

Updating the OMERACT Filter: Discrimination and Feasibility

George Wells, Dorcas E. Beaton, Peter Tugwell, Maarten Boers, John R. Kirwan, Clifton O. Bingham III, Annelies Boonen, Peter Brooks, Philip G. Conaghan, Maria-Antonietta D'Agostino, Maxime Dougados, Daniel E. Furst, Laure Gossec, Francis Guillemain, Philip Helliwell, Sarah Hewlett, Tore K. Kvien, Robert B. Landewé, Lyn March, Philip J. Mease, Mikkel Ostergaard, Lee Simon, Jasvinder A. Singh, Vibeke Strand and Désirée M. van der Heijde

J Rheumatol 2014;41;1005-1010
<http://www.jrheum.org/content/41/5/1005>

1. Sign up for TOCs and other alerts
<http://www.jrheum.org/alerts>
2. Information on Subscriptions
<http://jrheum.com/faq>
3. Information on permissions/orders of reprints
http://jrheum.com/reprints_permissions

The Journal of Rheumatology is a monthly international serial edited by Earl D. Silverman featuring research articles on clinical subjects from scientists working in rheumatology and related fields.

Updating the OMERACT Filter: Discrimination and Feasibility

George Wells, Dorcas E. Beaton, Peter Tugwell, Maarten Boers, John R. Kirwan, Clifton O. Bingham III, Annelies Boonen, Peter Brooks, Philip G. Conaghan, Maria-Antonietta D'Agostino, Maxime Dougados, Daniel E. Furst, Laure Gossec, Francis Guillemain, Philip Helliwell, Sarah Hewlett, Tore K. Kvien, Robert B. Landewé, Lyn March, Philip J. Mease, Mikkel Ostergaard, Lee Simon, Jasvinder A. Singh, Vibeke Strand, and Désirée M. van der Heijde

ABSTRACT. The “Discrimination” part of the OMERACT Filter asks whether a measure discriminates between situations that are of interest. “Feasibility” in the OMERACT Filter encompasses the practical considerations of using an instrument, including its ease of use, time to complete, monetary costs, and interpretability of the question(s) included in the instrument. Both the Discrimination and Reliability parts of the filter have been helpful but were agreed on primarily by consensus of OMERACT participants rather than through explicit evidence-based guidelines. In Filter 2.0 we wanted to improve this definition and provide specific guidance and advice to participants. (First Release April 1 2014; J Rheumatol 2014;41:1005–10; doi:10.3899/jrheum.131311)

Key Indexing Terms:

OUTCOME MEASURES IN RHEUMATOLOGY
OUTCOME AND PROCESS ASSESSMENT

FEASIBILITY
DISCRIMINATION

From the Cardiovascular Research Methods Centre, Department of Epidemiology and Community Medicine, University of Ottawa; Department of Medicine, University of Ottawa, Ottawa, Ontario, Canada; Departments of Epidemiology and Biostatistics, and Rheumatology, VU University Medical Center, Amsterdam, The Netherlands; University of Bristol Academic Rheumatology Unit, Bristol Royal Infirmary, Bristol, UK; Department of Occupational Sciences and Occupational Therapy, Institute for Health Policy Management and Evaluation, University of Toronto, Toronto, Ontario, Canada; Division of Rheumatology, Johns Hopkins University, Baltimore, Maryland, USA; Department of Internal Medicine, Division of Rheumatology, Maastricht University Medical Center and Caphri Research Institute, Maastricht University, Maastricht, The Netherlands; Australian Health Workforce Institute (AHWI), School of Population Health, University of Melbourne, Melbourne, Australia; Division of Musculoskeletal Disease, University of Leeds, and UK National Institute for Health Research (NIHR) Leeds Musculoskeletal Biomedical Research Unit, Leeds, UK; Versailles-Saint Quentin En Yvelines University, Department of Rheumatology, Ambroise Paré Hospital, APHP, Boulogne-Billancourt; Paris-Descartes University, Medicine Faculty, APHP, Cochin Hospital, Rheumatology B, Paris, France; Section of Rheumatology, Cardiff University School of Medicine, Cardiff, UK; Department of Rheumatology, Geffen School of Medicine at the University of California in Los Angeles; Los Angeles, California, USA; Université Pierre et Marie Curie (UPMC) - Paris 6, GRC-UMPC 08 (EEMOIS); AP-HP Pitié Salpêtrière Hospital, Department of Rheumatology, Paris; Université de Lorraine, Université Paris Descartes, Nancy, France; University of Leeds, Section of Musculoskeletal Disease, LMM Chapel Allerton Hospital, Leeds West Yorkshire; University of the West of England, Academic Rheumatology Unit, Bristol Royal Infirmary, Bristol, UK; Department of Rheumatology, Diakonhjemmet Hospital, Oslo, Norway; Department of Clinical Immunology and Rheumatology, Academic Medical Center, University of Amsterdam and Atrium Medical Center Heerlen, The Netherlands; Institute of Bone and Joint Research, Sydney Medical School and School of Public Health, University of Sydney, Sydney; Department of Rheumatology, Royal North Shore, St. Leonards, New South Wales, Australia; Seattle Rheumatology Associates, Swedish Medical Center Rheumatology Research Division, University of

Washington School of Medicine, Seattle, Washington, USA; Department of Rheumatology, Copenhagen University Hospital at Glostrup, Copenhagen, Denmark; SDG LLC, Cambridge, Massachusetts; Division of Immunology/Rheumatology, University of Alabama at Birmingham and Birmingham VA Medical Center, Birmingham, Alabama; Division of Immunology/Rheumatology, Stanford University School of Medicine, Palo Alto, California, USA; Department of Rheumatology, Leiden University Medical Center, Leiden, The Netherlands.

G.A. Wells, PhD, MSc, Director, Cardiovascular Research Methods Centre, Department of Epidemiology and Community Medicine, University of Ottawa; P. Tugwell, MD, Department of Medicine, University of Ottawa; M. Boers, MD, PhD, Departments of Epidemiology and Biostatistics, and Rheumatology, VU University Medical Center; J.R. Kirwan, MD, University of Bristol Academic Rheumatology Unit, Bristol Royal Infirmary; D.E. Beaton, BScOT, PhD, Department of Occupational Sciences and Occupational Therapy, Institute for Health Policy Management and Evaluation, University of Toronto; C.O. Bingham III, MD, Division of Rheumatology, Johns Hopkins University; A. Boonen, MD, PhD, Department of Internal Medicine, Division of Rheumatology, Maastricht University Medical Center and Caphri Research Institute, Maastricht University; P. Brooks, MD, Director, AHWI, School of Population Health, University of Melbourne; P.G. Conaghan, MB, BS, PhD, FRACP, FRCP, Division of Musculoskeletal Disease, University of Leeds, and NIHR Leeds Musculoskeletal Biomedical Research Unit; M-A. D'Agostino, MD, Versailles-Saint Quentin En Yvelines University, Department of Rheumatology, Ambroise Paré Hospital, APHP; M. Dougados, MD, Paris-Descartes University, Medicine Faculty, APHP, Cochin Hospital, Rheumatology B; D.E. Furst, MD, Department of Rheumatology, Geffen School of Medicine at the University of California in Los Angeles; L. Gossec, MD, PhD, UPMC, GRC-UMPC 08 (EEMOIS); AP-HP Pitié Salpêtrière Hospital, Department of Rheumatology; F. Guillemain, MD, PhD, Université de Lorraine, Université Paris Descartes; P. Helliwell MD, University of Leeds Section of Musculoskeletal Disease, LMM Chapel Allerton Hospital; S. Hewlett, PhD, RN, University of the West of England, Academic Rheumatology Unit, Bristol Royal Infirmary; T.K. Kvien, Department of Rheumatology, Diakonhjemmet Hospital;

Personal non-commercial use only. The Journal of Rheumatology Copyright © 2014. All rights reserved.

R.B. Landewé, MD, Department of Clinical Immunology and Rheumatology, Academic Medical Center, University of Amsterdam and Atrium Medical Center Heerlen; L. March, MBBS, MSC, PHD, FRACP, FAFPHM, Director, Institute of Bone and Joint Research and Conjoint Professor, Sydney Medical School and School of Public Health, University of Sydney, Senior Staff Specialist, Department of Rheumatology, Royal North Shore; P.J. Mease, MD, Seattle Rheumatology Associates, Chief, Swedish Medical Center Rheumatology Research Division, Clinical Professor, University of Washington School of Medicine; M. Østergaard, MD, DMSc, Department of Rheumatology, Copenhagen University Hospital at Glostrup; L.S. Simon, MD, SDG LLC; J.A. Singh, MD, MPH, Division of Immunology/Rheumatology, University of Alabama at Birmingham and Birmingham VA Medical Center; V. Strand, MD, Division of Immunology/Rheumatology, Stanford University School of Medicine; D.M. van der Heijde, MD, PhD, Department of Rheumatology, Leiden University Medical Center.

Address correspondence to Dr. Wells, University of Ottawa, Epidemiology and Community Medicine, 40 Ruskin St., Ottawa, Ontario K1H8M5, Canada. E-mail: gawells@ottawaheart.ca

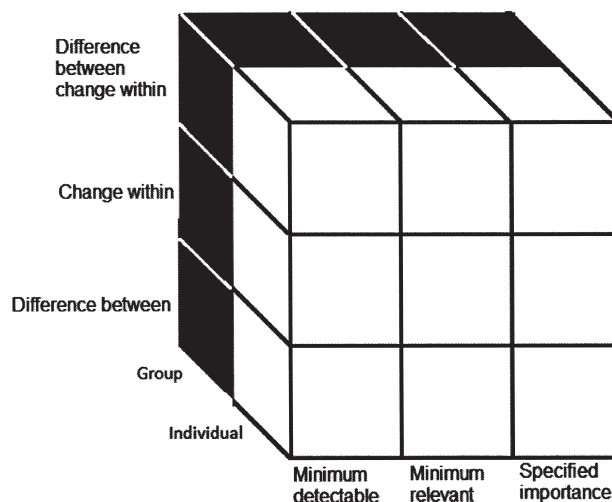
Discrimination

The “Discrimination” part of the Outcome Measures in Rheumatology (OMERACT) Filter asks whether the measure discriminates between situations that are of interest. The situations can be states at one time (for classification or prognosis) or states at different times (to measure change). The word captures the issues of reliability and sensitivity to change (responsiveness). The “Discrimination” part of the filter has been helpful but was agreed on primarily by consensus of OMERACT participants rather than through explicit evidence-based guidelines. In Filter 2.0 we want to improve this definition and provide specific guidance and advice to participants.

Various conceptual models have been considered in OMERACT for discrimination. For example, a classification system for studies of discrimination was designed to help organize the specific purpose of such studies, and identify those with the potential to provide information on minimal clinically important difference (MCID). A 3-dimensional cube was developed¹; and a simplified version of the cube is provided in Figure 1 into which studies of discrimination can be categorized based on their evaluation of 3 attributes: (1) Setting, which identifies whether the study results were targeted (a) to individuals; or (b) to groups; (2) Comparison, which identifies whether discrimination was considered as (a) differences between individuals or groups at 1 point in time; (b) change within individuals or groups over time; or (c) differences in the change within individuals or groups over time; and (3) Extent of difference, which identifies whether the difference being assessed is (a) the minimum detectable; (b) the minimum relevant or important; or (c) a higher and possibly specified level of importance.

This classification system helps to focus attention on the specific type of discrimination of interest in specific assessment circumstances. It reinforces an understanding that an instrument that is able to discriminate between states as represented by 1 cell within the cube will not necessarily

Comparison



Extent of Difference

Figure 1. Simplified cube of discrimination. From Beaton, et al. J Rheumatol 2001;28:406-12.

be able to discriminate between states as represented by another cell. In particular, it is easiest for an instrument to show discrimination in the bottom, left, front corner; and most difficult in the top, right, back corner of the cube.

In the context of clinical trials the setting is usually treatment groups and the comparison is change within groups — more particularly differences between changes within groups. The sensitivity required may depend on the person involved (e.g., physician, patient, or policy maker; group or individual) and the intended use (e.g., clinical service design or research exploration). Various measures have been proposed for considering important changes and states including the MCID², which is not without its critics^{3,4,5,6}; and more recently the patient acceptable symptom state (PASS)⁷. Most have been targeted to the individual (patient) level. Other novel methodologies may be informative, such as identifying a system of levels of improvements that, for example, patients are striving to achieve.

As part of the research agenda, various paradigms for considering discrimination that integrate the different measures, perspectives, and purposes were explored. Similarly, while different methods used to define MCID or clinically important changes or important differences were previously considered and categorized according to a version of the “cube” classification system⁸, literature on other methods within the derived paradigm will be considered. Also, OMERACT working groups and patient groups were consulted to identify different ways that results can be viewed and the extent of discrimination was assessed by physicians and patients in their area of interest.

Feasibility

Feasibility in the OMERACT Filter encompasses the practical considerations of using an instrument, including its ease of use, time to complete, monetary costs, and interpretability of the question(s) included in the instrument. Considerations include cost of equipment, training for observers, burden/difficulty for the patient, and in the case of patient self-report, the perceived length, wording of questions, reading level, and ease of response options (clarity, ease of retrieval of information, ease of responding on that scale).

Feinstein coined the term “sensitivity” to reflect an enlightened common sense appraisal of the instrument under consideration^{9,10}. He is credited with encouraging the clinical research community to accept that this is as important as some statistically based measurement property. In Feinstein’s framework, feasibility is a key concern addressed by 6 questions: (1) Is it easy to understand? (2) Are the items, their scaling, and the aggregate score simple and easy to use? (3) Does the data collection sheet conform to basic principles of questionnaire design; are there instructions and definitions provided and are procedures standardized? (4) Is it acceptable to the patient/participant and to the observer? (5) Is the format for administration appropriate for your purpose or does it require special tests or special skills? (6) Is the administration time suitable?

Auger, *et al* reviewed potential instruments to conduct this “common sense appraisal” and suggested the following main domains for assessment of feasibility (termed “applicability”): respondent burden, examiner burden, distributional issues, and format issues¹¹.

The questions asked within Feinstein’s feasibility assessment are consistent with the domains and subdomains described by Auger (Figure 2). Note that OMERACT has traditionally used the term “applicable” to a measurement

instrument that has passed all the filter requirements of Truth, Discrimination, and Feasibility¹².

Feasibility is most often appraised by a researcher or clinician who is selecting the instrument. And it is the most frequently and quickly endorsed step in the OMERACT Filter. In Filter 2.0 we are seeking a more thoughtful reflection on each of the components, and consideration of a combined point of view from the researcher/clinician and patients. We proposed a merger of Feinstein and Auger’s points.

Breakout Discussion Groups

Following a plenary presentation of the topics reviewed above, conference participants were divided into 5 pairs of breakout (discussion) groups. Examples of the conduct of some discrimination exercises that had already taken place in different areas of OMERACT activity were presented to each pair of breakout groups. These were taken from work on gout, ultrasound, psoriatic arthritis, MCID, and worker productivity (Table 1, column 1). They served to provide concrete examples of the discrimination issues being addressed, and to help the discussions focus on the main questions of discrimination and feasibility for each breakout group.

Report Back and Plenary Discussion

The summary of the discussions from the breakout groups is provided in Table 1.

Discrimination

Two groups deliberated on general methods and procedures for assessing discrimination when only non-inferiority head-to-head trial data are available. The following points were noted. If a current standard treatment is effective, then placebo-controlled trials may not be possible, because they

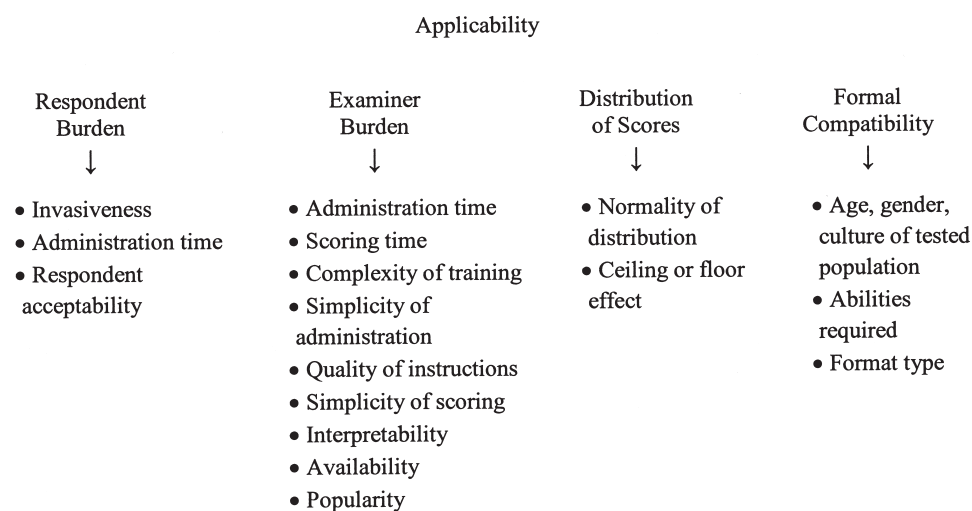


Figure 2. Domains of applicability, after Feinstein⁹ and Auger¹¹.

Table 1. Summary of the report back and discussion from the breakout groups.

Discrimination Topic	Specific Discrimination Topic	How Was the Discrimination Topic Addressed?
Determining differences when only non-inferiority trial evidence available	What general methods and procedures could be considered for determining differences (a) between groups; (b) response in a patient when only non-inferiority head-to-head trial data are available?	<ul style="list-style-type: none"> Assess assay sensitivity of treatment to active control and constancy of active control to placebo Use adjusted indirect methods for deriving indirect treatment comparison of treatment to placebo Use indirect estimates as one uses direct estimates for minimum detectable, minimum important, and major differences
Determining minimum detectable, minimum important, and major differences	What general methods and procedures could be considered for determining minimum detectable, minimum important, and major differences (a) between groups; (b) response in a patient?	<ul style="list-style-type: none"> Consider the “cube of discrimination” for looking at changes within patients and at differences between groups Anchor vs distribution based methods Consider contextual factors Importance of scaling method used and signal-to-noise ratio Determinations for improvement may not be same as for worsening Need for patient involvement
Determining the sensitivity to measure	What different “situations of interest” could be considered for assessing discrimination?	<ul style="list-style-type: none"> Consider in terms of providing examples for study designs and clinical practice for each situation in “cube of discrimination” Need for 2 RCT questioned
Determining a responder index	What general methods and procedures could be considered for determining a responder index?	<ul style="list-style-type: none"> Index must be sensitive to both improvement and worsening Standardization of the technique is needed in order to assess change
Practical checklist for discrimination	What items constitute a practical checklist for discrimination?	<ul style="list-style-type: none"> Filter 1 presentation best way but more examples recommended Basic principle for responsiveness — consider a clinical trial where there is a known treatment effect and then assess the change in the outcome Look for responsiveness first at group level then at individual level Anchor treatment effect to a PRO if want it to be understandable to patient COSMIN was not discussed in great detail and this must be more closely considered Role of effect size important

*How can the feasibility of a measure be made, taking into consideration aspects such as cost, burden, and interpretability? RCT: randomized controlled trial; PRO: patient reported outcomes; COSMIN: COnsensus-based standards for the selection of Health Measurement INstruments.

are likely unethical. New treatments can then only be compared with active treatments so there will be no comparison of the new treatment against placebo and hence no measure of the “actual” effect of the new treatment. If superiority of the new treatment is not anticipated, but the new treatment may be safer, cheaper, and/or easier to administer, then a head-to-head non-inferiority trial could be considered. A non-inferiority trial is designed to demonstrate the efficacy of a new treatment by showing that it is not less efficacious than the active control (standard treatment) by more than a specified margin, known as the non-inferiority margin. An important fact is that a well-designed and properly conducted non-inferiority trial that correctly demonstrates the treatments to be similar cannot be distinguished in itself from a poorly executed trial that fails to find a true difference. The ability of a trial to demonstrate a difference between treatments, if such a difference truly exists, is known as “assay sensitivity.” A non-inferiority trial that finds the effects of the treatments to be similar has not demonstrated assay sensitivity, and

must rely on an assumption of assay sensitivity on the basis of information external to the trial. Use of past placebo-controlled trials may accomplish this, and we must have available historical data in which it has been established that standard treatment is superior to placebo. Further, we must have constancy, namely, that the historical difference between the standard and placebo is assumed to hold in the setting of the new trial if a placebo control had been used. How to use information from these trials to determine pertinent differences between groups and within patients must be identified. We would now have available direct evidence comparing the new treatment to standard treatments, and standard treatment to placebo; and, in using the standard as the common linking treatment, we can consider indirect evidence of the new treatment to placebo on which we could base these differences.

In making this assessment, in addition to assay sensitivity and constancy, an adjusted indirect treatment comparison method must be used in which the comparison of the treatments of interest is adjusted by results of their direct

comparison with the standard group, thus partially using the strength of the randomized controlled trial (RCT). In the simplest, yet widely applicable setting, the method by Bucher, *et al*¹³, and generalized by Wells, *et al*¹⁴, is one such method.

Two breakout groups considered what general methods and procedures could be considered for determining minimum detectable, minimum relevant (important), and major differences. The breakout groups reviewed the “discrimination cube,” looking at changes within patients and at differences between groups. Considering the MCID, several issues were raised and discussed. More consideration and explanation is needed on whether the MCID applies only to patient-reported outcome (PRO) measures or whether it also applies to composite measures. Can physicians determine what an MCID is for an objective measure such as the erythrocyte sedimentation rate (ESR)? The scaling method used (e.g., numeric rating scale vs Likert scaling) may change the signal-to-noise ratio. For example, the MCID established using the anchor-based method and distribution-based method for the Health Assessment Questionnaire (HAQ) in psoriatic arthritis with etanercept was greater than that for rheumatoid arthritis, which may be a problem when considering a nonlinear score such as the HAQ. The MCID calculated for improvement may not be the same as the MCID for worsening; in this regard, it has been found that the MCID for improvement is more than that for deterioration for health-related quality of life in systemic lupus erythematosus¹⁵.

Patient involvement in the determination of MCID was raised, noting that the choice of anchors can determine the MCID, and there may be importance in obtaining patient input for determining the appropriate question. And it may make a difference, if asking about the state you would be comfortable in, if the inquiry takes the form of a global assessment, PASS, or the determination of the amount of change. Also the anchor question may be dependent on the study design or the primary outcome, and for a composite measure examining multiple aspects of disease (e.g., skin/joint), the MCID may be different for different areas involved.

Finally, the determination of MCID may be dependent on contextual factors such as the initial disease state and the disease experience (including duration and coping mechanisms), which could lead to a response shift in the MCID as well as expectations for treatment.

Two other breakout groups considered what different “situations of interest” could be considered for assessing discrimination. The difficulty in defining “situations of interest” may be most usefully considered in terms of providing examples for RCT, longitudinal observational studies, and clinical practice for each situation in the “cube of discrimination.” In particular, OMERACT has required information from 2 RCT before endorsing a measure or a

responder index, and the need for 2 RCT was questioned. There may be a situation where it is not possible to obtain results from 2 RCT, or there is a negative trial with no difference in outcome between groups.

Two other breakout groups deliberated on what general methods and procedures could be considered for determining a responder index. A responder index is a combination of a series of indicators, each a threshold value on a measurement instrument. American College of Rheumatology 20% response and European League Against Rheumatism response criteria are reasonable examples of responder indices. Such an index should ideally be discriminative in situations of both improvement and worsening.

The final pair of breakout groups each considered what items constitute a practical checklist for discrimination. While noting the CONsensus-based standards for the selection of Health Measurement INSTRUMENTS (COSMIN) checklist¹⁶, discussion on this was limited because necessary information was lacking and time was too short. One breakout group noted that discrimination as presented in Filter 1¹² appears to be the best way to proceed, but there is a need for more examples. To test responsiveness, the group believed that one should consider an RCT where there is a known treatment effect and then assess the change in the outcome of interest. Further, one should look for responsiveness first at group level, and then at individual level. The treatment effect should be anchored to a PRO to be best understood by patients. The discussion on diminishing the role of effect size in assessing discrimination was not resolved.

Feasibility

Three pairs of breakout groups also deliberated on how the feasibility of a measure or group of measures could be assessed, taking into consideration aspects such as cost, burden, and interpretability. Two groups felt feasibility should be considered early, as part of the “development loop” from the beginning to the end of the process in developing an instrument, and that pilot testing for feasibility should be conducted. The issue of capturing information on paper versus by computer in different sites globally was raised as an important issue related to feasibility. Longitudinal data capture and frequency of data collection can be burdensome issues related to feasibility and, for that matter, to validity. The possibility of developing a feasibility “score” was raised.

In summary, this OMERACT session was designed to evaluate key aspects of discrimination and feasibility proposed for the Filter 2.0¹⁷ framework that had been discussed prior to the meeting and/or arisen as issues in using the discrimination and feasibility descriptions of the original filter over the years. Using specific topics on discrimination and a general question on feasibility raised in questions, OMERACT 11 participants were able to probe

the theoretical and practical implications of the framework and examine areas of strength and weakness. Although specific aspects and issues were raised regarding the various topics related to discrimination and feasibility, providing guidance and research agenda topics, there was general agreement that the more explicit explanations of discrimination and feasibility to be included in Filter 2.0 would help developers of core outcome measures.

REFERENCES

1. Beaton DE, Bombardier C, Katz JN, Wright JG, Wells G, Boers M, et al. Looking for important change/differences in studies of responsiveness. OMERACT MCID Working Group. Outcome measures in rheumatology. Minimal clinically important difference. *J Rheumatol* 2001;28:400-5.
2. Jaeschke R, Singer J, Guyatt GH. Ascertaining the minimal clinically important difference. *Cont Clin Trials* 1989;10:407-15.
3. Kirwan JR. Minimum clinically important difference: The crock of gold at the end of the rainbow. *J Rheumatol* 2001;28:439-44.
4. Rennard SI. Minimal clinically important difference, clinical perspective: an opinion. *COPD* 2005;2:51-5.
5. Beaton DE, Boers M, Wells GA: Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. *Curr Opin Rheumatol* 2002; 14:109-14.
6. Gatchel RJ, Mayer TG. Testing minimal clinically important difference: consensus or conundrum? *Spine J* 2010;10:321-7.
7. Dougados M, Moore A, Yu S, Gitton X. Evaluation of the patient acceptable symptom state in a pooled analysis of two multicentre, randomised, double-blind, placebo-controlled studies evaluating lumiracoxib and celecoxib in patients with osteoarthritis. *Arthritis Res Ther* 2007;9:R11.
8. Wells G, Beaton D, Shea B, Boers M, Simon L, Strand V, et al. Minimal clinically important differences: review of methods. *J Rheumatol* 2001;28:406-12.
9. Feinstein AR. *Clinimetrics*. New Haven: Yale University Press; 1987.
10. Rowe BH, Oxman AD. An assessment of the sensibility of a quality of life instrument. *Am J Emerg Med* 1993;11:374-80.
11. Auger C, Demers L, Swaine B. Making sense of pragmatic criteria for the selection of geriatric rehabilitation measurement tools. *Arch Gerontol Geriatr* 2006;43:65-83.
12. Boers M, Brooks P, Strand V, Tugwell P. The OMERACT Filter for outcome measures in rheumatology. *J Rheumatol* 1998;25:198-9.
13. Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *J Clin Epidemiol* 1997;50:683-91.
14. Wells G, Sultan S, Chen L, Coyle D. Indirect evidence: indirect treatment. Comparisons in meta-analysis. Ottawa: Canadian Agency for Drugs and Technology in Health (CADTH) Health Technology Report; 2009.
15. Strand V, Crawford B. Improvement in health-related quality of life in patients with SLE following sustained reductions in anti-dsDNA antibodies. *Expert Rev Pharmacoecon Outcomes Res* 2005; 5:317-26.
16. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN Checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. *Qual Life Res* 2010;19:539-49.
17. 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 Feb 27 (E-pub ahead of print).