Patient Perspective: Choosing or Developing Instruments

JOHN R. KIRWAN, JAMES F. FRIES, SARAH HEWLETT, and RICHARD H. OSBORNE

ABSTRACT. Previous Outcome Measures in Rheumatology (OMERACT) meetings recognized that patients view outcomes of intervention from a different perspective. This preconference position paper briefly sets out 2 patient-reported outcome (PRO) instrument approaches, the PROMISE computer adaptive testing (CAT) system and development of a rheumatoid arthritis-specific questionnaire to measure fatigue; a tentative proposal for a PRO instrument development pathway is also made. (J Rheumatol 2011; 38:1716–19; doi:10.3899/jrheum.110390)

Key Indexing Terms:
PATIENT REPORTED OUTCOMES OUTCOME ASSESSMENT OMERACT

Previous Outcome Measures in Rheumatology (OMERACT) meetings recognized that patients view outcomes of intervention from a different perspective than do researchers, and this should be reflected in the way effects of new treatments are assessed and reported[1,2]. The "patient perspective" has become an integral part of OMERACT proceedings, and has led to the emergence of a broader concept of outcome assessment[3], incorporating the notion of measuring the impact of arthritis and its treatment on the lives of individual patients[4,5]. However, to do so requires a greater reliance on patient-reported outcomes (PRO) in domains not yet well characterized within rheumatology, and for which the choice of measuring instrument is not yet well defined. The regulatory authorities are interested in PRO, and the US Food and Drug Administration recently issued guidelines on how such PRO might be chosen, developed, or justified[6]. One aim of OMERACT 10 is to identify steps by which we might ensure that appropriate instruments are chosen or developed. If agreement can be reached on an appropriate and rigorous process for developing PRO instruments, then recognition of the validity of such instruments by the research community will be strongly facilitated. Our objective is to promote the emergence of consensus agreement on such a process.

This preconference position paper briefly sets out 2 PRO instrument approaches: the patient-reported outcome information system (PROMIS) computer adaptive testing (CAT) system and the development of a rheumatoid arthritis (RA)-specific questionnaire to measure fatigue, and makes a tentative proposal for a PRO instrument development pathway. The outlines will be elaborated during 3 plenary presentations (by J. Fries, S. Hewlett, R. Osborne, respectively) and, augmented by additional brief reports, will form the basis of the workshop discussion groups.

From the University of Bristol, Academic Rheumatology Unit, Bristol Royal Infirmary, Bristol, UK; Stanford University School of Medicine, Palo Alto, California, USA; University of the West of England, Academic Rheumatology Unit, Bristol Royal Infirmary, Bristol, UK; and Deakin University, Public Health Innovation, Melbourne, Australia.

J.R. Kirwan, MD, University of Bristol, Academic Rheumatology Unit, Bristol Royal Infirmary; J.F. Fries, MD, Stanford University School of Medicine; S. Hewlett, PhD, University of the West of England, Academic Rheumatology Unit, Bristol Royal Infirmary; R.H. Osborne, BSc, PhD, Deakin University, Public Health Innovation.

Address correspondence to Prof. Kirwan.
E-mail: john.kirwan@bristol.ac.uk.

The PROMIS Approach to Choosing or Developing Instruments

In 2004, the USA National Institutes of Health (NIH) initiated a multicenter cooperative group establishing the PROMIS[7,8]. The PROMIS network of clinicians, clinical researchers, and measurement experts is organized around 6 primary research sites and a statistical coordinating center with the aim of building and validating common, accessible item banks to measure key symptoms and health concepts applicable to a range of chronic conditions, enabling efficient and interpretable clinical trial research and clinical practice application of PRO. With this item bank as a basis, it aims to create a CAT system that allows efficient, psychometrically robust assessment of PRO in clinical trial research involving a wide range of chronic diseases, including musculoskeletal diseases. One application within rheumatology has been the assessment of physical function, wherein an important goal is to circumvent the rigidity of current questionnaires, and to eliminate their floor and ceiling effects when used in patients whose physical function differs from the average[9]. A recent report documents the qualitative and quantitative item-evaluation process for developing the PROMIS Physical Function item bank[10].

In the PROMIS approach to choosing or developing instruments, the process starts with the item. Instruments, in the PROMIS and OMERACT sense, are groups of questionnaire items which together can provide a reliable and valid estimate of a "latent trait," often called a "domain." The approach begins with selection and definition of the latent trait, perhaps from a hierarchy or table of possible domains. For example,

"Physical Function," which has a subdomain of "Mobility" and is itself a subdomain of "Physical Health." It is desirable to have as few domains as possible and for everyone to use compatible processes, but sometimes a project may require development of a new domain.

Next, items that have been used to estimate the domain historically are assembled by an exhaustive literature review that may yield thousands of items from hundreds of instruments. If gaps and omissions are suspected, new items may be written. Items are screened for clarity, translatability, patient importance, and other attributes employing focus groups and cognitive reviews, rewritten as necessary, and binned into similar categories, such as "Walking." For an item to function well in IRT (Item Response Theory), it needs (1) to contribute to a unidimensional construct and (2) to be nonredundant to other items (local independence). Thus, items are winnowed within bins to reduce numbers of items for study yet preserve content validity. Item stems, response options, and timeframes may be rendered uniform for consistency.

Then, item banks undergo validity testing in multiple diverse populations, with each item having a minimum of 250 and for some applications 750 or more respondents. From these, items are quantitatively calibrated using IRT for their degree of difficulty and their ability to discriminate. The resulting Physical Function Core Item bank currently has 154 items. From the calibrated item bank, the items with the greatest information content can be selected and aggregated into instruments, which will necessarily outperform legacy instruments on all measures[11,12]. Usually, however, this process is mediated by sentient beings, with the goal to select the strongest items that collectively are content-balanced across the domain.

There presently are 4 kinds of PROMIS instruments: First, improved generic short forms, such as the PROMIS PF-20, the improved Health Assessment Questionnaire Disability Index, or the improved PF-10. Second, improved focused short forms, which are tailored, for example, for a severely impaired population or a very healthy one; there can be a number of such short forms, each containing different, sometimes overlapping, items. The instruments are nearly as efficient as CAT. Third, the PROMIS profiles, where items from several domains are administered together, perhaps 3 to 7 items per included domain; results can efficiently profile the individual across disparate domains. The legacy Medical Outcome Study Short-Form 36 is essentially a profile instrument. Finally, there are CAT instruments, where items are sequentially administered to a subject based upon responses to prior items, in a dynamic test. In research mode a "simulated" CAT is generally used, where data on all items are collected, and then simulated runs made on the entire data bank. CAT instruments are extremely efficient, and ultimately will be required for populations with individuals at the extremes of scores. CAT can allow the same instrument to be administered in a rehabilitation setting and in a setting of healthy students. In full CAT mode the number of available items can be large, and the questionnaire burden quite low[13].

In validation studies, PROMIS instruments are run head to head with legacy instruments, looking specifically for greater sensitivity to change[14], with greater effect sizes, and smaller sample size requirements[15].

The Example of Fatigue

Fatigue, a major and common problem for people with RA, is recommended for assessment in all clinical trials alongside the core set[16,17]. Patients describe a range of features of RA fatigue and multiple consequences[18,19,20,21], yet most scales in use are generic and have not been adequately standardized or validated in RA[21]. While these may (overall) provide reasonable answers[22], most do not fulfil current, rigorous guidelines for PRO[6]. In addition, they yield global scores, yet there is the possibility of different facets of fatigue[18,19,20], which may potentially have important causal and therapeutic implications.

In order to develop and evaluate new RA short scales (for severity, impact, and perceived coping) and a multidimensional questionnaire (MDQ) assessing separate components, a series of studies were conducted[23], following recommended methodology[6], and with decisions discussed with and supported by a patient research partner at all timepoints. In study 1, qualitative analyses of interviews with patients identified descriptors and language (to inform questionnaire phrasing) and identified features of RA fatigue that might exemplify potential separate fatigue dimensions. In study 2, these descriptors and concepts were debated by patient focus groups, who helped design the short scales. Next, the short scales and a draft 45-item MDQ were subject to cognitive interviewing (study 3), in which patients completed them while "thinking aloud" to the researcher, allowing identification of potential problems with understanding the questions in the way the researcher had intended, or with response options[24].

Having used qualitative methods to address face and content validity, the resultant Bristol RA Fatigue (BRAF) short scales were evaluated for construct validity against appropriate variables in a large cohort (study 4)[25]. In the same cohort, the draft 45-item BRAF-MDQ was subject to an iterative process of Cronbach's alpha for internal consistency, factor analysis for dimensions, and bootstrapping for stability of the factor analysis, alongside clinical judgments to inform removal of less informative items and retention of more informative items. The resulting 20-item BRAF-MDQ was tested for construct validity (Spearman's correlation) with appropriate variables. This produced a robust scale that yields a global fatigue score and 4 dimension scores: physical fatigue (severity), living with fatigue (everyday life consequences), emotional fatigue, and cognitive fatigue[26]. Exploration of the BRAF short scales shows that patients with similar fatigue severity scores can differ greatly in their perceived coping and

impact scores. Similarly, patients with the same global BRAF-MDQ scores can have different profiles for the 4 subscales of Physical, Living, Emotional, and Cognitive fatigue.

Application of careful and thorough modern methodologies, grounded in the patient perspective, with decisions enhanced by a patient research partner throughout these studies, has produced scales with potential for individualized interventions and assessment.

A Methodological Proposal for Developing or Choosing Instruments (Questionnaires)

Patients complete questionnaires across clinical, survey, and experimental settings. Questionnaire data are used to make substantive decisions about patient care, the value of treatments in clinical trials, patient education programs, and the quality of healthcare professionals and the institutions in which they work. Questionnaires therefore carry enormous responsibility.

The adage "garbage in, garbage out" is pertinent in questionnaire assessments. Data are invalid if questionnaires comprise imprecise or misdirected questions, or do not reflect target concepts. How do we judge the fidelity of a question or questionnaire? The answer to this therefore requires input from several disciplines. Most importantly, we need to scrutinize the groundwork undertaken to develop the questionnaire, which should consist of high quality consultation with 2 groups of people — patients and clinicians.

A key aspiration of OMERACT is to improve endpoint outcome measurement through a data-driven, iterative consensus process involving relevant stakeholder groups[27]. The first of the OMERACT filters[28], Truth, asks: Is the measure truthful? Does it measure what it intends to measure? Is the result unbiased and relevant? This criterion captures issues of face, content, construct, and criterion validity. Questionnaire design must be founded on developing questions with face, content, construct, and criterion fidelity. If this step fails, subsequent steps are flawed. No amount of subsequent statistical calculation will make such a questionnaire valid. It is imperative that adequate resources are used in the initial development of the question concept, construct refinement, and verification of the fidelity between the intention of the question and patient responses.

Key Steps for Initial Question and Questionnaire Development

1. The construct ("thing") to measure. Motivation for questionnaire development in the OMERACT context generally evolves from an observed clinical or public health phenomenon, which forms the construct needing systematic evaluation. The precise scope and purpose of the proposed questionnaire must be explicit.
2. Verification of the construct in target patient populations. Qualitative processes such as direct observation, interviews, and focus groups are used to verify the construct in patient samples.
3. Does the construct have properties amenable to measurement? Subjective experiences such as fatigue, engagement in life, and self-efficacy are difficult to measure, while others (e.g., gait) are observable. Highly structured analytical processes are required to organize everyday patient experiences to generate a measurable construct.
4. Question wording should emanate from patients and the patient-clinician interaction, and be faithful to the purpose of the questionnaire. Explicit rules must govern question development to ensure they are appropriate for the target patient group (i.e., literacy level, gender, body part, age), and that the resulting data covers the breadth of the target construct and informs the target audience (clinicians, researchers, policymakers).
5. Drafted and checked questions undergo cognitive testing with naive patient samples. Cognitive testing questions can include "Can you tell me how you came to that answer?".

Questions should resonate with patient thinking and observable behavior, available treatments, and should be relevant, meaningful, and modifiable. Sufficient excellent questions permit straightforward construct validation procedures, with question choices reflecting the defined construct and the purpose of the questionnaire, not the desire for an elegant or parsimonious statistical model.

This approach starts with reflective clinician/researcher observations, followed by a process to ensure faithfulness to how patients view the world. The validation processes prioritize veracity for the intended construct and purpose of the nascent questionnaire.

REFERENCES

1. Kirwan J, Heiberg T, Hewlett S, Hughes R, Kvien T, Ahlmèn M, et al. Outcomes from the patient perspective workshop at OMERACT 6. J Rheumatol 2003;30:868-72.
2. Kirwan JR, Hewlett S, Heiberg T, Hughes R, Carr M, Hehir M, et al. Incorporating the patient perspective into outcome assessment in rheumatoid arthritis — progress at OMERACT 7. J Rheumatol 2005;32:2250-6.
3. Kirwan J, Newman S, Tugwell P, Wells G. Patient perspective on outcomes in rheumatology — A position paper for OMERACT 9. J Rheumatol 2009;36:2067-70.
4. Sanderson T, Kirwan J. Patient-reported outcomes for arthritis: time to focus on personal life impact measures? Arthritis Rheum 2009;61:1-3.
5. Kirwan J, Newman S, Tugwell P, Wells G, Hewlett S, Idzera L, et al. Progress on incorporating the patient perspective in outcome assessment in rheumatology and the emergence of life impact measures at OMERACT 9. J Rheumatol 2009;36:2071-6.
6. US Department of Health and Human Services, Food and Drug Administration. Guidance for Industry: Patient-reported outcome measures: Use in medical product development to support labelling claims. [Internet. Accessed March 29, 2011.] Available from: http://www.fda.gov/downloads/Drugs/GuidanceCompliance RegulatoryInformation/Guidances/UCM193282.pdf
7. Patient-Reported Outcomes Measurement System. [Internet. Accessed March 29, 2011.] Available from: http://www.nihpromis.org/default.aspx
8. Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, et al.

The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH roadmap cooperative group during its first two years. Med Care 2007;45 Suppl 1:S3-S11.
9. Aletaha D. From the item to the outcome: the promising prospects of PROMIS. Arthritis Res Ther 2010;12:104.
10. Bruce B, Fries JF, Ambrosini D, Lingala B, Gandek B, Rose M, et al. Better assessment of physical function: item improvement is neglected but essential. Arthritis Res Ther 2009;11:R191.
11. Fries JF. The promise of the future, updated: better outcome tools, greater relevance, more efficient study, lower research costs. Future Rheumatol 2006;1:415-21.
12. Fries JF, Bruce B, Bjorner J, Rose M. More relevant, precise, and efficient items for assessment of physical function and disability: moving beyond the classic instruments. Ann Rheum Dis 2006;65 Suppl 111:16-21.
13. Fries, JF, Cella D, Rose M, Krishnan E, Bruce B. Progress in assessing physical function in arthritis patients: PROMIS short forms and computerized adaptive testing (CAT). J Rheumatol 2009;36:2061-6.
14. Fries JF, Krishnan E, Rose M, Lingala B, Bruce B. Improved responsiveness of physical function (disability) scales based upon item response [abstract]. Arthritis Rheum 2009;60 Suppl:S229.
15. Rose M, Bjorner JB, Becker J, Fries JF, Ware JE. Evaluation of a preliminary physical function item bank supported the expected advantages of the patient-reported outcomes measurement information system (PROMIS). J Clin Epidemiol 2008;61:17-33.
16. Wolfe F, Hawley DJ, Wilson K. The prevalence and meaning of fatigue in rheumatic disease. J Rheumatol 1996;23:1407-17.
17. Kirwan J, Minnock P, Adebajo A, Bresnihan B, Choy E, de Wit M, et al. Patient perspective: fatigue as a recommended patient centered outcome measure in rheumatoid arthritis. J Rheumatol 2007;34:1174-7.
18. Tack BB. Fatigue in rheumatoid arthritis: conditions, strategies, and consequences. Arthritis Care Res 1990;3:65-70.
19. Hewlett S, Cockshott Z, Byron M, Kitchen K, Tipler S, Pope D, et al. Patients' perceptions of fatigue in rheumatoid arthritis: overwhelming, uncontrollable, ignored. Arthritis Rheum 2005;53:697-702.
20. Repping-Wuts H, Uitterhoeve R, van Riel P, van Achterberg T. Fatigue as experienced by patients with rheumatoid arthritis (RA): a qualitative study. Int J Nurs Stud 2008;45:995-1002.
21. Hewlett S, Hehir M, Kirwan JR. Measuring fatigue in rheumatoid arthritis: a systematic review of scales in use. Arthritis Rheum 2007;57:429-39.
22. Wolfe F. Fatigue assessments in rheumatoid arthritis: comparative performance of visual analog scales and longer fatigue questionnaires in 7760 patients. J Rheumatol 2004;31:1896-902.
23. Nicklin J. The development of scales to measure fatigue in people with rheumatoid arthritis [PhD dissertation]. Bristol, UK: University of West of England; 2009.
24. Nicklin J, Cramp F, Kirwan J, Hewlett S. Using "Think Aloud" to improve questionnaire design [abstract]. Arthritis Rheum 2008;58 Suppl:S868.
25. Nicklin J, Kirwan J, Cramp F, Hewlett S. Development and initial validation of short scales to measure severity, effect and ability to cope with fatigue in RA [abstract]. Rheumatology 2009;48:OP38.
26. Nicklin J, Kirwan J, Cramp F, Urban M, Hewlett S. Development and initial validation of the Bristol RA Fatigue Multi-dimensional scale (BRAF-MDQ). Arthritis Rheum 2009;60 Suppl 10:1204.
27. Tugwell P, Boers M, Brooks P, Simon L, Strand V, Idzerda L. OMERACT: An international initiative to improve outcome measurement in rheumatology. Trials 2007;8:38.
28. Boers M, Brooks P, Strand V, Tugwell P. The OMERACT filter for outcome measures in rheumatology. J Rheumatol 1998;25:198-9.