Workshop Report

Health Status Instruments / Utilities

NICHOLAS BELLAMY, MAARTEN BOERS, DAVID FELSON, JAMES FRIES, DANIEL FURST, DAVID HENRY, MATTHEW LIANG, DANIEL LOVELL, LYN MARCH, VIBEKE STRAND, and SJEF van der LINDEN

ABSTRACT. Rheumatologists and other interested professionals at the OMERACT II conference formed small groups to discuss whether it was sensible to use a generic health status instrument in musculoskeletal disease trials. These instruments promise the possibility of comparison of health status between disease states. However, data is lacking on validity of the current generation of instruments to support their use. Participants had little personal experience with these instruments. After inspection, many voiced strong concerns over comprehensiveness and responsiveness. Many dimensions of health relevant for patients with this group of diseases were felt to be underrepresented. The dimension of adverse effects was universally absent, although this is more a problem of state of the art in trial methodology than a problem of these measures. Although there is little data, the small number of response categories in the dimensions covered, plus the lack of comprehensiveness, make it likely that responsiveness will be low. Further research, especially the adoption of one or more generic measures alongside specific measures, both in trials and in observational studies, is necessary to validate and improve the current generic measures. Until that time, valid conclusions and health policy regarding musculoskeletal diseases cannot be based on generic measures of health status. (*J Rheumatol 1995;***22:1203-7**)

Key Indexing Terms:
GENERIC HEALTH STATUS INSTRUMENTS MUSCULOSKELETAL DISEASE VALIDITY

Conference participants split into groups to discuss the sensibility of a limited number of health status instruments. The groups were formed on the basis of interest in 6 rheumatic diseases: low back pain, ankylosing spondylitis (AS), osteoarthritis (OA), lupus, rheumatoid arthritis (RA) (2 groups), and osteoporosis. The constitution of the groups differed from that of Part 1 (toxicity), but was the same as that for Part 3 (economics).

In each subgroup, a sample trial design was provided in which the instruments to be discussed were candidate endpoints. To assess sensibility, the groups used a revised version of the sensibility questionnaire devised by Paul Peloso for toxicity indices[1]. Where possible, both before and after discussion, the group was asked to determine whether inclusion of the generic instrument in the trial would be useful (1) to put improvements seen in specific endpoints into perspective in the context of overall health; (2) to provide an assessment that incorporates both benefit and side effects, inconvenience; (3) to compare the relative benefit across conditions; (4) for Cochrane meta-analyses; or (5) when carrying out a cost-effectiveness study as part of the trial.

Conference materials provided to participants included full details of the instruments. After discussion in small groups, participants reconvened to report their findings in plenary session.

Low Back Pain *(Rapporteur: Matthew Liang)*

The trial scenario suggested to the low back pain group was a 6-month randomized controlled trial of nurses in a large chronic care hospital with low back pain of one or more weeks duration. The intervention was a self-exercise program booklet compared to intensive physiotherapy. We changed the scenario slightly so that we would not haggle over study design issues. We had time to rate only one generic instrument, the Short Form-36 (SF-36); this is a 13-component health status index. We lacked time to cover another generic measure, the European Quality of Life

N. Bellamy, MD, PhD, Professor, Department of Medicine, University of Western Ontario, London, Canada; M. Boers, MD, PhD, MSc, Associate Professor, Department of Internal Medicine/Rheumatology, University of Maastricht, Maastricht, The Netherlands; D. Felson, MD, MPH, Professor, Arthritis Center, Boston University, Boston, USA; J. Fries, MD, Department of Medicine, Stanford University School of Medicine, Stanford, USA: D. Furst, MD, PhD, Director of Arthritis Clinical Research, Virginia Mason Center, Seattle, USA; D. Henry, MB, ChB, FRCP(Edin), Centre for Clinical Epidemiology and Biostatistics, University of Newcastle, Newcastle, Australia; M. Liang, MD, MPH, Departments of Medicine and Rheumatology and Immunology, Harvard Medical School, Robert B. Brigham Multipurpose Arthritis and Musculoskeletal Diseases Center, Brigham and Women's Hospital, Boston, USA; D. Lovell, MD, Division of Rheumatology Juvenile Arthritis, Children's Hospital Medical Center, Cincinnati, USA; L. March, MD, Associate Professor, University of Sydney, Sydney, Australia; V. Strand, MD, Clinical Associate Professor, Division of Immunology, Stanford University, Stanford, USA: S. van der Linden, MD, Professor, Department of Medicine, Division of Rheumatology, University of Maastricht, Maastricht, The Netherlands.

Address reprint requests to Dr. M. Boers, Department of Internal Medicine/Rheumatology, University Hospital, PO Box 5800, Maastricht 6202 AZ, The Netherlands.

Measure (EUROQOL), or the Owestry low back pain score. Currently, some users look at the SF-36 as the *de facto* standard. We considered whether it would be the appropriate instrument for this mythical trial. We noted several problems. Even though the instrument has fairly good instruction, we thought myopic people would not be able to deal with it; it is unclear whether it is self-administered or interviewer-administered. As an aside, people with cognitive deficiencies will have problems answering any of these questionnaires.

We thought that many areas important for people with low back pain were not covered adequately or with enough response categories, including vital areas like sleep, sexual function, and ability to stand for any period of time. The pain category in SF-36 is too general and not specific to low back, with too few response categories. Another point is that the SF-36, the acute version, tries to assess symptoms over the past week. This creates problems for people with variable symptoms over a day or over a week: should they take the maximum pain that they had during the week or the average? Without specification, this complicates interpretation of the results, and most likely reduces reproducibility and reliability.

Finally, one of the important advantages of generic instruments could be that they can summarize total benefit minus the negatives. However, if you look at the famous Ds in terms of side effects or inconveniences, the SF-36 gives very skimpy if any coverage of those areas. In terms of a generic instrument allowing clinicians to compare the management of low back pain against other requirements, i.e., other rheumatic diseases or other medical conditions, the lack of coverage in those areas would create real problems.

However, we concluded no existing or theoretical instrument would be more satisfactory.

Ankylosing Spondylitis *(Rapporteur: Sjef van der Linden)*
Our task was to rate the sensibility of one disease specific instrument, the functional index by Dougados, and 2 generic instruments, the Canadian health Utility Index and the Nottingham Health Profile. The instruments were suggested as endpoints in a trial of group physiotherapy versus unsupervised exercises at home. All participants expressed satisfaction with the Dougados functional index. We went on to discuss the Canadian Health Utility Index, a 15 component health index that rates health utility on a 0-1 scale. We thought this instrument inappropriate for rheumatic diseases for the following reasons. It is not comprehensive enough, i.e., it does not capture most rheumatic problems, and it does not address expected side effects. We missed items on problems with sleep, stiffness, and also recreational activities. We considered whether other generic instruments such as Sickness Impact Profile, SF-36, looked more comprehensive but we did not come up with a definite answer. Also, we are worried about the responsiveness of the Canadian Health Utility Index.

The Nottingham Health Profile is a 38-item questionnaire in 6 dimensions. Generally, we thought it would be much better than the Health Utility Index, both in comprehensiveness and in responsiveness. However, this profile also does not capture adverse effects that might be expected, and we missed items concerning employment, inability to work, recreation, and side effects. Advantages of both indices could be that they would allow comparison with other diseases, including rheumatic diseases, but currently we question their validity in AS, and we think that neither can replace disease specific instruments.

Osteoarthritis *(Rapporteurs: Lyn March and Nicholas Bellamy)*
The OA group considered a hypothetical clinical trial and whether either of 2 generic health status instruments, the SF-36 and the Health Utility Index, would be useful in outcome measurement to supplement the following disease specific measures: pain and the Western Ontario and McMaster University Osteoarthritis Index (WOMAC), an instrument that assesses clinically important changes in pain, stiffness, and physical function. The hypothetical trial was of 6 months' duration in symptomatic patients with an osteoarthritic knee, and compared the effects of hyaluronic acid injections versus naproxen tablets.

Very few participants in our group had familiarity or any personal experience with the use of either the SF-36 or the Health Utility Index. No data on the clinimetric properties of these indices were available. Nevertheless, 90% of the group indicated it would be useful to include generic measures in the trial to improve the perspective of overall health. The majority also felt that generic instruments would be useful in comparing relative benefit across different disciplines.

Gillian Hawker had reported in plenary session on the relative sensitivities of the SF-36 and the WOMAC in a joint replacement study[2]. The SF-36 was a more sensitive generic measure, as might be expected, but the WOMAC was the superior disease specific measure. Although the statistical efficiency of the WOMAC has been well described in pharmacologic studies, it is not known whether the SF-36 or the Health Utility Index could detect changes in generic health status in such studies. Before the routine use of generic instruments can be recommended, further evaluation of their performance in this clinical setting is required. The group expressed enthusiasm for such evaluations.

Some concern was expressed whether the generic instruments could be completed by the elderly or infirm, and whether revalidation was required in populations with OA. Even if generic instruments are valid, reliable, and responsive, their use may not be necessary for all future studies. Use should be based on specific research questions and the

anticipated dimensionality of any response to treatment. Finally, it will be important to evaluate how to use generic instruments as outcome criteria for adjudicating the success or failure of treatment in individuals (as opposed to patient groups).

Lupus *(Rapporteur: Vibeke Strand)*
Our group studied measures as endpoints in a 12 month trial of patients with lupus with active glomerulonephritis, comparing intravenous cyclophosphamide plus a new biologic agent to intravenous cyclophosphamide only. The hope would be that at the end of 12 months, such patients would have stable or improved renal function and perhaps less toxicity, both in the short and the long term. We allowed ourselves multiple disease specific measures of activity, such as the Systemic Lupus Erythematous Disease Activity Index (SLEDAI) and other indices, besides renal function and global assessment by the patient and the physician. We agreed that we could also use either the Health Assessment Questionnaire (HAQ) or the Arthritis Impact Measurement Scales (AIMS) to get at some of the disease specific physical functions; we talked about using a Fatigue Severity Scale that has been validated in patients with lupus or multiple sclerosis specifically for fatigue and has shown to be sensitive. We then assessed the sensibility of a variety of generic assessments; we agreed that we might be able to use one — the SF-36 — although we do not know whether it would be responsive.

The Quality of Well-Being Scale is an instrument that classifies the patient's function level and then attaches a predetermined value or utility to this level. We thought that this scale was limited: it did not include cognitive or emotional scales, a lot of disability was implied, and there was a very significant ceiling effect. Several people felt that less than one percent of their patients with lupus or even rheumatic disease would actually be captured in this particular instrument. In contrast, we felt that the patients with lupus, in this trial, although relatively healthy, would show a score on Nottingham Health Profile. Again, there was doubt about the sensitivity to change. We looked at the Health Utility Index and felt that it had little broad applicability for lupus with the exception of Question 12 ("Which one of the following best describes your usual ability to think and solve day to day problems?"; answers: "Able to think . . .," "a little difficulty . . .," "some difficulty . . .," "great difficulty . . .," "unable to think . . ."). We thought this was a very nice cognitive question and captured a very important issue in patients with lupus. Finally, we looked at the SF-36 briefly and it seemed to have broader domains, it certainly talked about pain, but the concern was that it was described as bodily pain as opposed to the effects of, say, migraines. In general, it seemed to be the closest to capture the emotional and energy and fatigue issues, although it was a little bit light on the cognitive aspects. In general, most of us felt that, if we had to pick one, we would pick the SF-36 as a generic instrument.

Rheumatoid Arthritis *(Rapporteurs: Jim Fries and Dan Lovell)*
The scenario we had was an RA randomized clinical trial of patients caught early in the disease course, less than 2 years of disease duration, and they were all characterized as having severe disease. The treatment comparisons were steroids, sulfasalazine, and methotrexate in one group, versus sulfasalazine alone in the other group. The endpoints were the OMERACT 7; we included the Health Assessment Questionnaire as our functional assessment tool in our OMERACT 7, specifically the short, 2 page form, and we included in that the visual analog scale for pain and the visual analog scale for global.

The main question for our group was, "Would our randomized clinical trial design be improved by adding all or part of the EUROQOL and/or the short Sickness Impact Profile or SIP?" The EUROQOL is a combination of a patient global assessment of utility (rating scale) and 5 simple questions probing difficulties in 5 dimensions of health. No one in the group had used EUROQOL; we felt that the only unique aspect of this scale compared to a health assessment questionnaire short form was that the EUROQOL had an anxiety/depression dimension; we felt the global scale on the EUROQOL is quite important, however, the HAQ has one that is quite similar. The suggestion was made that we should standardize the wording of these global visual analog scales across different instruments, so it would be the same for the AIMS, and the HAQ, and the EUROQOL, in order to use that as a standard to compare the measurement characteristics of these various tools. In our pretest analysis of the EUROQOL, more than half of our group did not have a positive response to the EUROQOL questions; most of the time, at least 75% of the group felt the EUROQOL was not useful. Eight of the 18 felt they would be able to use it in early stage clinical use, and 10 of the 18 said that they would not.

When we did the main sensibility questionnaire for the EUROQOL, we had no information about reproducibility. Question 3 asks, "Does it assess physical, emotional and social functions?" - most people felt it did not. We felt that the items on the questionnaire did make sense. Important aspects of physical, emotional, and social functions would be captured by this questionnaire, although certain things about social aspects, drug side effects, costs, and mortality would not be quite caught on the EUROQOL. "Is there another generic instrument that does capture these things?" We believe there is. "Were the same items captured on the HAQ?" We say they were not, because the HAQ at least on the short form, really does not get at the drug side effects and some of the psychological concerns. We were evenly split about whether the EUROQOL would be responsive.

Some of the scores were available and it was easily calculated. We did not feel the EUROQOL would allow us to analyze the advantages and disadvantages of a particular intervention.

Our other questionnaire was the Sickness Impact Profile (SIP). There were 2 people in our group who had actually used the 130 questions, but it has now been shortened to 68 questions (in 6 dimensions of health) and none of us had used this short form. We felt that both questionnaires were quite long; the experience of clinicians was that patients did not like to fill it out. They showed a very large treatment effect in the hip replacement trial and a moderate effect in the erythropoietin trial in renal failure. SIP does a good job putting an overall aspect of quality of life in perspective. There is nothing on this scale about pain. It does augment the short form of the HAQ because it has dimensions dealing with emotional, social and psychological aspects. Because each of the 68 items is individually weighted, it is difficult or almost impossible to score by hand. But no one in our group had experience in a clinical setting with the 68 item questionnaire.

Rheumatoid Arthritis *(Rapporteurs: David Felson and Dan Furst)*

We evaluated the usefulness of the SF-36 in assessing patients in RA clinical trials. The SF-36 is a widely used health status instrument, felt to be valuable in that any mean improvements could be placed in perspective relative to other diseases. If widely used in the rheumatic diseases, the relative benefits of treatments in RA, OA, or lupus could be compared, as well as the relative benefits of treatments for arthritis versus other diseases such as stroke or mental illness. Obviously, this would provide advantages for the Cochrane collaborative metaanalysis project in providing a single standard of evaluation for all rheumatic diseases, and perhaps even all diseases.

Unfortunately, the SF-36 has many deficiencies that limit its use for evaluating RA. First, the brief list of physical disabilities included in the SF-36 raises serious questions about its content validity for RA. There was special concern that no upper extremity functions were assessed. Further, a variety of functional status and quality of life issues important to patients with RA are not included, such as questions on sexuality, fatigue, and sleep. Also, an evaluation of coping and self-efficacy are absent, although it is realized that other quality of life or functional status instruments also lack these items. There is little information that could be used to evaluate adverse events. Although comorbidity is not directly evaluated by the SF-36, one might imagine that overall health status as measured by the SF-36 would be affected by both RA and concomitant illness. Last, and critical for cost-effectiveness analyses, there is no utility measurement, although investigators are working to convert the scales produced by the SF-36 into a single measure of utility.

Osteoporosis *(Rapporteur: David Henry)*

We looked at the EUROQOL and the short SIP in the context of a patient that I will describe. I think the short SIP requires perseverance to complete. We had some difficulties with both these instruments, echoing the comments of the previous group, because no one in the group has actually used the EUROQOL or the short version of the SIP. The other points were that in the conference material, bits of EUROQOL were missing, and the description of the SIP was actually a study in which a fraction analysis was supported and the instrument itself was not included.

An important point is that in our view, the scenario for our discussion group was different from that of the other groups; this is probably relevant to the choice of instruments. The suggested scenario was a trial in women with post-menopausal osteoporosis, vertebral collapse, and bone mineral density outside 2 standard deviations below the mean. The issue that was not addressed in the written scenario was whether the patient was symptomatic. Often, such patients would not be symptomatic, so that the treatment would really be to prevent a future event, a vertebral fracture that may itself be symptomatic or not. In this setting, analogous to treating hypertension or lowering cholesterol, it may be that a disease specific instrument is actually not useful. Moreover, the main result of treatment from day to day may in fact be adverse effects on quality of life! If you were using estrogens, for example, most patients would be aware only of the side effects of the estrogens, rather than any benefits of the treatment.

We decided that in symptomatic patients, a disease specific instrument was going to have some advantages because it captures things like pain and fear of falling, which are going to be important to somebody in this setting. In asymptomatic patients, we would be more interested in an instrument that would capture the side effects of treatment. In our view, both generic instruments probably would be quite useful in the latter setting. We considered that somebody starting estrogen treatment, particularly today, would have been exposed to a lot of concerns about side effects of estrogens; thus, the emphasis of both versions of the SIP on emotional and social functioning would be quite appropriate.

The big disadvantage of SIP versus EUROQOL and some other generic instruments was that it does not look at pain at all. Pain would certainly be an important endpoint in an osteoporosis trial, even if the aim was prevention of further fractures in asymptomatic patients. Nevertheless, both instruments could perform quite well in terms of their comprehensiveness and their ability to capture different aspects of the person's illness, in particular their ability to capture concerns and anxieties about the side effects of treatment.

DISCUSSION

The main points from these discussions:

1. Conference participants, self-selected for their interest in this field, have little personal experience with generic health status instruments. This is worrying in view of the growing popularity of such instruments with policymakers.

2. There is little data on the use of these instruments in musculoskeletal diseases.

3. The main concerns voiced over these instruments is their lack of comprehensiveness and their supposed lack of responsiveness. The lack of comprehensiveness is evident both for disease effects (e.g., pain, stiffness, upper extremity function, disability, psychosocial) and for adverse effects due to treatment. The lack of responsiveness is presumed to follow from the lack of comprehensiveness and the small number of response categories in the dimensions that are represented in the instruments.

4. Participants expressed willingness to engage in research with these instruments.

In conclusion, rheumatologists should actively pursue research in this area, ideally by including one or more of these instruments alongside clinical trials and observational studies. Only in this way can experience be gained to improve these measures. Until such experience is available, valid conclusions and health policy regarding musculoskeletal diseases cannot be based on generic measures of health status.

ACKNOWLEDGMENT

We thank Diane Gagnon for preparing the transcripts of the recording.

REFERENCES

1. Peloso PM, Wright JG, Bombardier C: A critical appraisal of toxicity indexes in rheumatology. *J Rheumatol 1995;*22:989-94.
2. Hawker G, Melfi C, Paul J, Green R, Bombardier C: Comparison of a generic (SF-36) and disease specific (WOMAC) instrument in the measurement of outcomes after knee replacement surgery. *J Rheumatol 1995;*22:1193-6.