Pushing the Limits of Patient-Oriented Outcome Measurements in the Search for Disease Modifying Treatments for Osteoarthritis

MATTHEW H. LIANG

Let us consider that we have a new chondroprotective agent that, in elegant clinical trials, has been shown to be efficacious. How do we identify the people to whom we will give this magic compound? We could attempt a primary prevention strategy in which we try to identify subjects who have no symptoms but are destined to develop osteoarthritis (OA), or those who may have early OA seen on radiographic or magnetic resonance imaging. On the other hand, we might treat individuals with very early, minimally symptomatic OA in a strategy aimed at secondary prevention. Based on the little we know about the natural trajectory of OA of the knee, it may take some 20 to 40 years of treatment with a drug that holds promise for preventing cartilage destruction. During this period, the recipient might lose somewhat less than the 0.1–0.2 mm per year of joint space width (JSW) that we expect to be lost in the untreated individual — and we would have no assurance that this will be associated with clinical benefit. This presents us with a major challenge. Further, we will need to understand how to get people who have no symptoms to take a drug to prevent or delay the occurrence of OA over 20 to 40 years.

The data and experience we have gained in treating hypertension to prevent stroke and heart disease and in treating low bone density to prevent osteoporotic fractures provide some lessons in what will be needed for a public health strategy in OA. They show us the following:

1. Clinical trials are needed to demonstrate the efficacy and safety of new agents intended to prevent clinically meaningful endpoints;
2. Longterm compliance with a prescribed medication will likely exceed the level of compliance with a prescribed change in behavior;
3. Support and approval from the patient's family and social network will be highly important;
4. The identification of potential barriers to compliance and the development of realistic strategies to overcome them (including financial barriers) will be necessary; and
5. It will be important to ensure that people see how well they are doing, as has been achieved with home blood pressure monitoring or home glucose monitoring. The availability of such technology to measure one's status has made a big difference in the management of hypertension and osteoporosis. It will also be needed for use of a chondroprotective drug on a population scale.

Elsewhere in these proceedings, Paul Dieppe presented the latest version of the World Health Organization (WHO) taxonomy[1]. Its progenitor was created in 1979[2] and was a landmark taxonomy that distinguished impairment, disability, and handicap to describe the experience of illness from the patient's viewpoint (Table 1). It set in motion the patient-oriented outcomes movement.

These definitions can be illustrated with knee OA. Impairment would correspond to objective evidence of joint space narrowing (JSN) on a radiograph. Disability would be the impact of knee OA on a person's ability to function in daily life — in recreational activities or on the job. Handicap would be the impact of knee OA on that person's ability to carry out their social role, be it butcher, baker, or candlestick maker. In the clinical management of OA, radiographs and other measures of impairment do not help the physician very much and, certainly, they do not capture the human condition — i.e., x-rays don't weep.

The relationships between knee OA, functional activities, loading of the knee, and JSN are complicated. The severity of knee pain in patients with knee OA is inversely proportional to the difficulty reported by the patient with activities that require weight-bearing. When someone's knee OA gets worse, they do less.

This observation is supported by a study of how pain may affect JSN. Mazzuca, *et al*[3] performed careful measurements of the medial compartment interbone distance in standing anteroposterior (AP) radiographs of patients with knee OA. When the patient experienced a flare of joint pain, a significant decrease in JSW occurred, in comparison with measurements of images of the same knee obtained when the patient was less symptomatic (and, therefore, better able to stand with full knee extension). However, when special attention was paid to positioning of the knee in a concurrently obtained semiflexed AP view, in which standardized radioanatomic positioning of the joint was achieved fluoroscopically, JSW was unaffected by the severity of joint pain (Table 2).

The Western Ontario and McMaster University

From the Department of Medicine, Rheumatology, and Immunology, Brigham and Women's Hospital, Boston, Massachusetts, USA.

M.H. Liang, MD, MPH.

Address reprint requests to Dr. M.H. Liang, Department of Medicine, Rheumatology and Immunology, PBB-B3, Brigham and Women's Hospital, 75 Francis Street, Boston, MA 02115.
E-mail: mliang@partners.org

Table 1. World Health Organization taxonomy. From *International Classification of Impairments, Disabilities and Handicaps*, Geneva, World Health Organization, 1980.

Impairment	Loss or abnormality of psychological, physiological, or anatomical structure or function
Disability	Restriction or lack of ability to perform normal activity
Handicap	Disadvantage for the individual, limiting fulfillment of normal role in society

Table 2. Mean ± SEM change in WOMAC pain score and joint space width (JSW) in extended and semiflexed AP views of the knee, after adjustment for the within-subject correlation between knees. With permission from Mazzuca, *et al*. Arthritis Rheum 2002; 46:1223-7.

	Flaring Knees, n = 12	Nonflaring Knees, n = 15	p, Flaring vs Nonflaring*
WOMAC pain score	–8.7 ± 1.1	–3.5 ± 0.9	0.0004
JSW in extended view, mm	0.20 ± 0.96	–0.04 ± 0.04	0.0053
JSW in semiflexed view, mm	0.08 ± 0.05	0.02 ± 0.05	0.376

* $p \leq 0.005$ vs 0, by t test.

Osteoarthritis Index (WOMAC)[4] is, at least based upon its usage, the functional measure of choice in the OA field. Nick Bellamy, its inventor, has written about its 20-year history, modifications, and enhancements[5]. At its core, the WOMAC measures 3 dimensions: pain, stiffness, and function, scored categorically or on a visual analog scale. It has been translated into numerous languages and is used in almost every clinical trial of OA today. In many ways, it shows the influence of the WHO classification over 25 years. What was once a theoretical construct is now mainstream. There are, however, limitations to the WOMAC and other questionnaires and if we are to prevent or cure OA or know when we have done so, we will need to find solutions to these limitations.

Some caveats to the use of these measurements of function in patients with OA come from our experiences in using them for research.

We studied nearly 300 community-dwelling elderly subjects in Boston and Vermont, where I made home visits with either a physical or occupational therapist to establish a relationship between self-reported function and objective performance measures[6]. The subjects rated themselves with the Health Assessment Questionnaire (HAQ), in which a score of zero indicates a lot of difficulty and 3 indicates no difficulty with a specific task. Health status instruments such as the HAQ may have floor and ceiling effects and are insensitive to change in people who score near the extreme values on the scale. The floor effect refers to the inability to measure improvement in subjects who already have a minimal disability score; the ceiling effect refers to the inability to measure deterioration in individuals who start off with the worst possible score. Floor and ceiling effects are not unique to the HAQ. They have been shown to be a problem with every measure of function when this has been studied.

Among the elderly subjects in our study, there was a group who reported they had no problem with turning doorknobs (a question in the HAQ), but who could apply only minimal force when they were tested by the therapist with a hand-held dynamometer. Thus, a functional grip needed to open a tight door requires a certain amount of strength, but a number of the subjects who could not generate that much force claimed to have no difficulty with the task — i.e., the questionnaire was not telling us what we wanted to know.

A subject whom we encountered in another study illustrates yet another problem with functional assessment questionnaires: He was a homebound man about whom we were consulted for the question of knee and hip OA. When he was in his 70s, he was admitted to a local hospital with a severe urinary tract infection. He weathered that illness and returned to his apartment in a weakened state. Further, he was convinced that he was no longer able to walk as the result of something that had happened while in the hospital. Some 17 years later, when I visited him, his life was confined to a couch; he had fulfilled his own prophecy that he could not walk. He did not have OA, but had developed severe hip and knee flexion contractures from sitting for prolonged periods of time. However, he had a wonderful support system. Neighborhood teenagers purchased his groceries. Charitable organizations came to sing Christmas carols to him. He sponge-bathed, toileted, and did everything from his couch. The rest of his apartment was unused. However, on a functional assessment instrument, he expressed no difficulty. Again, the questionnaire did not give us an accurate picture of the patient.

All patient-oriented outcome questionnaires — whether generic or OA-specific — share the following problem: What patients report as functional ability lies in the interstices of the 3 circles in Figure 1, and their reported function is the result of their physical ability or "capacity" to perform an activity, the environmental factors or the requirement to do a specific task ("need"), and what I have labeled "will" in Figure 1, which includes expectations and motivation.

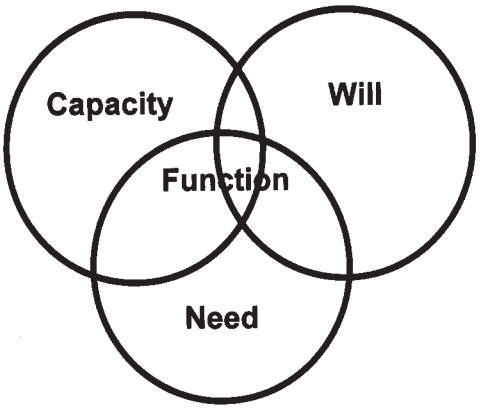

Figure 1. Determinants of disability. The function that patients report is a result of their physical, mental, and emotional capacity; will; and the physical need they have to perform the task.

In orthodox medicine we focus heavily on capacity, which corresponds to the WHO impairment level describing the impact of disease. In knee OA, we might strengthen the quadriceps or treat the occasional patient who has an inflammatory component with medications, lavage the joint, etc. If the patient's functional impairment is out of proportion to the severity of her "objective" disease, we blame her or we call in mental health services to improve her motivation or manage her depression. If things are bad and there is no medical or surgical solution, we essentially give permission to do less, revise their expectations downward, or accommodate the disability by the use of appliances or modification of the environment. We say, "Well, you don't have to do that anymore." The patient's report of function is dependent on these determinants — capacity, will, and need — and patient-oriented questionnaires cannot capture this complexity.

The major limitation of our current measures of function in OA is their inability to detect change. This has major significance for how we will evaluate chondroprotective agents and the sample size needed for these studies. The more sensitive a measure, the fewer subjects are needed to demonstrate whether a treatment is efficacious. As one might predict with respect to clinical outcomes in knee OA, joint arthroplasty has a huge effect on a patient's symptoms, nonsteroidal antiinflammatory drugs and medical management have less, and rehabilitation programs still less.

Angst, *et al*[7] provided quantitative data on the number of subjects required to demonstrate the effect of a rehabilitation intervention for a patient with knee OA. The intervention would not be common in America, but is available in Germany and might involve spa therapy, exercise, and hospitalization or admission to a controlled environment. Table 3 shows baseline and 3-month followup scores for the WOMAC and Medical Outcome Study Short Form-36 in that study[6]. The effect size (ES) and standardized response mean (SRM) statistics indicate the sensitivity of the measure, i.e., its ability to pick up any change. Responsiveness — the ability to detect a clinically meaningful change — was studied by comparing the WOMAC and SF-36 to a transition question that the investigators used to ask the patient whether she had changed over the 3-month period and, if so, how much change had occurred and whether it was meaningful. The higher the values for ES and SRM, the more sensitive the measure (Figure 2). The investigators' conclusion was that if the WOMAC is used in a randomized clinical trial, its sensitivity is sufficient to allow a sample size of fewer than 300 subjects per treatment arm. This is a sample size that can realistically be achieved and the change that was measured is clinically meaningful. There are few data on effect size with respect to chondroprotective drugs, and this study can give us an idea of its order of magnitude.

Figure 3 depicts what I believe is the relationship between impairment and disability in musculoskeletal disease and some other diseases as well.

If we measure impairment in the case of knee OA, JSW scores may range from "no impairment" to "endstage

Table 3. Patients with hip or knee OA, before and after inpatient rehabilitation. With permission from Angst, *et al*. Arthritis Care Res 2001; 45:384-91.

Measurement	Baseline, Mean ± SD	3–mo Followup, Mean ± SD	Change Between Baseline and 3-mo Followup		
			Mean ± SD	ES	SRM
WOMAC (n = 122)					
Pain	4.83 ± 2.25	4.18 ± 2.37	0.66 ± 1.96	0.29	0.34
Stiffness	4.61 ± 2.67	4.58 ± 2.40	–0.03 ± 2.55	0.01	0.01
Function	4.81 ± 2.18	4.33 ± 2.32	–0.47 ± 1.73	0.22	0.27
Global	4.80 ± 2.09	4.32 ± 2.26	–0.47 ± 1.72	0.23	0.27
SF-36 (n = 116)					
Bodily pain	27.1 ± 16.5	37.5 ± 20.3	10.5 ± 23.0	0.63	0.45
Physical function	37.5 ± 20.6	37.9 ± 22.1	0.4 ± 20.3	0.02	0.02
Physical component summary	28.6 ± 7.7	30.9 ± 9.1	2.3 ± 8.0	0.30	0.29

ES: effect size; SRM: standardized response mean; WOMAC: Western Ontario McMaster University Osteoarthritis Index; SF-36: Medical Outcomes Study 36-Item Short Form; WOMAC scale: 0 = no symptoms, 10 = extreme symptoms; SF-36 scale: 0 = extreme symptoms, 100 = no symptoms; ES = mean (effect) –SD (baseline); SRM = mean (effect) –SD (effect). Improvement occurred if WOMAC effect < 0, SF-36 effect > 0.

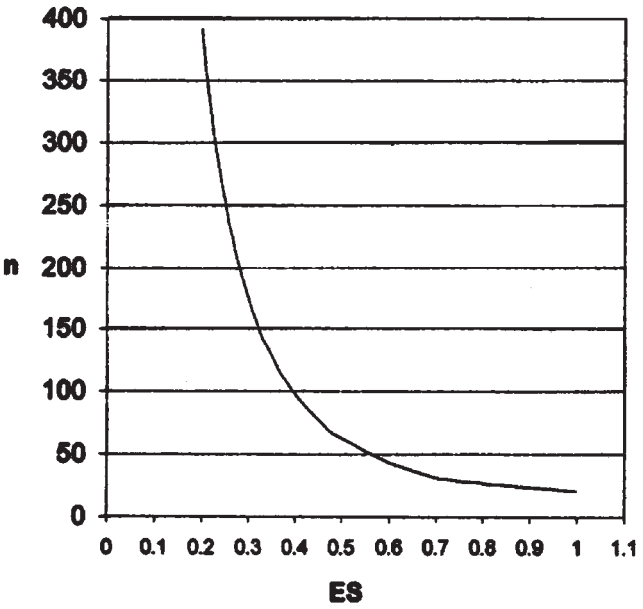

Figure 2. Relationship between sample size (n) and effect size (ES, i.e., sensitivity to change). The greater the ES, the smaller the n needed per treatment arm in a randomized controlled trial. With permission from Angst, et al. Arthritis Care Res 2001;45:384-91.

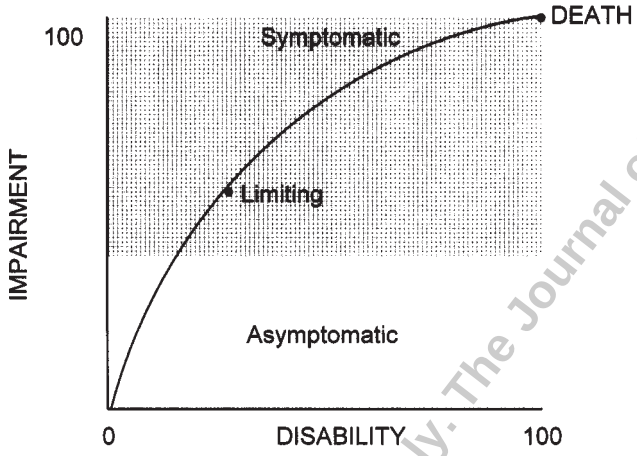

Figure 3. Relationship between impairment and disability.

disease." If we examine disability, it may range from no disability to 100% disability (i.e., being bedridden). Early on, when the subject is asymptomatic, it is possible to measure some level of impairment. There is a point or a threshold on the subject's trajectory of functionality, usually related to the patient's expectations and needs (Figure 3), where she becomes "symptomatic." Early on, in the trajectory of disability, the most sensitive measures are those that detect impairment; later in the course of the disease, patient-oriented measures are usually more sensitive to change.

JSW is not what people care about; they care about how they feel and what they can do. We need patient-oriented measures. However, the patient-oriented measures available to us are most relevant and useful in advanced knee OA. In fact, almost everything we know about knee OA is derived from patients with advanced disease. To evaluate the possibility of achieving major health effects with an intervention that prevents or delays the onset of knee OA, our outcome measures are insensitive, unresponsive and, basically, of little use.

How might we evaluate a putative chondroprotective agent or improve its dissemination? First, research aimed at enhancing adherence to the dosing regimen and retention of subjects in clinical trials, and at improving the reliability and sensitivity of structural outcome measures, such as magnetic resonance imaging and radiography, should be a priority. Second, a standardized technique for provoking symptoms could be devised for knee OA, analogous to the exercise treadmill test to detect subclinical ischemic heart disease or to evaluate symptoms of lumbar spinal stenosis. Measures are more sensitive when subjects are pushed to their limits. The sensitivity of questionnaires can be improved by decreasing noise or increasing the signal. Strategies for accomplishing this are illustrated in Table 4.

Third, computer adaptive testing could be employed. Formerly, the US Educational Testing Service's Graduate Record Examination was taken over several hours. Three or four months later, results would be sent. Today, students sit down before a computer, answer questions for less than 30 minutes, and promptly receive their score. The technique builds on a powerful methodology, item response theory, based on an enormous data bank of questions and answers. The response to a question of a certain difficulty triggers a followup question. For example, a candidate might know the number of points that Indiana-born basketball great Larry Byrd scored in a lifetime, but might not know the most points he ever scored in a game. In this manner, the final score to the test can be triangulated by individualizing the questions, using a vast bank of questions and the correct or incorrect responses of all people who have responded. The bank, of course, changes with the information collected. The SF-36 has moved to computerized testing. It may be possible to generate an unlimited number of questions, obtain experience with those questions by testing people

Table 4. Strategies to improve sensitivity of questionnaires.

Decrease noise
Have respondent see previous response
Eliminate unreliable questions
Reduce missing data and response bias
Pool outcome measures to create an index
Increase signal
Increase number of questions relevant to population studied
Increase number of response categories
Use goal attainment scaling techniques
Use item response theory method

with early OA and, thereby, create a questionnaire of reasonable length that can identify presymptomatic individuals or those with early symptomatic OA.

In summary, preventing or slowing the actual progression of OA, once a remote dream, is now considered attainable. If we are to ever know whether we have achieved this goal, we will need new knowledge of the determinants of why and how people adapt preventive behaviors or take medications for asymptomatic or barely symptomatic musculoskeletal conditions that take decades to become manifest. We will need cost-effective, sensitive measures for earliest and early OA, and we will need to develop measures of impairment and symptoms that are sensitive to the entire range of progressive OA.

REFERENCES

1. World Health Organization. International classification of functioning, disability and health [Internet]. 2001 [Accessed December 18, 2003.] Available from: www.who.int/classification/icf
2. World Health Organization. International classification of impairments, disabilities and handicaps. Geneva: World Health Organization; 1980.
3. Mazzuca SA, Brandt KD, Lane KA, Katz BP. Knee pain reduces joint space width in conventional standing anteroposterior radiographs of osteoarthritic knees. Arthritis Rheum 2002;46:1223-7.
4. Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt L. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes following total hip or knee arthroplasty in osteoarthritis. J Orthop Rheumatol 1988;1:95-108.
5. Bellamy N. WOMAC: a 20-year experiential review of a patient-centered self-reported health status questionnaire. J Rheumatol 2002;29:2473-6.
6. Liang MH. The measurement paradox of disability and its implication for gerontology. In: Bergnet M, Ermini M, Stahelin HB, editors. Challenges in aging. New York: Academic Press; 1990:223-31.
7. Angst F, Aeschlimann A, Stucki G. Smallest detectable and minimal clinically important differences of rehabilitation intervention with their implications for required sample sizes using WOMAC and SF-36 quality of life measurement instruments in patients with osteoarthritis of the lower extremities. Arthritis Care Res 2001;45:384-91.