weighting for RCT and prospective observational studies. Longitudinal studies would have to be consistent with the minimum standards for longitudinal study design advocated by the group (see below). Randomized surrogate endpoint trials were considered too high a hurdle for the objectives of this biomarker validation process. The proposed ranking recognizes the importance of biomarker validation with different drug classes. For example, it has now been consistently demonstrated that C-reactive protein has predictive validity for structural damage in patients with RA receiving methotrexate, but not in those receiving anti-tumor necrosis factor therapies[6-8]. Both longitudinal cohort studies and RCT are deemed essential. The former address validation in a wider spectrum of patients and over longer time periods, while the latter can more readily address validation with different drug classes and potential confounders.

3. Statistical strength domain. In evaluating the association between marker change and target change or the prediction of the effect of treatment on marker change and target change, regression-based modeling is the primary statistical technique, and the fitting of the model will lead to a goodness of fit statistic (such as, percentage of the variation explained by the model R^2). The statistical evidence of the association or prediction can be determined using a modification of the Sterne and Smith interpretation of the p value, taking the number of observations into consideration[9], whereas the statistical strength per se can be based on the model, with the effect size determined using the coefficient for the biomarker [e.g., the odds ratio (OR)]. This effect estimate can be translated using Cohen's standardized mean difference (SMD), i.e., an OR value can be transformed into an SMD[10]:

$$SMD = {}^{\sqrt{3}}/_{\pi} \log OR$$

and levels of strength can be derived based on the usual thresholds for interpreting "fair" (0.2), "good" (0.5), "very good" (0.8).

4. Penalties domain. The grading in the generic template proposal was largely adopted although with the stipulation that rather than being additive for different studies, the highest score would be applied as a penalty and that the same minimum standards be applied to the evaluation of study design.

5. Performance domain. This domain is not a component of the generic template but it was agreed that biomarkers should have been validated according to the criteria com-

prising this domain before proceeding with clinical validation studies. The criteria address standards of reproducibility, feasibility (readily accessible, availability of international standards, costs), biomarker stability, and evaluation of confounders that are defined in the OMERACT 9 biomarker validation draft criteria under the categories of feasibility and discrimination (criteria 1, 2, 4, and 5).

Longitudinal study design consensus. The following design issues and key recommendations were highlighted at the London meeting for consideration in the Delphi voting exercise: principal inclusion criteria, study design (RCT vs observational), treatment strategy, study duration, appropriate damage endpoints, frequency of assessment, and sample collection and processing. A total of 52 ASAS and 45 GRAPPA members provided additional input into the items proposed for the AS and PsA Delphi exercises, respectively. For the first round there were 130 OMERACT 9 participants, 46 ASAS members, and 53 GRAPPA members who participated in the Delphi voting exercise for RA, AS, and PsA, respectively. In the second round of voting, the corresponding number of participants was 113, 43, and 46, for OMERACT 9 participants, ASAS, and GRAPPA members. The final results of these 3 Delphi exercises are presented in Table 2.

Failure of consensus was evident for 2 key items that were discussed further at OMERACT 9. The first focused on the diagnostic inclusion criterion for a validation study of an RA biomarker. There were 2 principal schools of thought on this matter. Some considered it desirable for a validation study, especially the first, to stipulate the American College of Rheumatology (ACR) classification criteria on the premise that such patients would be not only relatively homogeneous but also more likely to demonstrate disease progression, which increases statistical power. Inclusion of patients with a wide spectrum of disease activity and severity was also considered desirable, since some biomarkers may reflect radiographic progression better in early versus late disease and vice versa. Other participants were support-

Table 2. Summary results of 3-stage Delphi consensus exercise addressing minimum standards for longitudinal study design for validation of biomarker reflecting damage endpoints in rheumatoid arthritis (RA), psoriatic arthritis (PsA), and ankylosing spondylitis (AS). Items lacking consensus are indicated in bold type (percentage of respondents voting in support of the item is indicated in parentheses).

	RA	PsA	AS
Inclusion Criteria	**ACR (48%)[17], EULAR early referral criteria (26%)[18], ANTI-CCP arthritis (25%)**	CASPAR[19]	**Modified New York (67%)[20] Pre-radiographic axial (24%)[21]**
Treatment strategy	**All treatments (68%)**	All treatments	All treatments
Selection of patient cohort	Consecutive cases	**Consecutive cases (67%) Inception cohort (33%)**	Consecutive cases
Study duration	2 yrs	**4 yrs (69%) 2 yrs (27%)**	4 yrs
Frequency of assessment	Every 3 mo	Every 6 mo	Every 6 mo
Analysis of radiographic endpoint	Blinded to timepoint	Blinded to timepoint	Blinded to timepoint
Allow steroid	Yes	Not considered	Not considered
Rules for changes in treatment	By predetermined DAS every 3 mo	Not considered	Not considered
Frequency of biomarker collection	Q6 mo and prior to new DMARD/anti-TNF	Q6 mo and prior to new DMARD/anti-TNF	Q6 mo and prior to new DMARD/ anti-TNF
Symptoms	Pain, patient global, **fatigue (50%), stiffness (51%)**	Pain, skin global, patient global, **stiffness (61%), fatigue (40%)**	Pain, stiffness, patient global, fatigue (38%)
Physical function	Patient self-reported function, **objective measures of function** (47%)	Patient self-reported function	Patient self-reported function, metrology
Psychosocial function	Quality of life	Quality of life	Quality of life
Other	Work status, **work productivity (53%), participation (42%)**	Work status, **work productivity (33%), participation (40%)**	Work status
Disease activity	Joint inflammation, global disease activity (patient/physician), general labs (ESR, CRP), **imaging/MRI (62%), imaging/US (31%)**	Joint inflammation, global disease activity, (patient/physician), clinical enthesitis, dactylitis, spinal, skin, general labs (ESR, CRP), **nail (56%), extraarticular disease (67%)**	Joint inflammation, global disease activity (patient/physician), clinical enthesitis, metrology, general labs (ESR, CRP), extraarticular disease, imaging/MRI
Radiographic damage endpoint	Modified Sharp	Modified Sharp	mSASSS

ACR: American College of Rheumatology; EULAR: European League Against Rheumatism; CCP: cyclic citrullinated peptide; CASPAR: CASPAR Study Group Criteria; DAS: Disease Activity Score; DMARD: disease modifying antirheumatic drugs; TNF: tumor necrosis factor; ESR: erythrocyte sedimentation rate; CRP: C-reactive protein; MRI: magnetic resonance imaging; mSASSS: modified Stoke AS Spinal Score; US: ultrasonography.

ive of differentiating patients on the basis of the anti-cyclic citrullinated peptide (CCP) antibody test in early disease on the premise that these patients are a distinct group both prognostically and on the basis of pathophysiology[11-13]. The latter could, therefore, imply quantitative and/or qualitative differences in the relationship between a particular biomarker and radiographic damage. A compromise proposal was to include patients on the basis of the ACR criteria but then to prespecify analysis stratified by anti-CCP status. Both RCT and observational studies were considered equally desirable to ensure generalizability of study findings, although RCT for AS were not considered feasible because progression of radiographic change is not reliably detected prior to 2 years in patients on standard therapy[14]. Validation in studies employing diverse and flexible treatment strategies was considered desirable since the real clinical utility of a biomarker is dependent on the demonstration that levels of the biomarker are independently associated with structural damage regardless of treatment approach.

Consensus was not achieved in regard to the minimum standards for the handling of biomarker samples because a substantial minority of respondents refrained from voting in the Delphi exercise as they assigned themselves the designation "not an expert." It was decided by consensus that the biomarker group should develop a proposal for the systematic handling of biomarker samples, which is presented in Table 3. First, the group has recommended the collection of both urine and serum. Although feasibility is an obvious advantage for serum, it is important to standardize the collection of serum in view of previous reports that preanalytical handling of serum influences certain biomarker levels, such as metalloproteinases, which are released from platelets and leukocytes particularly when using collection tubes that enhance clotting (kaolin-coated)[15,16]. Ideally, possible interfering factors should be identified as discussed under the Performance criterion assay-related confounders (Table 1) and recommendations for standardization of sample collection clarified prior to clinical studies. A practical problem is that several biomarkers are often tested simultaneously, and collection procedure may not be optimal for all biomarkers. In addition, samples are often collected as a routine during observational studies and RCT and then analyzed retrospectively. It is obvious that analysis of the individual biomarkers should only be done on samples that have been obtained and handled in a way that ensures reliable measurement with respect to diurnal variation, centrifugation, freezing temperature, stability to freeze/thaw cycles, etc. These characteristics may vary considerably from marker to marker. It will not always be known in advance which markers will later be analyzed. Therefore, a default approach is to recommend standardized operating procedures for sample collection as outlined in Table 3.

CONCLUSIONS AND FURTHER DIRECTIONS

The OMERACT 9 Soluble Biomarker Working Group has laid the groundwork for a systematic and standardized approach to biomarker validation studies. These recommendations constitute draft proposals until tested in prospective studies. These prospective studies will form the basis for further revision of these recommendations in preparation for subsequent OMERACT meetings.

REFERENCES

1. Lassere MN, Johnson KR, Boers M, et al. Definitions and validation criteria for biomarkers and surrogate endpoints: development and testing of a quantitative hierarchical levels of evidence schema. J Rheumatol 2007;34:607-15.
2. Lassere M, Johnson K, Hughes M, et al. Simulation studies of surrogate endpoint validation using single trial and multitrial statistical approaches. J Rheumatol 2007;34:616-9.
3. Wolfe F, Lassere M, van der Heijde D, et al. Preliminary core set of domains and reporting requirements for longitudinal observational studies in rheumatology. J Rheumatol 1999;26:484-9.
4. Garnero P, Landewé R, Boers M, et al. Association of baseline levels of markers of bone and cartilage degradation with long-term progression of joint damage in patients with early rheumatoid arthritis: the COBRA study. Arthritis Rheum 2002;46:2847-56.
5. Ostergaard M, Hansen M, Stoltenberg M, et al. Magnetic resonance imaging-determined synovial membrane volume as a marker of disease activity and a predictor of progressive joint destruction in the wrists of patients with rheumatoid arthritis. Arthritis Rheum 1999;42:918-29.
6. Smolen JS, van der Heijde DM, St. Clair EW, et al. Predictors of joint damage in patients with early rheumatoid arthritis treated with high-dose methotrexate with or without concomitant infliximab. Arthritis Rheum 2006;54:702-10.
7. Smolen JS, Han C, Bala M, et al. Evidence of radiographic benefit of treatment with infliximab plus methotrexate in rheumatoid arthritis patients who had no clinical improvement. A detailed

Table 3. OMERACT 9 Soluble Biomarker Working Group minimum standards for the handling and processing of biomarker samples.

	Recommendation	Comment
Type of sample	Serum, urine	Standardized sample collection (e.g., tubes) Fasting sample
Time of collection	2–4 hours after rising	Diurnal variation is common with musculoskeletal biomarkers[22]
Time of sample processing	Within 2 hours of collection	Standardized centrifugation procedure. Keep sample at 4°C prior to centrifugation[23]
Storage of samples	Sample storage in 300–400 μl aliquots in –70°C freezer	Avoid use of sample after 3 freeze/thaw cycles
Sample transport	Express delivery on dry ice	Use ample dry ice

Appendix 1. Ranking surrogate validity: domains, criteria, and ranks. From Lassere, *et al*. J Rheumatol 2007;34:607-15.

Domains	Ranks	Criteria
A. Target (for all studies ranked in Domain B)	0	All targets studied are **disease-centred** and **reversible**.
	1	At least one target studied that is **disease-centred** is **irreversible**
	2	At least one **patient-centered** target that is **reversible**
	3	At least one **patient-centered** target of **irreversible minor organ morbidity** or **minor** irreversible **clinical burden of disease**
	4	At least one **patient-centered** target of **irreversible major organ morbidity** or **major** irreversible **clinical burden of disease**
	5	**Death**
B. Study design (Requires as baseline appropriate study quality , study power and study duration)	0	Evidence from ***in vitro*** *OR* **animal studies** *OR* **Case reports** *OR* **Cross-sectional observational** *OR* **Retrospective observational cohorts** studies evaluating the relationship between marker and target.
	1	**At least one prespecified non-population based prospective observational study** with collection of all covariates needed to adjust for known confounding and effect modification evaluating the relationship between marker and target.
	2	**At least one prespecified population-based prospective observational study** with collection of all covariates needed to adjust for known confounding and effect modification evaluating the relationship between marker and target *OR* **One randomized controlled trial** of the **same drug class** of an intervention evaluating the relationship between marker and target.
	3	**At least two randomized controlled trials** of the **same drug class** of an intervention evaluating the relationship between marker and target.
	4	**At least two randomized controlled trials** in each of **two drug classes** of an intervention evaluating the relationship between marker and target.
	5	**At least three randomized controlled trials in each of three known drug classes** of an intervention evaluating the relationship between marker and target *OR* at least three **randomized surrogate objective** trials
C. Statistical Strength	0	**No** association / prediction **OR** no relevant data
	1	**At least fair** association or better between marker change and target change in most single study analyses
	2	**At least fair** association or better between marker change and target change in all single study analyses **OR** fair prediction in an across-study analysis evaluating the effect of treatment on marker change and target change.
	3	**At least good** association or better between marker change and target change in all single study analyses **OR** good prediction in an across-study analysis evaluating the effect of treatment on marker change and target change.
	4	**At least very good** association or better between marker change and target change in all single study analyses **AND** very good prediction in an across-study analysis evaluating the effect of treatment on marker change and target change.
	5	**Excellent** association between marker change and target change in all single study analyses **AND** excellent prediction in an across-study analysis evaluating the effect of treatment on marker change and target change.
D. Penalties due to lack of evidence or evidence to the contrary	-1	**No *in vitro* or animal study** evidence to support surrogacy validity OR no **epidemiological** evidence to support surrogacy validity
	-1	At least one **RCT** that does not demonstrate statistically significant surrogate validity (i.e. **evidence of no effect in at least one adequately powered RCT**)
	-1	At least one **epidemiological** study that supports **opposite assertion**.
	-1	At least one **epidemiological** study that does not demonstrate surrogacy validity (i.e. **evidence of no effect in at least one adequately powered epidemiological study**)
	-1	At least one RCT that demonstrated evidence of significant **clinical heterogeneity**
	-2	At least **one RCT** that supports **opposite assertion**
	-3	At least **one RCT** that demonstrates use of marker **confers patient harm**
	-3	**Does not meet the threshold criterion** of a rank of 3 in at least one domain if score is 7 or more
NB. Marker must meet minimum technical performance criteria as per OMERACT Filter		

subanalysis of data from the Anti-tumor Necrosis Factor Trial in Rheumatoid Arthritis with Concomitant Therapy Study. Arthritis Rheum 2005;52:1020-30.
8. Landewe R, van der Heijde DM, van Vollenhoven R, Fatenejad S, Klareskog L. A disconnect between inflammation and radiographic progression in patients treated with etanercept plus methotrexate and etanercept alone as compared to methotrexate alone: results from the TEMPO-Trial [abstract]. Arthritis Rheum 2005;52 Suppl:S343.
9. Sterne JAC, Smith GD. Sifting the evidence — what's wrong with significance tests? BMJ 2001;322:226-31.
10. Chinn S. A simple method for converting an odds ratio to effect size for use in meta-analysis. Stat Med 2000;19:3127-31.
11. Forslind K, Ahlmen M, Eberhardt K, Hofstrom, I, Svensson, B. Prediction of radiological outcome in early RA in clinical practice: role of antibodies to citrullinated peptides (anti-CCP). Ann Rheum Dis 2004;63:1090-5.
12. Kastbom A, Strandberg G, Lindroos A, Skogh T. Anti-CCP antibody test predicts the disease course during three years in early rheumatoid arthritis (the TIRA project). Ann Rheum Dis 2004;63:1085-9.
13. Ronnelid J, Wick MC, Lampa J, et al. Longitudinal analysis of anticitrullinated protein/peptide antibodies (anti-CP) during 5 year follow up in early rheumatoid arthritis: anti-CP status is a stable phenotype that predicts worse disease activity and greater radiological progression. Ann Rheum Dis 2005;64:1744-9.
14. Wanders AJ, Landewé RB, Spoorenberg A, et al. What is the most appropriate radiologic scoring method for ankylosing spondylitis? A comparison of the available methods based on the Outcome Measures in Rheumatology Clinical Trials filter. Arthritis Rheum 2004;50:2622-32.
15. Jung K, Laube C, Lein M, et al. Kind of sample as preanalytical determinant of matrix metalloproteinases 2 and 9 and tissue inhibitor of metalloproteinase 2 in blood. Clin Chem 1998;44:1060-2.
16. Jung K. Careful attention to blood sampling as a preanalytical determinant of circulating matrix metalloproteinase 9 to avoid misinterpretations: comment on the article by Ainiala et al [letter]. Arthritis Rheum 2005;52:673-4.
17. Arnett FC, Edworthy SM, Bloch DA, et al. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 1988;31:315-24.
18. 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.
19. Taylor W, Gladman D, Helliwell P, Marchesoni A, Mease P, Mielants H. CASPAR Study Group. Classification criteria for psoriatic arthritis: development of new criteria from a large international study. Arthritis Rheum 2006;54:2665-73.
20. van der Linden S, Valkenburg HA, Cats A. Evaluation of diagnostic criteria for ankylosing spondylitis: a proposal for modification of the New York criteria. Arthritis Rheum 1984;27:361-8.
21. Rudwaleit M, Metter A, Listing J, Sieper J, Braun J. Inflammatory back pain in ankylosing spondylitis: a reassessment of the clinical history for application as classification and diagnostic criteria. Arthritis Rheum 2006;54:569-78.
22. Kong SY, Stabler TV, Criscione LG, Elliott AL, Jordan JM, Kraus VB. Diurnal variation of serum and urine biomarkers in patients with radiographic knee osteoarthritis. Arthritis Rheum 2006;54:2496-504.
23. Skogstrand K, Ekelund CK, Thorsen P, et al. Effects of blood sample handling procedures on measurable inflammatory markers in plasma, serum and dried blood spot samples. J Immunol Methods 2008;336:78-84.