Rheumatology 2000;39:783–790

Use of health status measures in patients with low back pain in clinical settings. Comparison of specific, generic and preference-based instruments

M. E. Suarez-Almazor, C. Kendall[1], J. A. Johnson[2], K. Skeith[1] and D. Vincent[1]

Department of Medicine, Baylor College of Medicine, Houston, TX, USA, [1]University of Alberta, Edmonton, Alberta and [2]Faculty of Pharmacy and Pharmaceutical Sciences, University of Alberta, Institute of Pharmaco-Economics, Edmonton, Alberta, Canada

Abstract

Objective. To evaluate the discriminative performance over time of specific, generic and preference-based instruments in patients with low back pain (LBP) in clinical settings.

Methods. Forty-six consecutive patients with LBP participated in the study. Self-response questionnaires were administered at baseline and 3 and 6 months, including the following instruments: Oswestry (specific for LBP), SF-36 (generic), EuroQol (EQ-5D) and Health Utilities Index (HUI) (preference-based). EQ-5D and HUI weights were derived from previously published evaluations in the general population. Patients were asked to compare their health status with their baseline health and were categorized on the basis of an ordinal scale as: (a) improved; (b) stable; or (c) worse. Changes in the instruments were evaluated by rescaling the instruments over the same scale interval and by estimating standardized effect sizes between two time points for the three categories of change.

Results. Thirty-seven patients (80%) completed both the baseline and the 3-month questionnaire and 34 the baseline and 6-month questionnaires (74%). Overall, at both time points, approximately half of the patients reported no changes in their health status. Correlations between instruments were generally low, suggesting that they measure different health domains. The scales which discriminated best between patients who improved and those who deteriorated at 3 months were the Oswestry, the HUI, the EQ-5D and the SF-36 bodily pain and emotional role subscales. The SF-36 subscales appeared to have a floor effect for those patients who had deteriorated.

Conclusions. Most SF-36 subscales did not adequately reflect changes in the health status of patients with LBP, mostly for those who reported deterioration. Preference-derived quality-of-life scores appeared to discriminate among patients who improved and those who deteriorated, although not as consistently as the disease-specific measure (Oswestry). Additional research is needed to evaluate the role of generic measures of quality of life in the assessment of patients with LBP before they can be widely implemented in clinical settings or outcomes research.

Low back pain (LBP) is a major cause of discomfort and disability in developed countries, and is estimated to be the most prevalent pain complaint [1–2]. Overall, about three-quarters of the general population have experienced LBP at some time. Despite its high prevalence, the precise cause of LBP remains unidentified in the vast majority of patients. Physical examination and laboratory and imaging procedures are often non-specific, do not correlate well with symptoms and are generally of little help in the diagnosis or follow-up of most patients [3, 4].

Subjective patient-based assessments are increasingly being performed to evaluate the outcome of LBP [5]. Several instruments have been developed to specifically assess these patients, such as the Oswestry and the Roland–Morris Low Back Pain Disability Questionnaires [6, 7]. Also, a variety of generic instruments are used currently to evaluate health-related quality of life [5]. One advantage of these generic tools is

Submitted 21 May 1999; revised version accepted 24 January 2000.

Correspondence to: M. E. Suarez-Almazor, Health Services Research, Baylor College of Medicine, Veteran Affairs Medical Center (152), 2002 Holcombe Blvd, Houston, TX 77030, USA.

that they allow comparison of outcomes across diseases and can therefore be useful in policy decisions.

Preference-based quality-of-life measures are increasingly being used in clinical trials and outcome studies. These tools evaluate the health status of patients; the evaluation or rating of specific health states is obtained from the patients themselves or through surveys of groups of individuals (general population, patients, health-care providers). Different techniques are used for evaluation, most commonly visual analogue scales (VAS), time trade-off or standard gamble. Most often, these ratings are anchored at 0 (death) and 1 (perfect health), all other possible health states being rated within this interval. These ratings can then be used as quality-of-life adjustment weights to calculate, for instance, quality-adjusted life years and similar measures, which can be used in economic evaluations based on cost–utility analysis. Preference-based generic quality of life instruments include the EuroQol (EQ-5D) [8], the Health Utilities Index (HUI) [9] and the Quality of Well Being scale [10].

To ensure adequate coverage of health-related quality-of-life domains, general recommendations suggest the inclusion of generic profile, and preference-based in addition to condition specific measures [11]. A concern with generic quality-of-life and preference-based instruments is that they may not adequately reflect changes in health status in populations with specific disorders because of their broad scope [11]. Consequently, the discriminative ability of the generic tools should be evaluated specifically for each disease. Most studies evaluate responsiveness in clinical trials where most patients are expected to improve, and most instruments may perform reasonably well. It is not clear whether these instruments are also useful for patients in follow-up and routine care, who generally show less change in their health status. Desired measurement features in this population include not only discrimination between improvement and deterioration, but also stability in patients showing no change. The objective of this study was to compare the performance of specific, generic and preference-based quality-of-life instruments in a cohort of patients with LBP receiving clinical care at specialty clinics.

Methods

Consecutive patients with LBP seen by one of two physicians (rheumatologist and chronic pain specialist) at outpatient clinics in Edmonton (Alberta, Canada) were included in the study. Eighty-five patients were approached originally, but only 46 (54%) agreed to participate. Patients completed several self-administered questionnaires at baseline and at 3, 6 and 12 months. Most often the baseline assessment was completed at the clinic, but occasionally patients were allowed to complete it at home because of its length and duration ($\sim$45 min); the 3-, 6- and 12-month questionnaires were mailed to patients. Three-month questionnaires were sent to all patients; 6- and 12-month questionnaires were sent only to those patients who had completed either the baseline or the 3-month questionnaire. For the 12-month questionnaire the number of responses was too low to allow meaningful comparisons, and this time point was not included in the analysis.

Instruments

Disease-specific. The Oswestry disability questionnaire was used as a LBP-specific functional assessment [6]. It has been shown to be a valid indicator of disability in patients with LBP. It is based on 10 sections with six levels each, assessing the limitations of various activities of daily living. The range of possible values is from 0 (best health state) to 100 (worst health state).

Generic instruments. Two instruments were used: the SF-36 and the EuroQol visual analogue scale (EQ-VAS). The SF-36 is a 36-item general health questionnaire. Eight dimensions are measured: general health perception, physical function, physical role, bodily pain, social functioning, mental health, emotional role and vitality. Two summary scores have also been developed: the Physical Component Summary Score (PCS) and the Mental Component Summary Score (MCS). The validity and reliability of the SF-36 has been tested extensively [12, 13]. The EQ-5D has two components. One is a multidimensional instrument used for preference-based scores, which is discussed below [8]; the other component is the EQ-VAS, ranging from 0 to 100, where 0 is the worst imaginable and 100 the best imaginable health. The patient rates his or her current health within this interval, which becomes the score. This component is not preference-based, but can be considered as a simple, generic estimate of overall health.

Preference-based instruments. Two tools were used, the EQ-5D and the self-administered Health Utilities Index Mark 2/3 15-Q (HUI) [8, 9]. The EQ-5D is a new generic preference-based tool for the measurement of health-related quality of life, which was developed in Europe [8]. It evaluates five attributes or domains: mobility, self-care, activity, pain, and depression or anxiety (these two functions are evaluated together). Each of these domains has three possible levels (no impairment, mild to moderate impairment, and severe impairment). We calculated EQ-5D preference scores for each patient using general population evaluation weights obtained through time trade-off techniques in the UK (York weights) [14]. Time trade-off methods ask respondents to select between a given amount of time in a health state and a different amount of time in another health state. If one of these states is perfect health, the subject typically chooses a shorter period of time in perfect health than a longer time in a poorer health state. The time intervals are modified until it becomes difficult for the subject to choose between the two alternatives. This time point is then translated into a preference-based score. The resulting index score is based on a scale from 0 (death) to 1.0 (full health). Negative scores represent values for states considered worse than dead [14].

The other preference-based instrument used in this

study, the HUI, was developed in Canada. We used the self-administered version, HUI Mark 2/3 15-Q. This is a 15-item multi-attribute instrument, which evaluates sensation (vision, hearing, speech), cognition, mobility, self-care, emotion and pain [9]. Each of the items has between four and six possible levels, ranging from normal to highly impaired. The HUI was scored using the Mark 2 algorithm (The algorithm can also incorporate fertility, which in this study was assumed to be normal). The Mark 2 utility function is based on weights derived from the Canadian population using standard gamble. With this technique, individuals have to choose between two alternatives, one being risky. The risk alternative has the probability of a more preferred health state, the complementary alternative being a less preferred state. The other alternative has no risk and specifies the certainty of a given health state, which typically ranks between the two presented in the risk alternative. The score is obtained from the probabilities of the risk alternative when the subject has no preference between this and the other alternative offering a certain health state. Because of the number of possible health states described by the HUI, which cannot reasonably be evaluated independently, the weighting process developed by the authors is based on a multiplicative multi-attribute utility function [9]. Overall HUI scores are also scaled so that 0 represents death and 1.0 full health.

The 3- and 6-month questionnaires included a question (ordinal change scale) asking the patient to compare their current health status with their baseline status, with the following choices: (a) much better; (b) somewhat better; (c) mostly the same; (d) somewhat worse; and (e) much worse. Patients were also asked to rate their degree of pain on a scale of 0 to 10, 0 representing no pain and 10 the worst imaginable pain.

Analysis

Pearson and Spearman correlation coefficients were used to evaluate the association between the various measures at baseline, to determine to what degree the various tools measured similar attributes. To evaluate discrimination over time, baseline scores were compared with the 3-month or 6-month responses. As already mentioned, too few patients answered the 12-month questionnaire to allow meaningful comparisons, and this endpoint was not included in our analysis. For some analyses, raw mean score differences were rescaled to 0–100 for all scores, where 0 was the worst possible health state for each instrument and 100 the best possible health state. This allowed comparison of the magnitude of change across scales. Standardized effect sizes were also calculated using the standard deviation (S.D.) of the baseline score: $(\bar{X}_i - \bar{X}_0)/S_0$. Effect sizes were calculated so that a positive value indicated an improvement over the time interval and a negative value indicated deterioration. Standardized mean differences were also calculated using the S.D. of the difference, but no major discrepancies were observed between the two methods, so only effect sizes are presented. Relative changes, such as percent changes, were not used for a number of reasons. First, the EQ-5D can have negative values. These values, if used in the denominator, may reverse the direction of the health change, although this problem can be reduced by rescaling. Secondly, many of the scores were zero or approached zero. Such scores in the denominator would lead to outrageously high values. Thirdly, quality-of-life scores are based on interval and not ratio scales, so relative changes may be meaningless.

Spearman correlation coefficients were used to estimate the correlations of change in the various scales with (a) the 5-point change scale and (b) the 0–10 ordinal pain scale. Pearson correlation coefficients were used to evaluate the association between the Oswestry functional scores and the other instruments. The Oswestry tool was used as a benchmark because of our prior hypothesis that LBP-specific scales would be more sensitive to change.

Statistical differences between baseline and 3- or 6-month scores were evaluated using paired *t*-tests. We evaluated the discriminative properties of each instrument according to clinical improvement. Because of the small number of patients choosing some responses in the 5-point ordinal change scale, patients were aggregated into three categories when evaluating mean differences in instruments according to improvement: (a) improved (much better, somewhat better); (b) same (mostly the same); and (c) worse (somewhat worse, much worse). Mean effect sizes and rescaled mean differences were estimated for each category, for each scale. Differences between categories were compared using one-way analysis of variance. For variables with asymmetrical distributions, non-parametric tests were also used.

Intraclass correlation coefficients (ICCs) were used to evaluate the stability of the scales for patients who did not improve. In this situation, the scores at both time points are expected to be highly correlated (>0.80). Fixed-effects ICCs were calculated from two-way repeated measures analysis of variance tables.

Results

Forty-six patients completed the baseline questionnaire and were originally included in the study: 30 (65%) were female, and mean age was 49.9 years (S.D. = 14.8). All patients had chronic LBP (>3 months duration) with a mean duration of 10 years (S.D. = 11.4). Mean and distribution characteristics of the baseline scores are shown in Table 1. The Oswestry, the ordinal pain scale, the HUI and some of the SF-36 subscales (social functioning, mental health, energy, pain and general health) had a bell shape. The SF-36 physical functioning and role–physical subscales were positively skewed, and the emotional role subscale was bimodal, with cases clustering at both ends of the scale. The EQ-5D scores were negatively skewed. Correlations among the baseline scores for the various instruments were generally low to moderate ($r < 0.60$), suggesting that these scales measure related but different health domains.

TABLE 1. Baseline scores and distribution characteristics of the various scales

	(Possible range)	Mean (S.D.)	Median	Minimum	Maximum
Oswestry	(0 to 100)	42.2 (15.7)	44.4	11	80
SF-36 general health	(0 to 100)	48.2 (20.2)	50.0	10	92
SF-36 physical functioning	(0 to 100)	37.3 (22.4)	35.0	5	100
SF-36 role—physical	(0 to 100)	14.5 (27.4)	0.0	0	100
SF-36 bodily pain	(0 to 100)	30.1 (14.8)	31.0	0	62
SF-36 social functioning	(0 to 100)	43.6 (26.4)	37.5	0	100
SF-36 mental health	(0 to 100)	59.1 (20.3)	60.0	12	96
SF-36 emotional role	(0 to 100)	45.2 (42.8)	33.3	0	100
SF-36 vitality	(0 to 100)	38.4 (19.4)	40.0	12	96
SF-36 PCS	(0 to 100)	29.3 (8.1)	28.3	13.8	50.0
SF-36 MCS	(0 to 100)	43.1 (11.7)	41.9	20.9	64.2
EQ-VAS	(0 to 100)	55.4 (21.5)	57.5	19	90
EQ-5D	(− 0.59 to 1)	0.38 (0.33)	0.52	− 0.29	0.80
HUI	(0 to 1)	0.49 (0.19)	0.49	0.17	0.97

Increasing scores in the Oswestry scale are an indication of worse health status. For the other scales, increasing scores reflect better health states.

Thirty-seven patients (80%) completed both the baseline and the 3-month questionnaire, and 34 the baseline and 6-month (74%) assessments. No significant differences were observed in age, gender and duration of low back pain between participants and non-participants at 3 and 6 months. Only minor differences were observed in baseline health status (the EQ-5D and the SF-36 physical role subscale scores were higher in the participant group). After 3 months, one patient (3%) was much better, nine (24%) somewhat better, 20 (54%) mostly the same, three (8%) somewhat worse and four (11%) much worse; at 6 months, four (12%) were much better, five (15%) somewhat better, 18 (53%) mostly the same, five (15%) somewhat worse and two (6%) much worse. Overall, at both time points, approximately half of the patients reported no changes in their health status.

The PCS and MCS scores were analysed in 31 patients at 3 months and in 29 patients at 6 months. The scoring of these aggregate scales requires no missing values in the subscales. Six patients had one incomplete subscale each, either at baseline or at follow-up, which resulted in a smaller number of cases available for the aggregate score. Table 2 shows the correlations between the mean score differences for the various scales and (a) the 5-point ordinal change scale, (b) the mean score differences in the ordinal pain scale, and (c) the Oswestry questionnaire. In general, higher correlations were observed at 6 months. The highest correlations were observed for the EQ-5D, HUI, EQ-VAS and SF-36 bodily pain and physical functioning subscales, and the MCS. Most correlation coefficients were below 0.50.

Table 3 shows the rescaled mean score differences between baseline and 3 months, and baseline and 6 months for all patients combined, and for those patients who improved, those who remained stable, and those who deteriorated. All scores were rescaled within a possible range interval of 0–100, 0 being the worst possible health state and 100 the best possible state. Overall, only small differences in health status were observed when all patients were considered together. When categorizing patients according to global change (improved, stable or worse), most scales appeared to be most sensitive to change at 6 months.

Figure 1 shows the mean effect sizes at 3 months for all three groups. Figure 2 shows the mean effect sizes at 6 months. To illustrate the significance of the effect sizes from a clinical perspective, an effect size of 0.5 corresponds to an improvement of $\sim$10 points in the following SF-36 subscales: general health; physical functioning; mental health. It would also be equivalent to an improvement of 0.16 in the EQ-5D and 0.10 in the HUI. The most discriminative scales at 3 months were the Oswestry questionnaire, the SF-36 emotional role subscale and the HUI. At 6 months, the Oswestry, EQ-5D, HUI and EQ-VAS discriminated between the three groups. For the SF-36, only the emotional role subscale and MCS discriminated between the three groups in the adequate direction. Most SF-36 subscales appeared to have a floor effect for those patients who had deteriorated, with smaller effect sizes or change in the opposite direction. Intraclass correlation coefficients for those patients who remained stable are shown in Fig. 3. The EQ-VAS and the SF-36 emotional role subscale were the least stable measures.

Discussion

Most studies evaluating the performance of health status instruments examine their responsiveness in clinical trials, after patients initiate specific therapies which are expected to produce improvement. In the past few years, however, there has been an emphasis on effectiveness as opposed to efficacy, highlighting the importance of longer-term outcome studies evaluating patients in clinical settings. Very little is known about the performance of specific instruments for patients with ongoing disease. In addition, some of the more recently introduced preference-based instruments have seldom been evaluated for specific diseases. The purpose of this study was to evaluate the discriminative performance over time of various health status measurement tools in patients with LBP seen in clinical settings. The sample size was too

TABLE 2. Correlations between change in the various scales and ordinal change scores, ordinal pain scores and LBP-specific Oswestry scores

	3 months			6 months		
	Ordinal change scale†	Ordinal pain scale†	Oswestry‡	Ordinal change scale†	Ordinal pain scale†	Oswestry‡
Oswestry	0.22	0.20	—	0.37*	0.45*	—
SF-36 general health	0.01	0.11	0.09	0.24	0.27	0.31*
SF-36 physical functioning	0.18	0.06	0.13	0.28	0.20	0.55*
SF-36 role—physical	− 0.13	0.11	− 0.06	0.09	0.02	0.10
SF-36 bodily pain	0.30*	0.03	0.42*	0.48*	0.40*	0.44*
SF-36 social functioning	0.15	0.10	0.31*	0.25	0.20	0.43*
SF-36 mental health	0.19	− 0.34*	0.04	0.17	0.28	0.19
SF-36 emotional role	0.24	0.15	0.06	0.14	0.33*	0.35*
SF-36 vitality	0.05	− 0.14	− 0.13	− 0.06	0.19	0.26
PCS	− 0.02	0.10	0.06	0.23	− 0.04	0.32*
MCS	0.37*	0.07	0.01	0.33*	0.49*	0.41*
EQ-VAS	0.30*	0.08	0.48*	0.46*	0.45*	0.22
EQ-5D	0.10	0.12	0.36*	0.53*	0.39*	0.46*
HUI	0.06	0.13	0.55*	0.27	0.15	0.30*

Instruments are rescaled so that positive correlations indicate agreement: *$P \leqslant 0.05$ (one-tailed); †Spearman correlation coefficient; ‡Pearson correlation coefficient.

TABLE 3. Rescaled mean score differences from baseline for all patients according to change

	3 months				6 months			
	All Mean (S.D.)	Improved $n = 10$	Stable $n = 20$	Worse $n = 7$	All Mean (S.D.)	Improved $n = 9$	Stable $n = 18$	Worse $n = 7$
Oswestry	− 1.4 (10.6)	3.2	− 2.1	− 5.8	0.8 (9.1)	5.2	0.9	− 5.2†
SF-36 general health	− 0.6 (17.0)	− 3.0	1.6	− 4.5	0.4 (12.4)	2.1	3.1†	− 9.3†
SF-36 physical functioning	− 2.2 (12.8)	− 2.8	0.0	− 7.6	4.2 (14.5)	11.3	1.7	2.9
SF-36 role—physical	− 2.0 (23.8)	2.5	− 6.3	3.6	5.9 (28.2)	2.7	11.1	− 3.6
SF-36 bodily pain	− 1.5 (10.5)	2.4	− 3.0	− 2.7	6.9 (11.4)*	12.2†	8.0†	− 2.9†
SF-36 social functioning	9.5 (16.0)*	15.0	6.2	10.7	11.0 (17.9)*	15.3	10.4	7.1
SF-36 mental health	2.0 (13.0)	7.6	− 0.2	0.0	3.8 (16.8)	10.0	1.6	2.3
SF-36 emotional role	4.8 (39.7)	14.8	5.3	− 9.5	2.9 (50.8)	11.1	1.6	− 4.8
SF-36 vitality	0.3 (22.1)	0.5	− 2.3	7.1	3.5 (18.3)	2.2	3.3	5.7
PCS	− 1.1 (5.5)	− 2.5	− 0.7	− 1.0	1.8 (5.0)	2.5	2.9	− 2.3
MCS	2.5 (7.5)	9.0	1.7	− 1.2	1.7 (9.7)	5.7	1.0	− 1.0
EQ-VAS	− 7.3 (3.4)*	2.4	− 11.4	− 11.8	− 4.7 (28.8)	8.2†	2.2†	− 38.1†
EQ-5D	− 4.8 (17.4)	− 5.6	− 1.6	− 13.0	1.8 (16.4)	10.7†	2.0	− 10.4†
HUI	− 1.8 (16.1)	2.5	− 2.6	− 6.3	5.8 (14.5)*	10.4	6.6	− 3.2
Ordinal pain scale	− 4.2 (18.7)	0.5†	− 1.0†	− 20.0†	2.2 (27.7)	23.3†	− 2.2†	− 13.6†

All scales range from 0 to 100, 100 representing the best possible health state for each scale. Positive values denote improvement. *$P \leqslant 0.05$ for differences between baseline and 3 or 6 months for all patients; †$P \leqslant 0.05$ when comparing mean score differences according to change.

small to evaluate statistical differences among the various instruments, and our objective was not to determine which was the best or most responsive tool, but whether the various instruments appear to discriminate adequately among groups of patients who had improved, remained stable or deteriorated.

As expected, the LBP-specific questionnaire, the Oswestry, performed well for all types of patients after both 3 and 6 months, and had a high correlation with the ordinal change and pain scales. This instrument evaluates functional activities that are affected by LBP, and has been shown previously to be reliable and sensitive to change in this population [6]. Our objective in including the Oswestry was not so much to evaluate it but to use it as a benchmark for the assessment of generic instruments.

In general, most of the instruments appeared to perform better at 6 months than at 3 months, and to measure improvement better than deterioration. The best tools appeared to be the Oswestry and the HUI, but our sample was too small to detect statistically significant differences. Most scales were able to discriminate between patients who improved and patients who did not. However, the results in relation to patients who remained stable or worsened were inconsistent. This suggests that these scales are adequate for clinical trials in LBP, but that additional research is required to evaluate their role in clinical settings or as endpoints in outcomes research.

In this study, the SF-36 had some limitations, particularly in relation to patients who experienced deterioration. In these patients, the mean baseline scores for most subscales was $\leqslant 30$, indicating a smaller interval to score deterioration (floor effect). The emotional role

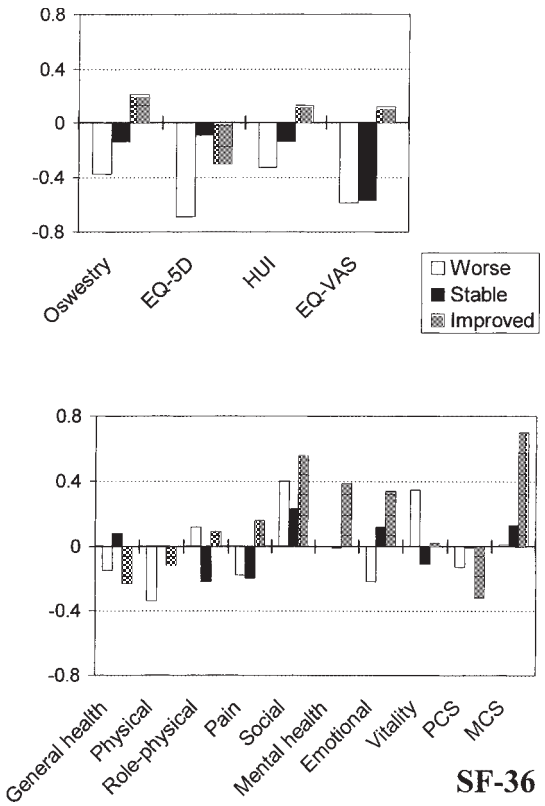

FIG. 1. Effect sizes according to perceived change at 3 months.

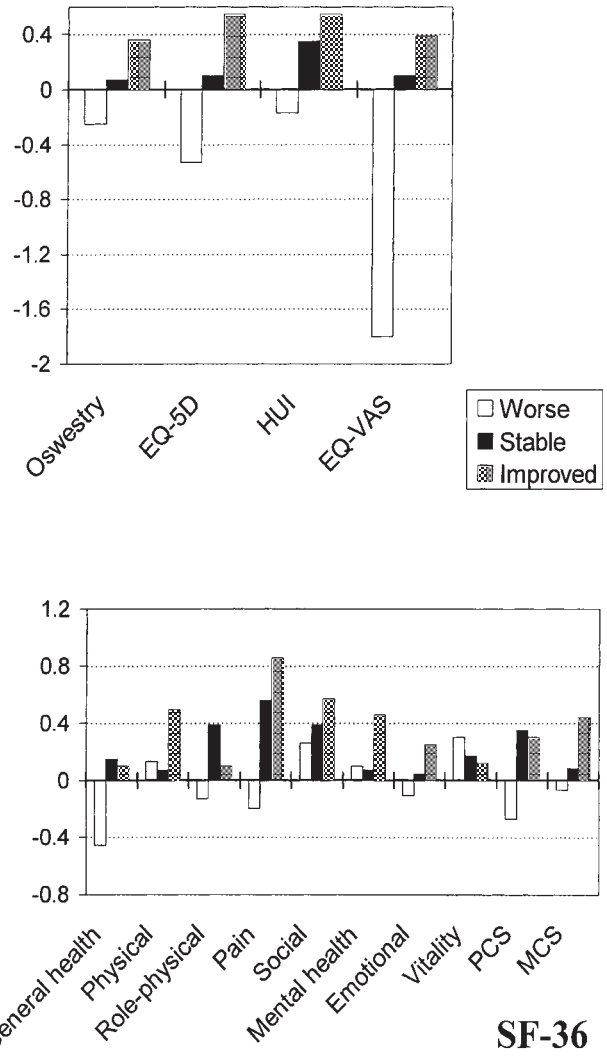

FIG. 2. Effect sizes according to perceived change at 6 months.

subscale discriminated well between patients, at both 3 and 6 months. This subscale is based on three items regarding the interference of emotional problems with daily activities; each item has two possible responses, yes or no. Despite the limited scope and possible variation of the subscale, it discriminated adequately among patients who reported improvement and those who reported deterioration. The baseline score for this variable was close to the midpoint in the interval, allowing increase or decrease in the scores. Since emotional role carries a large weight in the aggregated MCS, this composite score also performed well. On the other hand, the SF-36 physical role limitations subscale, which includes four yes/no dichotomous items regarding the interference of physical problems with daily activities, had poor responsiveness. Baseline scores were low, suggesting a floor effect, yet the subscale did not adequately reflect improvement either, suggesting low variability in these items. The SF-36 physical function subscale was inconsistent, with a good response for deterioration at month 3 but a good response only for improvement at month 6. The performance of the PCS composite score was mediocre, which could be attributed partly to the poor discrimination observed with the physical subscales, which carry substantial weight in the scoring of the PCS. The SF-36 pain subscale differentiated between improvement and deterioration but was not stable for patients with no change. The other subscales did not adequately or consistently discriminate between patients. Some of these inconsistencies in the SF-36 may reflect random variation due to the small sample size, yet the Oswestry and some of the other subscales showed adequate discrimination. Despite being a much simpler measure, the generic EQ-5D VAS appeared to perform better than most SF-36 subscales in discriminating among those who improved and those who became worse. The EQ-VAS, however, was not very stable for patients reporting no change.

Generic health status instruments are increasingly being used to compare health status across diseases. A concern with these tools is that they do not adequately evaluate specific symptoms and, therefore, cannot be sensitive to change in those dimensions which are important to the patient. The major concern of patients with LBP is likely to be their pain. Specific instruments, such as the Oswestry, evaluate pain and specific functional aspects that can be impaired by spinal pain. The SF-36 subscales evaluate various activities which perhaps are not affected as strongly by LBP. The most discriminative scales in the SF-36 were those related to mental and emotional health, suggesting that this com-

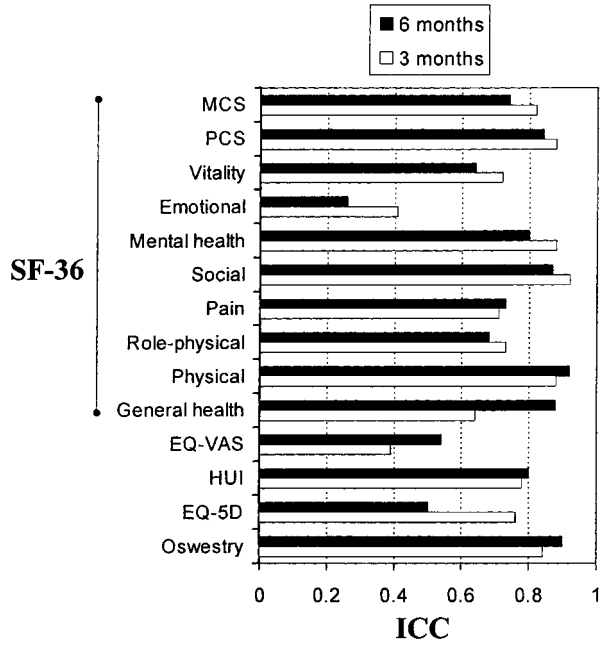

FIG. 3. Intraclass correlation coefficients (ICC) for patients reporting no change.

ponent of the health status is very important in the patient's perception of improvement, or that perception of improvement has a major effect on mental health.

The use of preference-based tools is being advocated to provide a societal view on the value of health states. In this study we used two instruments, the EQ-5D and the HUI, with imputed weights from population groups. Both instruments appeared to perform well at 6 months, although the EQ-5D did not respond well at 3 months for improved patients. We used UK York weights for the EQ-5D, which may not reflect the health state preferences in the Canadian population. Unfortunately, there are currently no valuation studies in the North American population that could be considered to be more satisfactory than the UK weights. It is conceivable that the general public weights pain more heavily than other aspects of health. This would therefore translate into better responsiveness to change when the major manifestation in a specific condition is pain. In fact, the EQ-5D and HUI weights assign the largest disutility function (which translates into lower utility scores) to the worst level of pain, compared with the worst levels of the other attributes. This suggests that, in fact, pain is valued as having a large impact on quality of life by the general population, more so than other domains. Beaton *et al.* [15] evaluated five generic health status instruments in workers with musculoskeletal disorders, including LBP. They concluded that the SF-36 was the most responsive questionnaire, performing better than the Nottingham Health Profile, the Ontario Health Survey, the Duke Health Profile and the Sickness Impact Profile. However, their study did not include LBP-specific measures or preference-based instruments. Most of the previous research comparing the performance of the SF-36 with specific measures in patients with musculoskeletal disorders has been conducted in patients undergoing arthroplasty and patients with inflammatory arthritis, such as rheumatoid arthritis [16–18]. The results in the arthroplasty population suggest that specific and generic instruments provide distinct but complementary information, and that both types of tools are very responsive to change. Most of these patients, however, show marked improvement after surgery, which can be detected easily. In fact, the major problem with the SF-36 subscales appeared to be a floor effect for patients who had deteriorated. In the general population as well, the SF-36 behaves differently, with mostly a ceiling effect. Our findings also suggest that these various instruments measure different aspects of health, since the correlation coefficients between scales were modest. The lack of responsiveness of some subscales suggests that they may not measure domains as relevant to the patient with LBP as those assessed by other, more discriminative tools.

In summary, in this study the SF-36 had some limitations in discrimination among patients with LBP in relation to their reported improvement. Preference-based quality-of-life instruments appeared to discriminate among patients who improved and those who deteriorated, although not as consistently as the disease-specific tool. Additional research is needed to evaluate the role of generic and preference-based measures of quality of life in the assessment of patients with LBP in clinical settings.

Acknowledgements

The authors would like to thank Dr David Feeny for allowing the use of the Health Utilities Index for this study and for providing the scoring algorithms. We are grateful to Mrs Brenda Topliss for assisting in the preparation of this manuscript. At the time of the study, Dr Maria Suarez-Almazor was an Arthritis Society Scholar. Drs Maria Suarez-Almazor and Jeffrey Johnson were recipients of Alberta Heritage Foundation for Medical Research Population Health Investigator Awards. Mr Chris Kendall was a summer student supported by the Alberta Heritage Foundation for Medical Research.

References

1. Abenhaim L, Suissa S. Importance and economic burden of occupational back pain: a study of 2500 cases representative of Quebec. J Occup Med 1987;29:670–4.
2. Deyo RA, Cherkin D, Conrad D, Volinn E. Cost, controversy, crisis: low back pain and the health of the public. Annu Rev Public Health 1991;12:141–56.
3. Deyo RA, Rainville J, Kent DL. What can the history and physical examination tell us about low back pain? J Am Med Assoc 1992;12:760–5.
4. Michel A, Kohlmann T, Raspe H, Main CJ: The association between clinical findings on physical examination and self-reported severity in back pain: Results of a population-based study. Spine 1997;22:296–304.

5. Deyo RA, Andersson G, Bombardier C, Cherkin DC, Keller RB, Lee CK *et al.* Outcome measures for studying patients with low back pain. Spine 1994;19:2032S–6S.
6. Fairbank JCT, Davies JB, Mbaot JC, O'Brien JP. The Oswestry Low Back Pain Disability Questionnaire. Physiotherapy 1980;66:271–3.
7. Roland M, Morris R. A study of the natural history of back pain. Part I: development of a reliable and sensitive measure of disability in low back pain. Spine 1983;8:141–4.
8. Brook R with the EuroQol group. EuroQol: the current state of play. Health Policy 1996;37:53–72.
9. Feeny DH, Torrance GW, Furlong WJ. Health Utilities Index. In: Spilker B, ed. Quality of life and pharmacoeconomics in clinical trials. 2nd edition. Philadelphia: Lippincott-Raven, 1996:239–52.
10. Kaplan RM, Anderson J. A general health policy model: update and applications. Health Serv Res 1988;23:203–35.
11. Guyatt G, Feeny D, Patrick D. Measuring health-related quality of life. Ann Int Med 1993;118:622–9.
12. Ware JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992; 30:473–83.
13. McHorney CA, Ware JE, Lu JF, Sherbourne CD. The MOS 36-item short-form health survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. Med Care 1994;32:40–66.
14. Dolan P. Modeling valuations for EuroQol health states. Med Care 1997;35:1095–108.
15. Beaton DE, Hogg-Johnson S, Bombardier C. Evaluating changes in health status: Reliability and responsiveness of five generic health status measures in workers with musculoskeletal disorders. J Clin Epidemiol 1997;50:79–93.
16. Stucki G, Liang MH, Phillips C, Katz JN. The short form-36 is preferable to the SIP as a generic health status measure in patients undergoing elective total hip arthroplasty. Arthritis Care Res 1995; 8:174–81.
17. Katz JN, Larson MG, Phillips CB, Fossel AH, Liang MH. Comparative measurement sensitivity of short and longer health status instruments. Med Care 1992; 30:917–25.
18. Ruta DA, Hurst NP, Kind P, Hunter M, Stubbings A. Measuring health status in British patients with rheumatoid arthritis: Reliability, validity and responsiveness of the short form 36-item health survey (SF-36). Br J Rheumatol 1998;37:425–36.