

## Is a Condition-Specific Instrument for Patients with Low Back Pain/Leg Symptoms Really Necessary?

The Responsiveness of the Oswestry Disability Index, MODEMS, and the SF-36

Thomas L. Walsh, MS, PT, OCS, Dip MDT,\*§ Brett Hanscom, MS,†  
Jon D. Lurie, MD, MS,\*‡§|| and James N. Weinstein, DO, MS\*†‡§||

**Study Design.** Analysis of longitudinal data collected prospectively from patients seen in 27 National Spine Network member centers across the United States.

**Objective.** To evaluate the responsiveness of the Oswestry Disability Index, MODEMS scales, and all scales and summary scales of the MOS Short-Form 36 (SF-36) for patients with low back pain/leg symptoms.

**Summary of Background Data.** The responsiveness of general and condition-specific health status instruments is a key concept for clinicians and scientists. Various authors have explored responsiveness in common surveys used to assess spine patients. Although it is generally believed that condition-specific measures are more responsive to change in the condition under study, in the case of low back pain, most authors agree that further exploration is necessary.

**Methods.** Patients with diagnoses of herniated disc, spinal stenosis, and spondylosis from the National Spine Network database who completed baseline and 3-month follow-up surveys were analyzed. Patient-provider consensus regarding improvement, worsening, or no change in the condition was selected as the external criterion. Responsiveness was evaluated using ROC curve analysis and effect size calculations.

**Results.** Nine hundred and seventy patients had complete data at baseline and 3 months. At follow-up, 68% of the patients had consensus improvement. Based on ROC analysis, scales assessing pain were significantly more responsive than scales assessing function. There were no significant differences between the condition-specific scales and their equivalent general-health counterpart. The scales with the highest probabilities of correctly identifying patient's improvement were: the condition-specific

pain scale from MODEMS (PAIN, ROC = 0.758); the combined pain and function scale from MODEMS (MPDL, ROC = 0.755); the general pain scale from the SF-36 (BP, ROC = 0.753); the combined pain and function scale from the SF-36 (PCS, ROC = 0.745); the condition-specific function measure from the Oswestry (ODI, ROC = 0.723); and the physical function measure from the SF-36 (PF, ROC = 0.721). A similar rank order was typically maintained with effect size calculations. Results were nearly identical in patients with multiple non-spine-related comorbidities and in patients with high degrees of perceived disability. The BP scale was most responsive to worsening of symptoms.

**Conclusion.** For studies of patients with low back problems, the general SF-36 may be a sufficient measure of health status and patient function, without the need for additional condition-specific instruments. Pain scales appear to be the most responsive measures in patients with low back pain. [Key words: responsiveness, SF-36, Oswestry Disability Index, National Spine Network, low back pain] **Spine 2003;28:607-615**

Low back pain (LBP), often with associated leg symptoms, is a common disorder that causes significant disability. Unfortunately, physiologic measurements have shown little relevance for patients with back and leg symptoms. It is therefore difficult to apply traditional clinical measures of "success" and "effectiveness" in this patient population. For these reasons, surveys of patient self-reported health and function have become common tools for assessing patient outcome. Although these surveys are becoming ubiquitous, it remains unclear which surveys are most useful and how many surveys are necessary to adequately assess whether clinical treatments have been effective.<sup>8</sup>

Health status assessment surveys typically fall into two categories, general and condition-specific. General health surveys measure overall health with a broad range of questions covering a variety of aspects of health. Condition-specific instruments focus on the specific symptoms or functional impact of a particular condition. It is commonly felt that the usefulness of general measures lies in their ability to allow comparisons among patients with the same condition as well as between patients with different conditions. Moreover, general measures may be able to identify unsuspected side effects from a new treatment. Advocates for condition-specific surveys theorize that these measures should be more responsive to change in the condition under study compared to general health

From \*The Spine Center and †Section of Orthopedics at Dartmouth Hitchcock Medical Center, Lebanon, ‡Department of Community and Family Medicine, Dartmouth Medical School, Hanover, and §The Center for the Evaluative Clinical Sciences at ||Dartmouth Medical School, Hanover, New Hampshire.

Supported in part by: The National Institute of Arthritis and Musculoskeletal and Skin Diseases and the Office of Research on Women's Health, the National Institutes of Health, and the National Institute of Occupational Safety and Health, the Centers for Disease Control and Prevention NIAMS # AR45444-01A1, and the members of the National Spine Network.

Acknowledgment date: March 25, 2002. First revision date: July 29, 2002. Acceptance date: August 19, 2002.

The manuscript submitted does not contain information about medical device(s)/drug(s).

No funds were received in support of this work. No benefits in any form have been or will be received from a commercial party related directly or indirectly to the subject of this manuscript.

Address correspondence and reprint requests to Thomas Walsh, MS, PT, OCS, Dip MDT, The Spine Center, Dartmouth Hitchcock Medical Center, One Medical Center Drive, Lebanon, NH 03756, USA; E-mail: Thom@Hitchcock.org

measures, especially in patients with multiple comorbidities. These goals—measuring general health across differing diseases and measuring the specific impact of a particular condition in a specific population—have the potential to be mutually exclusive, and it is therefore typically recommended that both types of measures be used clinically and for research.

Utilization of both types, however, increases respondent burden, creates redundancy, and adds to data collection and analysis burdens. If a general measure were as responsive as a condition-specific measure for a particular condition, the benefits of the general measure could be maintained, and the condition-specific measure could be culled without risk of losing information.

A commonly used set of outcomes instruments in patients with LBP/leg symptoms is the condition-specific, function assessing Oswestry Low Back Pain Disability Questionnaire, the condition-specific pain and function assessing MODEMS questionnaire, and the general health, pain, and function assessing SF-36. The SF-36 has been systematically developed and rigorously analyzed, meeting criteria for internal consistency, reliability, and validity for use in the general population and in patients with symptoms of back pain and sciatica with and without surgical intervention.<sup>6,14,15,19,22,27,28</sup> The Oswestry is a common condition-specific tool that has been used in over 200 published articles since its inception in 1980.<sup>1,9,10,12</sup> The MODEMS questionnaire is also widely used, but its reliability and validity have not been as rigorously assessed.<sup>25</sup> Whether the condition-specific measures (Oswestry and MODEMS) offer increased responsiveness over the SF-36 to changes in spine conditions has not been conclusively demonstrated.

Using the National Spine Network (NSN) database, we endeavored to quantify the ability of these surveys to detect changes in spine conditions, *i.e.*, their responsiveness, and to assess whether the condition-specific surveys are necessary in order to adequately measure changes in spine conditions.

## ■ Methods

**The National Spine Network and the NSN Health Status Survey.** The National Spine Network is a consortium of spine-care centers from across the United States that collaborate in collecting outcomes data on their patients. The NSN has developed an outcomes survey specifically for spine patients—the *NSN Health Status Survey*—that contains both patient- and clinician-based questions. The NSN was formed in 1995 and since that time, the organization has collected over 100,000 questionnaires. Data collection is ongoing.

The patient portion of the NSN survey includes the Oswestry Low Back Pain Disability Questionnaire, version 2.0 with the pain question (item 1) removed, a portion of the MODEMS low back questionnaire, and the SF-36 Health Survey, as well as a variety of questions regarding demographics, symptom history, and comorbid illness. The clinician survey contains a brief reporting of clinical symptoms, diagnoses, and treatment plans. Participating clinicians administer the survey to patients when they arrive at a clinic for the first time and subsequently at

follow-up visits. In addition to the original set of questions, the follow-up visit survey asks both the patient and the clinician to evaluate progress, or worsening, of the spine condition.

**Study Cohort.** Patients with diagnoses of herniated nucleus pulposus, spinal stenosis, or spondylosis from the NSN Database with an initial-visit survey and a follow-up visit survey at 3 months were selected for this analysis. The Oswestry questions were not added to the NSN Health Status Survey until 1998, so only patients who completed their initial visit after January 1998 were selected. Limiting to these criteria produced 970 patients with 3-month follow-up. Among patients in our study sample, 27% had surgery, 31% had injections, 61% took medications, 53% had physical therapy, 18% had manipulation (*e.g.*, chiropractic), 52% did home exercise, 10% used a splint or brace, 6% had other interventions, and 7% had reported having no interventions. Most patients had multiple forms of treatment.

**Patient and Clinician Report as Criteria for Change.** Patient and clinician consensus was used as a measure of the “true” change in the patient’s condition from initial evaluation to follow-up. At the time of follow-up, patients provided an overall assessment of their progress on a five-point scale: “Compared to when you last completed this questionnaire, is your musculoskeletal condition 1) much better now; 2) somewhat better now; 3) about the same; 4) somewhat worse now; 5) much worse now.” Physicians provide an independent assessment on a five-point scale: “Patient progress 1) major improvement; 2) minor improvement; 3) no change; 4) minor worsening; 5) major worsening.”

These two scales were reduced to a single dichotomous “improvement” variable as follows: if both the patient and the clinician agreed that some improvement had occurred (patient reported either “much better” or “somewhat better” and clinician reported “major improvement” or “minor improvement”), then the patient was considered to have improved. If only the patient or the clinician rating was available, that single rating was used to determine whether the patient had improved. If both clinician and patient reported no change or worsening, or if there was disagreement between the patient and clinician, the patient was considered to have not improved.

Patient “worsening” was defined in a similar manner, with patients who had both self-reported worsening and clinician-reported worsening being defined as “worse.” Two subgroups were analyzed as well: 1) patients with self-reports of two or more comorbidities; and 2) patients reporting severely low physical functioning, defined as a baseline Physical Component Summary and Oswestry Disability Index scores below the population median.

**Oswestry.** The Oswestry Low Back Pain Disability Questionnaire is a 10-question, low back-specific instrument designed to measure disability in spine patients.<sup>10</sup> The NSN Health Status Survey uses 9 of the 10 items (items 2–10) from the American version of the questionnaire developed by the American Academy of Orthopedic Surgeons for the MODEMS project. Each question has 6 possible responses that are scored from 0 to 5. All scores are summed and divided into the highest possible score of 45, producing a 0 to 100-percentage scale, with 0 representing no disability and 100 representing complete disability. For missing responses, the total possible score is reduced (*e.g.*, the highest possible score for 8 responses would be

40). This scale is referred to as the Oswestry Disability Index (ODI). For ease in clinical interpretation, this score was then subtracted from 100. This inversion yields a scale that “moves” similarly to the SF-36 and MODEMS scales, where 0 represents complete disability and 100 represents normal function. The inversion of the score does not affect any statistical calculations or mathematical relationships. Clinicians/scientists accustomed to the traditional reporting of the ODI should pay special attention to this inversion when interpreting the absolute disability of the subjects in this study.

**MODEMS.** The Musculoskeletal Outcomes Data Evaluation and Management System (MODEMS) is a collection of outcomes instruments for musculoskeletal conditions. The Lumbar Spine Instrument contains a variety of spine-specific pain and symptom questions, as well as the Oswestry questionnaire. Many of the pain and symptom items on the NSN Health Survey are identical to the MODEMS items. Several measurement scales can be calculated from responses to the survey, including a comorbidity index, a measure of patient expectations, a neurogenic symptom scale, and a pain/disability scale. For the purposes of this study, only the pain/disability scale was used.

The MODEMS Pain and Disability Lumbar (MPDL) scale is a combination of the nine Oswestry questions (described earlier), the question “In the past week, how often have you suffered: low back and/or buttocks pain?” (six-item response from “none of the time” to “all of the time”), and the question “In the past week, how bothersome have these symptoms been? Low back and/or buttocks pain?” (six-item response from “not at all bothersome” to “extremely bothersome”). The sum of patient responses to these 11 questions is divided into the maximum possible score, yielding a percent (0–100) scale.

The PAIN scale (not an official MODEMS scale) is a combination of the two (non-Oswestry) questions from the MPDL scale, and the two additional questions: “In the past week, how often have you suffered: leg pain?” (six-item response from “none of the time” to “all of the time”), and “In the past week, how bothersome have these symptoms been? Leg pain?” (six-item response from “not at all bothersome” to “extremely bothersome”). The sum of patient responses to these 4 questions is divided into the maximum possible score, yielding a percent (0–100) scale.

**The SF-36.** The Short Form 36 Health Survey is a 36-item general health instrument that measures 8 domains of health.<sup>28</sup> Multi-item scales have been developed to represent aspects of the following eight physical and mental health domains: Physical Functioning (PF), Role Physical (RP), Bodily Pain (BP), General Health (GH), Vitality (VT), Social Functioning (SF), Role Emotional (RE), and Mental Health (MH). Each scale is a weighted combination of between 2 and 10 items and is scored as a 0 to 100% score with 0 representing severe disability and 100 representing no disability. Scoring only occurs if the patient has responded to at least half of the relevant items.<sup>27</sup>

In addition to the eight scales, the SF-36 authors have developed two summary scales that provide a more concise measure of overall physical and mental health.<sup>27</sup> The Physical Component Summary (PCS) and Mental Component Summary (MCS) are linear combinations of all eight of the original scales, with the PCS heavily weighting physical measures and the MCS heavily weighting mental health measures. The PCS and MCS are adjusted by the population mean and standard

deviation to produce norm-based scores with a common mean of 50 and standard deviation of 10. Thus, any score below 50 represents a decrement from “normal” health and functioning. The summary scales are only scored if sufficient data exists to compute all eight SF-36 scales.<sup>27</sup>

**ROC Curve Analysis.** Our primary measure of responsiveness for the SF-36, Oswestry, and MODEMS is based on Receiver Operating Characteristic (ROC) Curve analysis. Deyo and Centor have argued that assessing responsiveness in a scale is akin to measuring the discriminatory properties of a diagnostic test where both true and false positives and negatives may occur.<sup>7</sup> A scale, like the diagnostic test, can be described in a clinically meaningful way, based on its ability to correctly identify a clinically important change. As with diagnostic tests, a survey instrument’s sensitivity is of primary importance, but its specificity must be considered as well. Receiver Operating Characteristic analysis, by assessing both sensitivity and specificity, provides clinically meaningful data and a more complete assessment of a scale’s discriminatory properties. Receiver Operating Characteristic analysis requires the use of a gold standard for comparison. Our study makes use of a “gold standard” measure of actual change: patient–clinician consensus improvement. Although this is not a true gold standard, it is difficult to argue that no improvement has occurred if both the patient and the clinician independently and simultaneously report improvement. The ROC technique estimates the probability that a particular scale will correctly identify patients who have improved (as measured by the gold standard: consensus improvement). Probabilities range from 0.50 to 1.00, with 0.50 representing a scale with no ability to identify improved patients and 1.00 representing the ability to perfectly discriminate between improved and unimproved patients. The change score calculated for each survey by comparing the initial appointment and 3-month follow-up was entered into the ROC calculation to assess the scale’s ability to predict improvement as measured by our gold standard.

**Effect Sizes.** As an additional measure of responsiveness we calculated Effect Size statistics for each SF-36 scale, MODEMS scale and the ODI. For patients who have improved (as measured by the patient–clinician consensus variable) the effect size may be calculated as the mean change in score divided by the standard deviation of the score at baseline. Generally referred to as the Standardized Effect Size, this ratio is intended to capture the degree to which a scale changes in value above and beyond the standard fluctuation in baseline scores.<sup>2,4</sup> Effect size calculations measure the ability of the scale to detect a signal (improvement) among the noise (normal variation) of a patient population.

Two other measures of effect size are commonly used. The Standardized Response Mean is computed as the mean change score divided by the standard deviation of the change scores themselves. This method represents a more traditional statistical standardization of the change scores.<sup>2</sup>

A third method was proposed by Guyatt and involves dividing the mean change score by the standard deviation of the change scores from a stable group of patients (*i.e.*, patients who got neither worse nor better).<sup>16</sup> This is an attempt to detect change above and beyond the random changes experienced by patients who do not experience detectable improvement.

We examined all three versions of Effect Size in our analysis. The rank order of responsiveness for each scale as measured by

**Table 1. Patient Characteristics of NSN Database Compared to the Study Sample**

Patient Characteristics	NSN Database* (n = 14,708)	Study Cohort† (n = 970)	P Value
Age [mean (SD)]	51.2 (15)	49.4 (14)	0.000
Gender (% female)	48.7%	48.0%	0.675
Education (% college graduates)	34.5%	31.9%	0.010
Work (% working)	47.2%	43.7%	0.000
Compensation (% receiving)	10.6%	16.4%	0.000
Comorbidities [mean (SD)]	1.7 (1.9)	1.5 (1.6)	0.001
Duration of spine problems at baseline			0.955
0–6 mos	23.5%	23.6%	
6–12 mos	9.1%	9.7%	
1–2 yrs	11.9%	11.7%	
>2 yrs	55.4%	55.0%	
Prior surgeries (%)	17.8%	21.1%	0.009
Baseline health status [mean (SD)]			
SF-36–PF	42.4 (28)	38.1 (27)	0.000
SF-36–BP	29.4 (20)	24.8 (18)	0.000
SF-36–RP	16.2 (30)	10.6 (24)	0.000
SF-36–GH	60.8 (22)	61.0 (22)	0.805
SF-36–MH	64.5 (21)	61.6 (21)	0.000
SF-36–VT	39.5 (22)	36.2 (21)	0.000
SF-36–SF	51.1 (29)	46.1 (27)	0.000
SF-36–RE	57.1 (44)	55.1 (44)	0.082
SF-36–PCS	30.4 (9)	28.8 (8)	0.000
SF-36–MCS	46.4 (12)	45.1 (12)	0.001
ODI	57.3 (21)	52.3 (20)	0.000
MPDL	53.3 (21)	48.0 (20)	0.000
PAIN	38.5 (30)	33.7 (29)	0.000
Diagnosis			0.000
Herniated disk	35.6%	46.9%	
Spinal stenosis	21.8%	21.7%	
Spondylosis	42.6%	31.3%	

PF = Physical Function; BP = Bodily Pain; RP = Role Physical; GH = General Health; MH = Mental Health; VT = Energy and Fatigue/Vitality; SF = Social Functioning; RE = Role Emotional; PCS = Physical Component Summary; MCS = Mental Component Summary; ODI = Oswestry Disability Index; MPDL = MODEMS Pain/Disability Lumbar; PAIN = MODEMS Back/Leg Pain Scale.

\* All patients from the National Spine Network Outcomes Database with a diagnosis of herniated disk, spinal stenosis, or spondylosis.

† All patients from the NSN Outcomes Database with herniated disk, spinal stenosis, or spondylosis who returned and completed a follow-up survey at approximately 3 months.

ROC curve analysis was compared to the rank order of responsiveness for each scale as measured by Effect Size calculations.

**Statistical Analysis.** Comparisons of patient characteristics were performed using  $\chi^2$  analyses for categorical data and *t* tests for continuous data. Estimation of ROC curve areas and comparisons of ROC curve probabilities were performed as described by Hanley and McNeil.<sup>17,18</sup> Calculations for effect sizes were performed using methods described by Beaton<sup>2</sup> and Guyatt *et al.*<sup>16</sup> All analyses were conducted using Stata 7 (College Station, TX).

## ■ Results

The 970 patients eligible for this study were roughly comparable to the 14,708 patients in the NSN database with the same diagnoses (herniated nucleus pulposus [HNP], spinal stenosis, spondylosis). Forty-eight percent of the study group was female, the average age was 49, and 32% were college graduates (Table 1). Patients in

**Table 2. Responsiveness of Key SF-36 Scales, MODEMS, and the ODI, as Measured by the Area Under the ROC Curve\* (n = 970)**

Responsiveness to Improvement		Responsiveness to Worsening	
Scale	ROC Area (95% CI)	Scale	ROC Area (95% CI)
PAIN	0.758 (0.728–0.788)	SF-36–BP	0.773 (0.719–0.828)
MPDL	0.755 (0.725–0.786)	PAIN	0.761 (0.712–0.810)
SF-36–BP	0.753 (0.722–0.783)	MPDL	0.759 (0.707–0.810)
SF-36–PCS	0.745 (0.714–0.775)	SF-36–PCS	0.743 (0.685–0.801)
ODI	0.723 (0.691–0.755)	ODI	0.737 (0.682–0.791)
SF-36–PF	0.721 (0.689–0.753)	SF-36–PF	0.736 (0.680–0.792)
SF-36–SF	0.684 (0.650–0.717)	SF-36–VT	0.698 (0.643–0.754)
SF-36–VT	0.669 (0.634–0.704)	SF-36–GH	0.693 (0.631–0.755)
SF-36–RP	0.642 (0.613–0.672)	SF-36–SF	0.665 (0.602–0.729)
SF-36–MH	0.625 (0.588–0.662)	SF-36–RE	0.647 (0.588–0.707)
SF-36–GH	0.615 (0.579–0.651)	SF-36–RP	0.641 (0.593–0.690)
SF-36–MCS	0.584 (0.547–0.621)	SF-36–MCS	0.628 (0.560–0.696)
SF-36–RE	0.560 (0.525–0.595)	SF-36–MH	0.615 (0.548–0.683)

ROC = Receiver Operating Characteristic; CI = confidence interval; PAIN = MODEMS Back/Leg Pain Scale; MPDL = MODEMS Pain/Disability Lumbar; BP = Bodily Pain; PCS = Physical Component Summary; ODI = Oswestry Disability Index; PF = Physical Function; SF = Social Functioning; VT = Energy and Fatigue/Vitality; RP = Role Physical; MH = Mental Health; GH = General Health; MCS = Mental Component Summary; RE = Role Emotional.

\* Scales are ranked according to ROC curve area (the higher the ROC area, the better a scale can discriminate between changed and unchanged patients).

the study group reported similar symptom duration at baseline, with approximately 55% having had symptoms for more than 2 years and reported a similar number of comorbidities. Our study cohort consisted of a proportionally greater number of patients with HNP and fewer patients with spondylosis compared to the full NSN sample. Our cohort tended to have lower baseline health status. Due to the large sample size, all differences, except for the measures of General Health and Role Emotional, were statistically different.

Based on ROC analysis, pain scales were significantly more responsive, statistically, than function scales (Table 2). There were no significant differences between the condition-specific scales and their equivalent general health counterpart. The statistical difference between the pain and function measures is small and of questionable clinical significance. The areas under the ROC curves are nearly identical (Figure 1).

The scales with the highest probabilities of correctly identifying patients' improvement were: the condition-specific pain scale from MODEMS (PAIN, ROC = 0.758); the combined pain and function scale from MODEMS (MPDL, ROC = 0.755); the general pain scale from the SF-36 (BP, ROC = 0.753); the combined pain and function scale from the SF-36 (PCS, ROC = 0.745); the condition-specific function measure from the Oswestry (ODI, ROC = 0.723); and the physical function measure from the SF-36 (PF, ROC = 0.721) (Table 2). A similar rank order was typically maintained with effect size calculations. Results were nearly identical in patients with multiple non-spine-related comorbidities and in patients with high degrees of perceived disability.

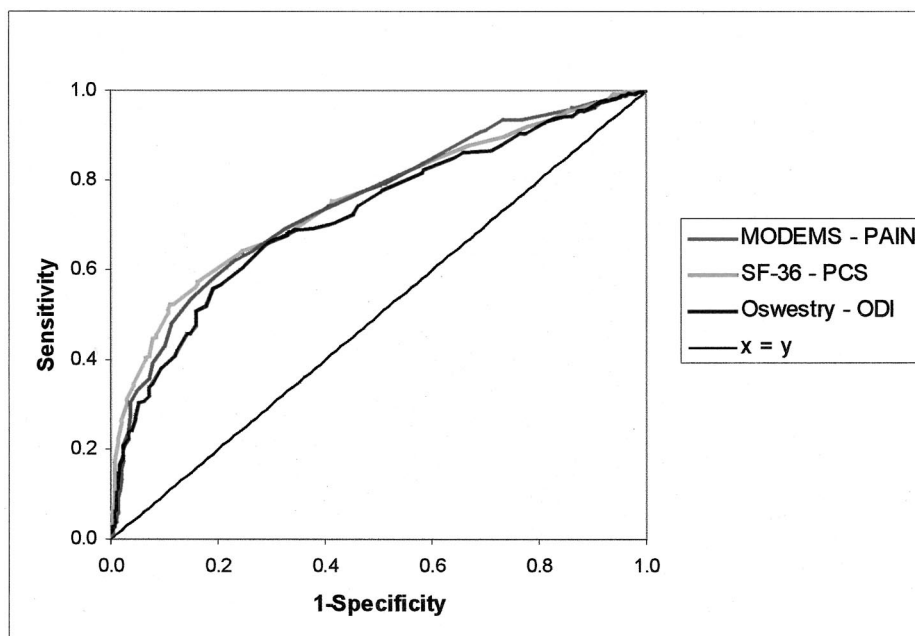


Figure 1. Receiver Operating Characteristic Curves for key pain and pain/function scales.

Responsiveness to worsening was similar to responsiveness to improvement in that pain measures were more responsive than measures of function. Surprisingly, most ROC scores were higher, revealing a higher degree of responsiveness in patients with consensus worsening. And here, the SF-36's BP scale was the most responsive (ROC = 0.773). Following were the nearly identical PAIN (ROC = 0.761) and MPDL (ROC = 0.759) scales from MODEMS. The PCS score (ROC = 0.743) was next. The condition-specific measure of function ODI (ROC = 0.737) and its general counterpart from the SF-36, PF (ROC = 0.736) were nearly identical.

Results for the two subgroups—patients with multiple comorbidities and patients with low function—were surprisingly similar to those for the entire group. Measures of patient self-reported pain were statistically superior to patient reported function. No condition-specific measure significantly outperformed its general health counterpart. The responsiveness of all measures was slightly less in patients with two or more self-reported comorbidities but the rank order was relatively unchanged (Table 3). In patients reporting very low function, all scales were slightly more responsive.

Effect size calculations were performed using three standard methods. The rank orders of responsiveness by

**Table 3. Responsiveness of Key SF-36 Scales, MODEMS, and the ODI, as Measured by the Area Under the ROC Curve, Among Patients Who Have Multiple Comorbid Illnesses and Patients Who Have Low Baseline Physical Function (Below Median PCS and ODI)**

	General Health Scales		Condition-Specific Scales		
	Scale	ROC Area	Scale	ROC Area	P Value
All patients (n = 970)					
Pain scales	SF-36-BP	0.753	PAIN	0.758	0.78
Pain and function scales	SF-36-PCS	0.745	MPDL	0.755	0.52
Function scales	SF-36-PF	0.721	ODI	0.723	0.89
		(P = 0.06)*		(P = 0.04)†	
Patients with 2+ comorbidities (n = 384)					
Pain scales	SF-36-BP	0.714	PAIN	0.743	0.36
Pain and function scales	SF-36-PCS	0.706	MPDL	0.707	0.96
Function scales	SF-36-PF	0.694	ODI	0.667	0.30
		(P = 0.46)*		(P = 0.01)†	
Patients with low baseline function (n = 380)					
Pain scales	SF-36-BP	0.765	PAIN	0.782	0.51
Pain and function scales	SF-36-PCS	0.764	MPDL	0.793	0.24
Function scales	SF-36-PF	0.753	ODI	0.767	0.56
		(P = 0.62)*		(P = 0.54)†	

ROC = Receiver Operating Characteristic; BP = Bodily Pain; PCS = Physical Component Summary; PF = Physical Function; PAIN = MODEMS Back/Leg Pain Scale; MPDL = MODEMS Pain/Disability Lumbar; ODI = Oswestry Disability Index.

\* P value for test comparing the ROC areas of the BP scale and the PF scale.

† P value for test comparing the ROC areas of the PAIN scale and the ODI scale.

**Table 4. Effect Sizes for Key SF-36 Scales and the ODI—Rank Ordered**

Rank	Effect Size 1 (SRM)		Effect Size 2 (SES)		Effect Size 3 (Guyatt)	
	Scale	ES	Scale	ES	Scale	ES
1	PAIN	1.08	SF-36-BP	1.46	SF-36-BP	1.53
2	SF-36-BP	1.00	PAIN	1.01	PAIN	1.12
3	MPDL	0.95	MPDL	0.99	MPDL	1.02
4	SF-36-PCS	0.82	SF-36-PCS	0.97	SF-36-PCS	0.96
5	ODI	0.82	ODI	0.87	ODI	0.87
6	SF-36-PF	0.74	SF-36-PF	0.76	SF-36-PF	0.75

ES = effect sizes; PAIN = MODEMS Back/Leg Pain Scale; BP = Bodily Pain; MPDL = MODEMS Pain/Disability Lumbar; PCS = Physical Component Summary; ODI = Oswestry Disability Index; PF = Physical Function.

The three effect sizes calculation techniques include: 1) Standardized Response Mean (SRM); 2) Standardized Effect Size (SES); and 3) Guyatt effect size.

these measures were essentially the same as the results presented above (Table 4).

### Discussion

Our data do not support the contention that there is significant added value in terms of responsiveness from using the condition-specific Oswestry and/or MODEMS instruments in addition to the SF-36 in patients with lumbar spine/leg complaints within the National Spine Network. In our analysis of 970 patients, the scales assessing pain were always statistically superior to the scales measuring function; however, the condition-specific measures were never statistically superior to their general counterparts. These findings were true for patients reporting improvement and worsening, for patients with multiple comorbidities, as well as for patients with low baseline function. Thus, even in the samples of patients in which the condition-specific measure would seem best suited, these measures were not superior to the SF-36.

Responsiveness is only one characteristic to consider when choosing a survey instrument. The ODI is shorter than the SF-36 and if computerized survey administration is not available, it may be easier to administer and score. If ease of administration and scoring is a top priority, then one may choose to utilize the ODI, and responsiveness would not significantly suffer. The cost of such a decision at least includes the loss of general health information, the ability to compare disability between patients with differing conditions, and the possible identification of unintended side effects from new treatment.

Our sample of patients statistically differed from patients in the NSN as a whole (Table 1); however, the differences are not large. These statistical differences are not surprising given our sample size and given the fact that our sample consisted of patients whose symptoms and disability persisted for the 3-month duration of our investigation and remained severe enough to warrant a return visit.

The concept of responsiveness has proven challenging. Beaton *et al*<sup>3</sup> have suggested that the responsiveness of an instrument should be viewed within the context of “who” is being studied, “which” scores are being contrasted, and “what” type of change is being assessed. The current study evaluates within-person clinically esti-

mated change over time per Beaton and colleague’s proposed taxonomy.

The issue of condition-specific *versus* general measures has been studied previously by Stucki *et al*.<sup>24</sup> The authors utilized the Roland-Morris Questionnaire (RMQ) as the condition-specific tool and The Sickness Impact Profile (SIP) as the general tool in an observational cohort study of surgical outcomes for degenerative spinal stenosis. Responsiveness was calculated using the standardized response mean, the standardized effect size, and the Guyatt responsiveness statistic. All measures were found to be responsive, the RMQ slightly more than the SIP. This is not surprising, as the RMQ was culled from the SIP and created specifically for back pain patients.

In 2001, Garratt *et al*<sup>13</sup> compared the condition-specific Aberdeen Back Pain Scale and the Roland Disability Questionnaire to the general utility scale, the EuroQol. They concluded that the condition-specific tools were more responsive than the EuroQol. However, as Lurie<sup>21</sup> points out in the following commentary, the authors’ methodology finds the condition-specific surveys to show “improvement” in the scores even in patients who rated themselves as deteriorating. The external criterion defining improvement used in this line of research is crucial. Not all clinically defined changes, be they physiologic or psychometric, are important changes to patients. We dealt with this dilemma by using patient-provider consensus as the gold standard for whether improvement had taken place.

In both the Stucki and Garratt articles, the condition-specific measures (RMQ and Aberdeen) were found to be more responsive than the general (SIP and EuroQol) measures they were compared to. Both sets of authors compared different measures than were used for the current study. Other possible reasons why the SIP and EuroQol may not have performed as well as the RMQ and Aberdeen may be if the RMQ and/or Aberdeen were more responsive than the ODI, or if the SIP and EuroQol were less responsive than the SF-36. Prior work can help shed light on these two possibilities.

The issue of which condition-specific survey is “best” has been studied. The two most common and well-studied instruments are the RMQ and the Oswestry. In 1997, Leclaire *et al*<sup>20</sup> compared the ODI and the Roland

Morris disability scale. The authors concluded that both scales measure the same element: functional ability. Although both were able to identify differing levels of clinical severity in the patient samples used (symptoms of low back pain only *vs.* sciatic symptoms), the ODI appeared more sensitive in the severely disabled patients.

Beurskens *et al*<sup>5</sup> compared the ODI and RMQ using ES calculations and the ROC method with a patient rating of improvement as their external criterion. The RMQ was superior to the ODI in both methods (effect sizes of 2.02 and 0.80 and ROC of 0.93 and 0.76, respectively). The baseline RMQ and ODI scores suggest the patients' function was not severely limited.

These analyses reveal that of the two disease-specific measures, the RMQ may be more suitable for patients with less limitation in function and the ODI may be more suitable for patients with greater limitation. Fanuele *et al* have shown that the average physical composite summary score (PCS) for patients in the NSN is roughly two standard deviations below the population norm.<sup>11</sup> Fanuele *et al* conclude that the real physical toll of conditions producing LBP-related symptoms is, perhaps, underestimated. In addition, it is possible that clinics proclaiming to be a "spine clinic" may have patient populations more severely compromised by their complaints. For these reasons, it can be argued that the condition-specific measure most sensitive to change in patients with low function should be used, and the NSN has chosen to utilize the ODI over the RMQ.

Thus, our findings—that several of the SF-36 scales are as responsive as the ODI and MODEMS scales—do not appear to result from having a poor condition-specific measure; indeed previous work suggests the ODI would be the best measure given our patient population. The other possibility as to why previous work has shown condition-specific measures to be more responsive is if the general measures used were less responsive than the SF-36.

Prior work reporting the superiority of condition-specific measures to general measures had not utilized the SF-36 until an attempt to assess the responsiveness of the RMQ compared to the SF-36 was undertaken by Patrick *et al*.<sup>23</sup> The RMQ and the BP and PF scales of the SF-36 all showed improvement over the 3-month period, the BP scale and RMQ demonstrating the most improvement.

In 1999, the responsiveness of the ODI and SF-36 scales was evaluated in work from the UK.<sup>26</sup> The authors assessed responsiveness using the three methods of calculating effect size. The results are displayed graphically but the exact ES scores are not reported. They found the rank order of effect sizes to vary depending on the method of calculation. The authors found that the majority of the SF-36 scales were responsive, the physical function, social function and bodily pain scales being most consistent, and the other five scales and two composite scores, less so. The ODI was also responsive. However, the sensitivity to change of the ODI was always matched or improved by one or more of the SF-36 scales. The authors suggest this implies the health concepts

in the general questionnaire are "indeed important to patients with low back pain." Our interpretation is that the ODI never "outperforms" some facets of the SF-36. Our own data reveal that by using the general measure alone, it is possible to gain the added value of using a general measure without sacrificing responsiveness.

### Limitations

Several limitations of our findings must be recognized. Our results are limited by the gold standard utilized in our analyses. Truthfully, there is no gold standard for measuring change. However, if both the patient and the provider agree that a change has occurred, and agree on the direction of change, it would seem difficult to argue that nothing has happened. In view of the fact that no truly objective gold standard yet exists, the consensus model first proposed by Deyo<sup>7</sup> would seem to be the best available measure. And notably, this method produced very similar results to those achieved using three common methods of calculating effect size.

Our consensus model is dependent on data from both the patient and the provider. Frequently one of those was missing. Either the provider forgot to enter the data (2.3% of the time) or the patient failed to complete that portion of the follow-up survey (62% of the time). In either case, we utilized the single available data point to classify the patient as improved, unchanged, or worsened. To assess what effect this missing data might have on our analysis and conclusions, we recalculated using only patients for whom both the patient and the clinician ratings were available. Our results were largely unchanged. Therefore, we separately performed and compared our analyses using the patient-only reported improvement and, for comparison, the clinician-only reported improvement as our external criterion. The resulting ROC areas were within 0.02 points of the ROC areas generated using patient-clinician consensus as the gold standard. The same rank order of scales was observed, with pain scales being the most responsive, pain-function scales somewhat less responsive, and function scales generally the least responsive. Differences between ROC areas for the various scales were small, ranging from 0.00 to 0.05. In no case were the results appreciably or significantly different from the results presented.

Also, in regards to our consensus model, the patient and clinician questions used to assess improvement or worsening were somewhat generally defined. The clinician question was quite general, asking only for a rating of the patient progress, leaving the specificity of the patient's spine problem to be implied by the clinical setting. The patient question did not specify a spine problem but did refer to the patient's specific musculoskeletal condition as the frame of reference for rating improvement. It is unclear to what extent the lack of specific reference to the spine condition in these questions contributes to the equivalent responsiveness between the disease specific and general health status measures in this study. It is important to remember, however, that only the BP, PCS,

and PF scales of the SF-36 showed good responsiveness in this study; the GH, MH, RP, and RE scales were much less responsive. This finding argues that a specific improvement in pain and physical function was indeed being measured and not a general improvement in overall well being.

Generalizing these findings beyond the issue of back pain/leg symptoms is probably unwise. Condition-specific measures utilized in other orthopedic conditions such as knee, hip, or shoulder pains may indeed provide additional information inaccessible with current general measures. It may be that low back pain/leg symptoms are so central to a patient's function and sense of well being that the SF-36 pain and function scales can sufficiently detect changes in patients with these complaints as well as their overall health. In fact, the pain and function questions imbedded in the SF-36 are very well suited for assessing patients with low back pain/leg symptoms. Other types of orthopedic complaints may not have the same dramatic impact on overall function. In addition, because the wording of the SF-36's physical questions is biased toward lower extremity function, the findings of this study may not be replicable in patients with cervical spine/arm-related symptoms.

### ■ Summary

We utilized ROC analyses and ES calculations to examine the responsiveness of the Oswestry, MODEMS and the SF-36 in patients with low back pain/leg symptoms. We found the measures assessing pain were more responsive than those assessing function. The condition-specific measures were not superior to their general health counterparts. These findings were consistent among patients with multiple comorbidities and patients with low baseline function. Our findings are consistent with prior work in which the SF-36 has performed similarly to or better than the ODI in patients with low back pain syndromes. However, these findings differ from prior work that showed condition-specific measures to be more responsive than their general counterparts. We conclude that using the SF-36 without a condition-specific supplement could lessen responder time and reduce the burdens of data collection and analysis.

### ■ Key Points

- For studies of patients with low back pain/leg symptoms, the SF-36 may be a sufficient measure of health status and patient function, without the need for an additional condition-specific instrument.
- Self report of pain is statistically more responsive than self report of function for patients with low back pain/leg symptoms.
- The ROC method of assessing responsiveness, first proposed by Deyo, appears to function similarly to effect size calculations in the assessment of responsiveness.

### Acknowledgments

The members of the National Spine Network are: Alegent Health System/Nebraska Spine Surgeons, Omaha, NE; Alliant Health System/Kenton D. Leatherman Spine Center, Louisville, KY; Dartmouth Hitchcock Medical Center, Hanover, NH; Emory HEALTHCARE/Emory Spine Center, Atlanta, GA; Georgetown University Medical Center, Washington, DC; Hospital for Joint Diseases Spine Institute, New York, NY; Lakewood Orthopaedics, Denver, CO; Medical College of Wisconsin, Milwaukee, WI; Medical University of South Carolina, Charleston, SC; New England Baptist Hospital, Boston, MA; North Carolina Spine Center, Chapel Hill, NC; North Kansas City Hospital/Spine and Scoliosis Surgery, Ltd., Kansas City, MO; Rush-Presbyterian/St. Luke's Medical Center, Chicago, IL; University Hospital Spine Institute, Cleveland, OH; University of Iowa Hospitals and Clinics, Iowa City, IA; University of Missouri Health Sciences Center, Columbia, MO; The Hospital for Special Surgery, New York, NY; Thomas Jefferson University Hospital/Rothman Institute, Philadelphia, PA; University of California, San Francisco, San Francisco, CA; University of Pittsburgh Medical Center/University Orthopaedics, Inc., Pittsburgh, PA; University of Utah School of Medicine, Salt Lake City, UT; Washington University School of Medicine; Barnes-Jewish and St. Louis Children's Hospitals, St. Louis, MO; Vanderbilt University Spine Center, Nashville, TN; SUNY Health Science Center, Syracuse, NY; Tulane University Medical Center, New Orleans, LA; William Beaumont Hospital, Detroit, MI.

### References

1. Baker DJ, Pynsent PB, Fairbanks JCT. The Oswestry disability index revisited: its reliability, repeatability and validity, and a comparison with the Thomas's disability index. In: Roland MO, Jenner JR, eds. *Back Pain: New Approaches to Rehabilitation and Education*. Manchester, England: Manchester University Press; 1989:174-86.
2. Beaton D. Understanding the relevance of measured change through studies of responsiveness. *Spine* 2000;25:3192-9.
3. Beaton D, Bombardier C, Katz J, et al. A taxonomy for responsiveness. *J Clin Epidemiol* 2001;54:1204-17.
4. Beaton D, 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.
5. Beurskens A, de Vet H, Koke A. Responsiveness of functional status in low back pain: a comparison of different instruments. *Pain* 1996;65:71-6.
6. Brazier J, Harper R, Jones N, et al. Validating the SF-36 health survey questionnaire: new outcome measure for primary care. *Br Med J* 1992;305:160-4.
7. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. *J Chron Dis* 1986;39:897-906.
8. Deyo RA, Weinstein JN. Outcomes research for spinal disorders. In: Herkowitz HN, Balderston GS, Eismont FJ, et al., eds. *Rothman-Simeone: The Spine*. Philadelphia: W.B. Saunders; 1999:173-83.
9. Fairbank J, Pynsent P. The Oswestry Disability Index. *Spine* 2000;25:2940-52.
10. Fairbank JCT, Davies JB, Couper J, et al. The Oswestry low back pain and disability questionnaire. *Physiotherapy* 1980;66:271-3.
11. Fanuele J, Birkmeyer N, Abdu W, et al. The impact of spinal problems on the health status of patients: have we underestimated the effect? *Spine* 2000;25:1509-14.
12. Fisher K, Johnston M. Validation of the Oswestry Low Back Pain Disability questionnaire, its sensitivity as a measure of change following treatment and

- its relationship with other aspects of the chronic pain experience. *Physiother Theory Prac* 1997;13:667–80.
13. Garratt A, Moffett J, Farrin A. Responsiveness of generic and specific measures of health outcome in low back pain. *Spine* 2001;26:71–7.
  14. Garratt AM, Ruta DA, Abdalla MI, et al. The SF-36 health survey questionnaire: an outcome measure suitable for routine use within the NHS? *Br Med J* 1993;306:1440–4.
  15. Grevitt M, Khazim R, Webb J, et al. The short-form 36 health survey questionnaire in spine surgery. *J Bone Joint Surg Br* 1997;79:48–52.
  16. Guyatt GH, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. *J Chron Dis* 1987;40:171–8.
  17. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143:29–36.
  18. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 1983;148:839–43.
  19. Jenkinson C, Coulter A, Wright L. Short-form 36 (SF-36) health survey questionnaire: normative data for adults of working age. *Br Med J* 1993;306:1437–40.
  20. Leclaire R, Blier F, Fortin L, et al. A cross-sectional study comparing Oswestry and Roland-Morris functional disability scales in two populations of patients with low back pain of differing levels of severity. *Spine* 1997;22:68–71.
  21. Lurie J. Point of view: responsiveness of generic and specific measures of health outcome in low back pain. *Spine* 2001;26:77.
  22. McHorney CA, Ware JE Jr, Raczek AE. The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. *Med Care* 1993;31:247–63.
  23. Patrick DL, Deyo RA, Atlas SJ, et al. Assessing health-related quality of life in patients with sciatica. *Spine* 1995;20:1899–1909.
  24. Stucki G, Liang M, Fossel A, et al. Relative responsiveness of condition-specific and generic health status measures in degenerative lumbar spinal stenosis. *J Clin Epidemiol* 1995;48:1369–78.
  25. American Academy of Orthopaedic Surgeons. Outcomes data collection instruments, version 2.0. Rosemont, IL: American Academy of Orthopaedic Surgeons; 1997.
  26. Taylor S, Taylor A, Foy M, et al. Responsiveness of common outcome measures for patients with low back pain. *Spine* 1999;24:1805–12.
  27. Ware J, Snow K, Kosinski M, et al. SF-36 Health Survey: Manual and Interpretation Guide. Boston, MA: The Health Institute; 1993.
  28. Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care* 1992;30:473–83.