J Neurosurg: Spine 2:639–646, 2005

Guidelines for the performance of fusion procedures for degenerative disease of the lumbar spine. Part 2: assessment of functional outcome

Daniel K. Resnick, M.D., Tanvir F. Choudhri, M.D., Andrew T. Dailey, M.D., Michael W. Groff, M.D., Larry Khoo, M.D., Paul G. Matz, M.D., Praveen Mummaneni, M.D., William C. Watters III, M.D., Jeffrey Wang, M.D., Beverly C. Walters, M.D., M.P.H., and Mark N. Hadley, M.D.

Department of Neurosurgery, University of Wisconsin, Madison, Wisconsin; Department of Neurosurgery, Mount Sinai Medical School, New York, New York; Department of Neurosurgery, University of Washington, Seattle, Washington; Department of Neurosurgery, Indiana University, Indianapolis, Indiana; Departments of Orthopedic Surgery and Neurosurgery, University of California at Los Angeles, California; Department of Neurosurgery, University of Alabama at Birmingham, Alabama; Department of Neurosurgery, Emory University, Atlanta, Georgia; Bone and Joint Clinic of Houston, Texas; and Department of Neurosurgery, Brown University, Providence, Rhode Island

Key Words • **fusion** • **lumbar spine** • **practice guidelines** • **treatment outcome**

Recommendations

Standards. It is recommended that functional outcome be measured in patients treated for low-back pain due to degenerative disease of the lumbar spine by using reliable, valid, and responsive scales. Examples of these scales in the low-back pain population include the following: The Spinal Stenosis Survey of Stucki, Waddell–Main Questionnaire, RMDQ, DPQ, QPDS, SIP, Million Scale, LBPR Scale, ODI, the Short Form–12, the JOA system, the CBSQ, and the North American Spine Society Lumbar Spine Outcome Assessment Instrument.

Guidelines. There is insufficient evidence to recommend a guideline for assessment of functional outcome following fusion for lumbar degenerative disease.

Options. Patient satisfaction scales are recommended for use as outcome measures in retrospective case series, where better alternatives are not available. Patient satisfaction scales are not reliable for the assessment of outcome following intervention for low-back pain.

Abbreviations used in this paper: CBSQ = Curtain Back Screening Questionnaire; DPQ = Dallas Pain Questionnaire; DRI = Disability Rating Index; FSQ = Functional Status Questionnaire; JOA = Japanese Orthopaedic Association; LBPR = Low Back Pain Rating; ODI = Oswestry Disability Index; QOL = quality of life; QPDS = Quebec Pain Disability Scale; RMDQ = Roland–Morris Disability Questionnaire; SF-36 = Short-Form–36; SIP = Sickness Impact Profile; VAS = visual analog scale.

Rationale

Lumbar spinal fusion is an increasingly common procedure performed as an adjunct in the surgical management of patients with degenerative lumbar disease and instability. As the frequency and complexity of lumbar fusion surgery increases, there is a tendency for costs and complication rates to increase as well.[20] With fewer hospital resources available, the ability to assess objectively the functional outcome following lumbar fusion and to correlate patient outcome with the economic consequences of treatment is important.

Various assessment tools are available for measuring functional outcomes in patients who have undergone lumbar fusion. These outcomes may vary widely in the same population depending on whether subjective or objective measures have been used.[17] Examples of objective outcome measures include physiological, anatomical, economic, health-related QOL, and mortality measurements.[10] Objective outcome measures may be classified into functional questionnaires, global ratings (satisfaction), economic factors (employment, disability, and cost), and physical factors (activities).[21] The purpose of this review was to identify valid, reliable, and responsive measures of functional outcomes after lumbar fusion for degenerative disease.

Search Criteria

A computerized search of the National Library of Medicine database of the literature published between 1966

and 2003 was performed. A search using the subject heading "lumbar fusion" yielded 3708 citations. The following subject headings were combined: "lumbar fusion and outcomes." Approximately 204 citations were acquired. Only citations in English were selected. A search of this set of publications with the key words "functional outcome" and "satisfaction" resulted in 107 matches. Alternative searches included each disability index by name. Titles and abstracts of the articles were reviewed and clinical series dealing with adult patients treated with lumbar fusion for degenerative lumbar disease were selected for detailed analysis. Additional references were culled from the reference lists of remaining articles. Among the articles reviewed, 30 studies were included that dealt with lumbar fusion, functional outcomes, and satisfaction surveys. Nineteen of these articles were studies in which the authors examined the reliability of functional outcome measures. In another seven articles investigators examined the utility of these functional outcome measures in the setting of lumbar fusion. Two articles were overviews on functional outcome and lumbar degenerative disease. All papers providing Class I medical evidence are summarized in the evidentiary table (Table 1).

Scientific Foundation

Assessment of Functional Outcome

To assess outcome following treatment properly, a functional instrument must fulfill three criteria.[11,21] First, it must be reliable.[10,11] Repetition of the functional assessment should be consistent within (internal reliability) and between (external reliability) observers. If a functional instrument contains multiple domains, each should correlate with the final outcome (internal consistency). Second, a functional instrument must be valid.[21] It should measure the property intended. For example, an instrument assessing dysfunction due to leg pain would be expected to correlate with a reduction in the ability to walk a given distance. Finally, the instrument should be responsive.[21] The instrument should be able to detect differences in severity among populations. If an instrument measures low-back pain and this pain improves with physical therapy, the instrument should reflect that improvement quantitatively. When evaluating the utility of a functional tool, the initial assessment should emphasize reliability. If a functional instrument does not produce reliable results, its validity and responsiveness are irrelevant.

In terms of grading the quality of outcomes instruments, κ and α values are used. The κ value refers to the degree of correlation of interrater observations (reliability). In patient-based assessments, it indicates consistency in response at a given time point. The α value, often calculated using the Cronbach α test, reflects the degree to which each domain of a multidomain outcome measure correlates with the final result.[7] For example, an assessment tool for pain may contain physical, psychological, and social domains. Each domain score should correlate with the final score. For a study to provide Class I medical evidence regarding functional outcomes, the outcomes tool used must have a κ value greater than 0.8. Class II medical evidence requires an outcomes tool to have a κ greater than 0.6. Any outcome scale with a κ value less than 0.6 is considered to provide Class III medical evidence for the assessment of outcomes following an intervention.[18]

Roland and Morris[30,31] followed 230 patients of whom 193 were studied up to 4 weeks after their initial presentation. Functional disability was assessed using a 24-item disability questionnaire (the RMDQ) with statements derived from the SIP and relating to the lower back. Reliability was ascertained in 20 patients with an external reliability greater than 0.91. Internal consistency appeared to be greater than 0.8. Validity was confirmed after comparisons to a six-point pain rating scale and physical signs ascertained by an examining physician.[31] In this group, 60% of patients appeared to improve over the 4-week period, whereas 20% worsened. Absence from work appeared to correlate less well with disability, as only 8% of the employed were unable to work.[30] Using the ODI, Fairbank and colleagues[12] followed 25 patients with acute low-back pain in whom a reasonable prognosis was expected. The questionnaire has 10 categories with six gradations each, for a total score of 50. It was completed at weekly intervals over a period of 3 weeks. Reliability ($\kappa > 0.95$) was confirmed in 22 patients who repeated the questionnaire over 2 days. Validity was demonstrated as patients improved over 3 weeks. Paired t-tests revealed a significant improvement in ODI scores during this time period ($p < 0.005$).

Leclaire and colleagues[24] observed patients who presented with acute low-back pain alone (100 cases) or accompanied by radiculopathy (100 cases). The cohort was followed using the RMDQ and ODI questionnaires. In the radiculopathy group, ODI and RMDQ scores were significantly more severe (higher) than in the low-back pain–alone group ($p < 0.0001$). The two scales had a moderate correlation to each other in each subgroup ($r = 0.72$ [radiculopathy]; $r = 0.66$ [lumbago]; $p < 0.0001$). In a cohort of patients with low-back pain, the JOA score was used as a psychometric measure. External reliability was strong ($\kappa > 0.90$) when 15 patients reassessed their status with no change in their symptomatology. Interobserver external reliability among physicians was also sound ($\kappa > 0.90$) in 30 patients reassessed using the JOA. Validity was established by a strong correlation to the RMDQ, ODI, and the SF-36.[15] In several different groups with lumbar degenerative disease, the North American Spine Society Lumbar Spine Outcome Assessment tool was used to assess patients who had undergone conservative or decompressive therapy.[8] In this study, 136 of 206 questionnaires were successfully completed. External reliability was assessed in 64 patients. Both internal and external reliability was strong ($\kappa > 0.90$). The test was determined to be a valid measure compared with existing instruments.

The SIP is a traditional general functional outcome measure, with 136 items in 12 categories, that has been evaluated in the general populace for a variety of conditions. It has been applied to patients with low-back pain and degenerative lumbar disease. Bergner, et al.,[1] examined the use of this general health instrument in 1108 patients with multiple medical problems including rheumatoid arthritis and hip osteoarthritis.[1] Simultaneous with this questionnaire were a clinician's assessment of physical function and patients' self-assessment of the severity of sickness and dysfunction. In this setting, the test–retest (external) reliability of SIP was greater than 0.90, and its

TABLE 1

*Evidentiary table summarizing published studies involving Class I medical data**

Authors & Year	Class	Description	Results	Conclusions
Fairbank, et al., 1980	I	25 patients w/ acute LBP & reasonable prognosis were studied at wkly intervals for 3 wks w/ a functional disability survey. The ODI has 10 categories each w/ 6 responses graded 0–5. A total of 50 points are possible.	Test–retest reliability was $\kappa > 0.95$ ($p < 0.001$) in 22 patients. Over the 3-wk interval, significant improvement was noted clinically & was detected using the ODI. A paired t-test revealed a significant improvement on the ODI over 3 wks ($p < 0.05$).	The ODI is a reliable & valid measure in detecting changes in the LBP & its functional severity.
Bergner, et al., 1981	I	1108 patients in a general populace w/ multiple problems including RA & hip osteoarthritis. Patients were evaluated using the SIP. Assessment was done by a clinician for physical measures. Self-assessment was completed for severity of sickness & dysfunction.	External reliability w/in & btwn observers was $\kappa > 0.90$. Internal consistency was $\alpha > 0.90$. Self-assessment of sickness & dysfunction had a reliability of $\kappa > 0.60$. The SIP appeared to correlate w/ the self-assessment of sickness & dysfunction (correlation > 0.50).	SIP measures independent function, physical wellness, & psychosocial wellness. It is reliable & valid. Reasonable measures to use for outcome are SIP & self-assessment of sickness & dysfunction.
Million, et al., 1982	I	19 patients w/ chronic LBP. Their functional disability was studied using the Million Scale which was a VAS examining 15 subjective variables reflecting the severity of lumbago. A soft corset w/ & w/o support was used to test the responsiveness of the Million Scale.	External reliability was strong btwn & w/in observers $\kappa > 0.90$. As a validity measure, the Million Scale appeared to reflect changes in physical measurements. At 4 & 8 wks after rigid bracing, patients improved clinically, & this responsiveness was detected by the Million Scale ($p < 0.05$ at 4 wks & $p < 0.01$ at 8 wks).	The Million Scale is a reliable indicator of the severity of lumbago & is responsive in the early phase of treatment. Its responsiveness appears better than that of objective measurements including lumbar motion & straight leg raising.
Roland & Morris, 1983	I	230 patients w/ acute lumbago; 193 were studied at 0, 1, & 4 wks after the episode. Test–test reliability was done on 20/230 patients. The construct validity was qualitatively assessed by comparing this functional questionnaire to the pain rating scale.	External reliability was $\kappa > 0.90$ & internal consistency $\alpha > 0.80$. Construct validity demonstrated that the Roland–Morris questionnaire was able to detect qualitatively patients w/ poorer outcomes from acute lumbago; however, no specific analysis was done.	The RMDQ is reliable for assessment of acute LBP.
Roland & Morris, 1983	I	230 patients w/ acute lumbago who were studied at 0, 1, & 4 wks. The disability questionnaire was administered & completed at all time intervals in 193 patients. Correlation was qualitatively done w/ back-to-work status.	$> 60\%$ of patients had improvement over the 4-wk period, whereas 20% had an increase in disability. These changes appeared to be reflected in the disability questionnaire. Absence from work appeared to correlate less well as only 8% of employed were unable to work 4 wks after acute lumbago.	No specific statistics tested the correlation in this study. The RMDQ is reliable but this manuscript did not assess its responsiveness to a standard measure in statistical fashion.
Waddell & Main, 1984	I	160 patients w/ 12 wks of lumbago (chronic) w/ severity studied by a 9-category disability index & physical characteristics. Reliability determined using a subgroup of 30 patients.	Disability as determined by functional outcome on questionnaire had a reliability > 0.80 & correlated w/ the ODI ($r = 0.70$). For physical characteristics (lumbar flexion, straight leg raising, root compression signs) reliability was > 0.90.	Waddell Scale describes functional disability w/ chronic LBP. All 9 scales correlate w/ final score (content validity) & the scale is reliable. It also has construct validity as it correlates w/ ODI.
Deyo, 1986	I	136 patients who were examined in a clinic for a chief complaint of lumbago. Evaluation was done using SIP & the modified RMDQ Scale (shortened version of SIP) initially & 3 wks later.	Reliability for both scales was $\kappa > 0.80$ in patients (10) who had no change in pain. For patients who did not resume full activity (47), the reliability was $\alpha > 0.60$. A strong correlation existed btwn the scales ($r = 0.85$) & between the physical dimension of the SIP & the modified RMDQ ($r = 0.89$). The modified RMDQ correlated less well w/ the psychosocial dimension of the SIP ($r = 0.56$).	The SIP & the modified RMDQ (shorter) are reliable scales for the assessment of lumbago, which seem to follow the physical dimension of functional disability. The modified RMDQ is less well suited to follow the psychosocial dimension of functional disability.
Lawlis, et al., 1989	I	143 patients overall (24 normal, 15 chronic lumbago but working, 104 chronic lumbago undergoing inpatient therapy). Functional assessment performed using the DPQ which assesses daily activities, work & leisure activities, anxiety/depression, & social interest. Reliability tested on 15 chronic pain patients & 13 normal patients.	External reliability was $\kappa > 0.90$. Construct validity was shown by correlation of the 1st 2 categories of DPQ w/ functional capacity scores relating to the physical demands of work. Responsiveness was assessed by comparing DPQ scores in the 104 chronic lumbago patients to the 24 normal patients. DPQ scores were significantly higher in the former.	The DPQ is a reliable test in assessing chronic LBP & appears responsive in defining differences btwn patients w/ chronic lumbago & those w/o.

continued

TABLE 1 *Continued*

Authors & Year	Class	Description	Results	Conclusions
Manniche. et al.. 1994	I	58 patients who underwent lumbar disc op were surveyed 14–60 mos postop. The assessment was an LBPR scale that examined physical impairment. disability. & pain intensity. Comparison was done against a doctor's global assessment & a patient's global assessment.	The LBPR scale comprised 60 points for pain. 30 for level of function. & 40 for physical impairment. Interrater reliability was $\kappa > 0.95$. Using contingency tables. the scale correlated with the doctor's assessment and patient's assessment ($p < 0.00005$).	The LBPR Scale combines elements of physical function. pain intensity. & overall disability. It is a reliable indicator of dysfunction & appears valid compared w/ objective measures (doctor's assessment) & subjective/satisfaction measures (patient's assessment).
Ruta. et al.. 1994	I	354 patients w/ lumbago initially examined in clinic & surveyed shortly thereafter to assess functional disability. 273 patients were retested for reliability of whom 183 reported no change in clinical severity. Correlation to the SF-36 general health profile was done for construct validity.	183 patients had no clinical changes & underwent external reliability testing ($\kappa > 0.90$). The questionnaire correlated well w/ all 8 domains of the SF-36 using linear regression ($p < 0.001$) & w/ perceptions of disease severity.	This LBP scale is a reliable & valid indicator of the functional disability relating to lumbago. No usage described in the setting of lumbar fusion. No acuity given for the lumbago.
Salen. et al.. 1994	I	1445 patients were divided into 3 groups: 1092 volunteer controls. 306 w/ axial skeletal pain. & 47 w/ joint pain. Patients were evaluated using the DRI & an FSQ.	External reliability for the DRI was $\kappa > 0.80$. There was a correlation to the FSQ. The DRI was responsive in detecting improvement after joint replacement.	The DRI is a reliable. valid. & responsive measure in patients w/ axial skeletal pain.
Harper. et al.. 1995	I	150 patients were divided into 3 groups (Group I: chronic lumbago >4 wks/disabled; Group II: acute lumbago/working; Group III: normal). Evaluation of functional disability was done using the CBSQ & the SIP. The CBSQ tests 11 categories of functional disability. Test–retest correlation & correlation btwn tCBSQ and SIP was done using the Pearson correlation test.	External reliability in all 3 groups was $\kappa > 0.90$. Internal reliability was $\alpha > 0.80$. There was a strong correlation btwn each category in CBSQ & its similar category in the SIP ($r = 0.56$–0.72). Finally. CBSQ scores appeared responsive w/ higher scores in the more severely affected groups.	The CBSQ is a reliable & valid measure for determining the functional disability associated with LBP. No testing of responsiveness was undertaken.
Kopec. et al.. 1995	I	242 patients with a history of lumbago in Quebec. 80% had prior lumbago w/ 29% receiving compensation. Patients were assessed for functional disability using the QPDS. Reliability was examined in a 98-patient sample w/in 1–14 days after initial survey. Construct validity was done by comparing results to functional scales of ODI. RMDQ. & SF-36.	External reliability was $\kappa > 0.90$ w/ internal consistency of $\alpha > 0.90$. Construct validity was shown by a strong correlation in this functional index w/ the ODI ($r = 0.80$). RMDQ ($r = 0.77$). & SF-36 ($r = 0.72$).	The QPDS is suitable for the reliable functional measurement of LBP.
Daltroy. et al.. 1996	I	206 patients in 6 orthopedic practices were evaluated. Patients were in several categories including those w/ LBP & sciatica. Also included were patients who underwent lumbar decompression but not fusion.	External & internal reliability were strong ($\kappa > 0.90$) when assessed in 64 patients. The measure was valid compared w/ known instruments.	The NASS LSOA is a valid & reliable outcome measure for functional evaluation of the lumbar spine.
Stucki. et al.. 1996	I	193 patients w/ lumbar degenerative stenosis undergoing decompression. Prospective multicenter study of self-administered outcome measure assessed w/in 6 mos. Likert response scales used in domains of physical dysfunction. symptom severity. & satisfaction. Results compared w/ SIP & VAS.	23/193 studied for reliability w/ $\kappa > 0.80$. Internal consistency $\alpha > 0.80$; 130/193 studied for responsiveness. Responsive & valid over 6 mos to detect improvement postop.	This outcome questionnaire was reliable in lumbar stenosis patients who underwent op & had construct validity compared w/ established scale & was responsive in detection of differences w/in 6 mos for functional improvement.
Fujiwara. et al.. 2003	I	97 patients observed clinically w/ LBP & followed using JOA. ODI. and RMDQ. Correlation was calculated btwn these measures & external reliability was assessed by repeated physician & patient observation.	Test–retest reliability was $\kappa > 0.90$ when patients (15) or physicians (30) did repeat measurements. Strong correlation was observed btwn JOA & ODI & RMDQ.	The JOA is a reliable & valid indicator of LBP.
Luo. et al.. 2003	I	2520 patients w/ LBP; 506 patients assessed over 3–6 mos. SF-12 survey was used & compared w/ subjective quantification of LBP intensity.	External reliability of the SF-12 was performed by Ware. et al.. in a different patient group; however. internal reliability & responsiveness was found in this study.	The SF-12 is capable of assessing & following LBP reliably.

* LBP = low-back pain; NASS LSOA = North American Spine Society Lumbar Spine Outcome Assessment; RA = rheumatoid arthritis.

internal consistency was greater than 0.90. Self-assessment of sickness and dysfunction had a reliability greater than 0.60. The SIP appeared to correlate (> 0.50) with the self-assessment of sickness and dysfunction. Deyo[9] used the SIP and a modified RMDQ when evaluating 136 patients with a chief complaint of low-back pain at an initial index visit and 3 weeks later. Reliability was examined in 10 patients who claimed no interval improvement in pain and in 47 patients who did not resume full activity. For patients with no change in pain, the correlation was greater than 0.80. In those patients who may have improved but did not resume normal activity, reliability was greater than 0.60. A strong correlation was observed between the SIP and the modified RMDQ (r = 0.85). The physical dimension of the SIP (r = 0.89) correlated more strongly with the RMDQ than the psychosocial dimension (r = 0.56). The SIP appears to be a reliable and valid measure of the severity of low-back pain in the acute phase.

Million and colleagues[27] assessed 19 patients with chronic low-back pain by using a VAS examining 15 subjective variables reflecting its severity. External reliability among and within observers was greater than 0.90. To determine validity, they compared their results with physical measurements of spinal movements and straight leg raising. These objective assessments had a reliability greater than 0.90 and correlated with the Million Scale. After bracing with a rigid support, low-back pain improved clinically and this responsiveness was detected by the Million Scale. The Waddell–Main Disability Index was used to evaluate chronic low-back pain (duration > 12 weeks) in a 160-patient cohort.[37] Reliability in this study was evaluated in a random subgroup of 30 patients. Measures were also obtained of objective physical characteristics including lumbar flexion, straight leg raising, and root compression signs. The external reliability on the Waddell–Main Disability Index was greater than 0.80, and its validity was established by a strong correlation with the ODI (r = 0.70). The physical characteristics, when evaluated for objective reliability, had a correlation greater than 0.80.

Using the DPQ, Lawlis and colleagues[23] studied 143 patients of whom 119 had chronic low-back pain. Fifteen patients in this group were working, whereas the remaining 104 were undergoing inpatient therapy. Twenty-four healthy volunteers served as controls. The DPQ was used to assess daily activities, work/leisure activities, anxiety/depression, and social interest. Reliability was tested in 15 patients with chronic back pain and 13 controls. External reliability was greater than 0.90. Construct validity was shown through a positive correlation to other assessments of functional capacity relating to the physical demands of work. The DPQ was responsive to differences between patients with chronic low-back pain and controls.

Ruta, et al.,[32] devised an outcome measure based on questions commonly used in the clinical assessment of patients with low-back pain. A total of 354 patients with low-back pain seen by primary and specialty practitioners were studied. Within this group, 273 patients were tested for reliability. One hundred eighty-three reported no clinical changes over a 2-week interval. External reliability was tested in these 183 patients with correlations greater than 0.90. Validity was demonstrated by a strong correlation ($p < 0.001$ on regression) with the SF-36 general health assessment. Harper and colleagues[19] examined 150 patients in three subgroups (chronic low-back pain [50 cases], acute lumbago [49 cases], and control [51 cases]). They employed the CBSQ, which evaluated 11 categories of functional disability and compared results with those of the SIP. External reliability for the CBSQ was greater than 0.90, with internal reliability greater than 0.80. A strong correlation was observed between each category in the CBSQ and its similar category in the SIP (r = 0.56–72), and the CBSQ appeared responsive in distinguishing the severity of dysfunction among the three groups of patients.

Several other groups undertook studies on the functional assessment of chronic low-back pain. Using the QPDS, Kopec, et al.,[22] analyzed 242 patients with a history of chronic low-back pain. Twenty-nine percent of this group were disabled and receiving compensation. This scale contains 48 items assessing the difficulty in simple daily activities pertaining to domains relevant to low-back pain. Reliability was gauged using a random sample (98 cases) who were retested after 14 days. External reliability was greater than 0.90, with an internal consistency coefficient greater than 0.90. Construct validity was determined by a strong correlation with the ODI (r = 0.80), RMDQ (r = 0.77), and SF-36 (r = 0.72) Scales. Using the LBPR scale, Manniche and colleagues[26] surveyed 58 patients 14 to 60 months after they underwent lumbar disc surgery. This scale comprises 60 points for back and leg pain, 30 points for level of function, and 40 points for physical impairment. External reliability had a coefficient greater than 0.95. Validity was determined by dichotomizing the scale into good and bad outcomes. The mean score of the study population was 39, and therefore a value greater than 39 implied greater dysfunction than the mean. The results on the LBPR Scale correlated ($p < 0.00005$) with a Global Assessment Scale (a graded evaluation tool) performed by both patient and physician.

Stucki, et al.,[35] evaluated 193 patients with degenerative lumbar stenosis from multiple centers who were to undergo lumbar decompression. A functional survey was undertaken preoperatively and 6 months after surgery. Interobserver reliability was studied in a random sample of 23 patients. Correlation (κ) was greater than 0.80 in this group. Internal consistency was greater than 0.80. This lumbar outcome scale was responsive to functional improvement in this cohort of patients when reassessed 6 months following surgery. Comparison to the SIP and the VAS for pain confirmed the validity of this instrument in detecting overall dysfunction associated with lumbar stenosis.

Bernstein and colleagues[2] followed 291 patients with chronic low-back pain by using the 90-item Symptom Checklist, which measures psychological dysfunction. It has nine major scales with one common factor—general psychological discomfort. The somatization scale covers general physical discomfort. The reliability of this checklist was not reported in this study, but validity was ascertained by comparison with the Minnesota Multiphasic Inventory and the McGill Pain Inventory. In this group of patients, the scale had a high correlation with the Minnesota Multiphasic Inventory and McGill Inventory scales for detecting general discomfort; however, external reliability was not reported. In a 5-year period, Greenough and Fraser[16] studied 300 patients with low-back pain by using a Low-Back Outcome Score that examined 13 functional factors related to pain. Comparison was made to the ODI

and Waddell–Main scales. Despite a statement that external reliability was studied, no mention was made of the statistical analysis in their study. This scale had a high correlation with the ODI (-0.87; $p < 0.001$) and Waddell–Main scale (-0.74; $p < 0.001$). Moffroid and colleagues[28] assessed 115 patients undergoing physical therapy referred for low-back pain, 112 asymptomatic volunteers were used as a control group. The physical capabilities of both groups were quantified using the National Institute for Occupational Safety and Health Low Back Atlas score. Although external reliability was described, it was not specifically reported in this study. The authors did find clusters of patients with imbalances in muscle strength and symmetry. Those patients were more apt to suffer from low-back pain.

General health may be measured in addition to low-back pain. In addition to the use of the SIP as a general health measure, Brazier and colleagues[3] studied the SF-36 Scale in 1582 patients in a general medical practice. The SF-36 Scale focuses on functional status, general wellness, and an overall assessment of health in eight domains by asking 36 questions. Results were compared for validity with the Nottingham Scale. In the general population, the external reliability coefficient was greater than 0.60. Construct validity was determined through a correlation with the Nottingham Scale ($r > 0.50$). Ware and colleagues[38] used regression methods to shorten the SF-36 to a 12-item format (SF-12) focusing on physical and mental aspects. Reliability in an initial evaluation of two different sets of patients was strong ($\kappa > 0.80$). Luo and colleagues[25] used the SF-12 in 2520 patients with low-back pain. Although no external reliability was performed in this setting, internal consistency was sound, and the SF-12 appeared valid and responsive to changes in patients with low-back pain.

Salen, et al.,[33] assessed 1092 healthy volunteers and compared observation with 306 patients with axial skeletal pain or 47 with joint pain by using a DRI. External reliability for this group was greater than 0.80. The DRI was valid with correlation to the FSQ. The DRI was responsive in detecting improvement after joint replacement.

Examples of the Application of Functional Assessments to Lumbar Fusion

The appropriateness of an outcome instrument designed to assess low-back pain does not necessarily generalize to the assessment of patients treated with lumbar spinal fusion procedures. Despite this fact, these same outcome measures have been used to assess outcome following lumbar fusion procedures. In an attempt to correct this apparent deficiency, many investigators have used multiple outcome instruments for correlation.

Several groups have used more formalized methods of assessing patient outcome. Moller and Hedlund[29] studied 111 patients with isthmic spondylolisthesis and a 1-year history of back or leg pain. Patients were randomized to surgery (80 cases) or exercise (34 cases). Evaluation was completed at 1 and 2 years by using the DRI and a patient assessment survey involving broad categories (much better, better, unchanged, or worse). In this patient population, the DRI appeared responsive with improvement in the surgical group at 12 and 24 months ($p < 0.0001$, Mann–Whitney U-test). Similarly, the broad patient assessment survey revealed that a higher proportion of "good" responses occurred in the surgery group ($p < 0.01$). In a similar cohort study, Christensen and colleagues[6] followed 129 patients with chronic low-back pain and either isthmic spondylolisthesis, primary lumbar degeneration, or secondary lumbar degeneration. Comparison was made between posterior fusion with and without instrumentation by using the DPQ and LBPR Scale in a 5-year period. Patients in both groups improved significantly from their preoperative status on the DPQ during this period. With the exception of patients with isthmic spondylolisthesis, no differences were observed between groups when using the DPQ or LBPR Scale. For patients with isthmic spondylolisthesis, fusion without instrumentation resulted in significantly better results as measured by the DPQ.

In a different cohort study, Fritzell and colleagues[14] studied 294 patients with L4–S1 disc degeneration and low-back pain who underwent surgical (222) or expectant (72) management during a 6-year period. Evaluation was completed at 6, 12, and 24 months by using the ODI, Million, and General Function Score Scales. Disability significantly decreased in the surgical group over a 2-year period when assessed using all of these scales ($p < 0.02$). Using a general, subjective assessment, 63% in the surgical group indicated they were better or much better compared with 29% in the nonsurgical group ($p < 0.0001$). Burkus, et al.,[4] reported on 46 patients randomized to anterior interbody fusion with or without bone morphogenetic protein–2. Outcome was recorded over a 24-month period by using the ODI, SF-36, and satisfaction scales. Neurological function, satisfaction, and general health measures were no different between groups. The ODI score indicated an improvement in the bone morphogenetic protein–2 group as early as 3 months after surgery. These outcome measures were responsive to low-back pain after lumbar fusion, and the use of multiple outcome measures conferred apparent validity.

Other Outcome Measures

Turner and colleagues[36] undertook a metaanalysis of all lumbar fusion Medline literature published between 1966 and 1991. Studies were required to have a minimum 1-year follow-up period and classification of clinical outcome as satisfactory or unsatisfactory in at least 30 patients. Forty-seven articles met their inclusion criteria. No randomized trials were identified at that time. A mean of 68% of the patients had a satisfactory outcome (range 16–95%). Substratification revealed outcomes of excellent/good in 66% (range 16–93%), fair in 22% (range 5–68%), and poor in 13% (range 2–54%). No defined criteria were reported for external reliability. Their analysis demonstrates that outcomes may be dichotomized into broad categories to assess overall outcome following lumbar fusion.

Patient satisfaction has been used as an outcome measure for patients undergoing lumbar fusion. Patient satisfaction surveys are frequently used in the setting of retrospective series because preintervention data may not be available. Patient satisfaction is easily surveyed but is dependent on multiple external factors independent of the surgical procedure. Furthermore, satisfaction outcome measures are hampered by the inherent inability to measure

responsiveness. The validity of satisfaction measures has been examined but their external reliability has not.

Slosar and colleagues[34] followed 141 patients who underwent circumferential lumbar fusion. A satisfaction survey was used as a follow-up instrument, as was return to employment. Patients were asked if: 1) surgery met their expectations; 2) surgery improved their condition; 3) surgery improved their condition but they would not repeat it; and 4) surgery worsened their condition. One hundred thirty-three patients were followed for more than 37 months. The outcomes were classified as follows: 10.5% in Category 1, 51.1% in Category 2, 19.5% in Category 3, and 18.8% in Category 4. Christensen and colleagues[5] followed 148 patients who underwent posterior lumbar fusion with or without supplemental anterior interbody fusion. Satisfaction surveys and the DPQ and LBPR Scale were used. In addition to improvements on the LBPR Scale and DPQ, satisfaction was high in both groups, with 77% of patients in the posterior fusion group and 79% of patients in the circumferential fusion group stating they would undergo surgery again if indicated.

In a study of 388 Workers' Compensation patients in Washington state, Franklin and colleagues[13] undertook an assessment of broad satisfaction surveys. Simple surveys examined back and leg pain, QOL, and the decision to undergo surgery at 2 years following lumbar fusion. Patients were dichotomized into two outcome groups: poor (receiving Workers' Compensation) and good (not receiving Workers' Compensation) at 2 years. There was a higher incidence of poor outcomes among those who stated that back or leg pain was worse than expected (76% compared with 54%; $p < 0.0003$) and in those whose QOL was no better or worse than expected (69% compared with 34%; $p < 0.0001$). There was a lower incidence of poor outcomes in patients who would undergo surgery again for the same indications (52% compared with 80%; $p < 0.0001$).

Although patient satisfaction surveys are easy and are intuitively valuable, they have never been validated and the responsiveness of such measures cannot be measured. Furthermore, wide discrepancies exist when results of patient satisfaction surveys are compared with validated outcome measures. These inadequacies limit their ability to provide high-quality medical evidence for or against any treatment modality.

Summary

Functional disability secondary to acute low-back pain, chronic low-back pain, lumbar stenosis, and lumbar disc disease may be reliably and validly assessed using functional outcome surveys that are valid, reliable, and responsive. Outcome instruments supported by Class I and Class II medical evidence for the evaluation of low-back pain include the Spinal Stenosis Survey of Stucki, Waddell–Main, RMDQ, DPQ, QPDS, SIP, Million Scale, LBPR Scale, ODI, and CBSQ. Many of these outcome measures have been applied to patients who have been treated with lumbar fusion for degenerative lumbar disease and have proven to be valid and responsive; however, the reliability of these instruments has never been specifically assessed in the lumbar fusion patient population. Patient satisfaction surveys have been used to measure outcome following lumbar fusion. Their usefulness resides in their insight into patient attitudes toward the treatment experience but is limited because of their inability to measure responsiveness and the lack of information on their reliability.

Key Issues for Future Investigation

Although the functional outcome instruments discussed in this review appear valid and responsive in the low-back pain patient population, their external reliability has not been confirmed in the clinical setting of lumbar fusion. This may be important for the comparison of different lumbar fusion techniques. Another key issue appears to be the timing of administration of the outcomes instruments. The aforementioned functional outcome measures appear to be responsive both initially and over a few years. Whether the benefits associated with any sort of intervention for low-back pain are durable beyond this period has not been established.

References

1. Bergner M, Bobbitt RA, Carter WB, et al: The Sickness Impact Profile: development and final revision of a health status measure. **Med Care 19:**787–805, 1981
2. Bernstein IH, Jaremko ME, Hinkley BS: On the utility of the SCL-90-R with low-back pain patients. **Spine 19:**42–48, 1994
3. Brazier JE, Harper R, Jones NM, et al: Validating the SF-36 health survey questionnaire: new outcome measure for primary care. **BMJ 305:**160–164, 1992
4. Burkus JK, Transfeldt EE, Kitchel SH, et al: Clinical and radiographic outcomes of anterior lumbar interbody fusion using recombinant human bone morphogenetic protein-2. **Spine 27:** 2396–2408, 2002
5. Christensen FB, Hansen E, Eiskjaer SP, et al: Circumferential lumbar spinal fusion with Brantigan cage versus posterolateral fusion with titanium Cotrel-Dubousset instrumentation: a prospective, randomized clinical study of 146 patients. **Spine 27:** 2674–2683, 2002
6. Christensen FB, Hansen ES, Laursen M, et al: Long-term functional outcome of pedicle screw instrumentation as a support for posterolateral spinal fusion. **Spine 27:**1269–1277, 2002
7. Cronbach LJ: Coefficient alpha and the internal structure of tests. **Psychometrika 16:**297–334, 1951
8. Daltroy LH, Cats-Baril W, Katz JN, et al: The North American Spine Society Lumbar Spine Outcome Assessment Instrument: Reliability and Validity Tests. **Spine 21:**741–748, 1996
9. Deyo RA: Comparative validity of the sickness impact profile and shorter scales for functional assessment in low-back pain. **Spine 11:**951–954, 1986
10. Deyo RA, Andersson G, Bombardier C, et al: Outcome measures for studying patients with low back pain. **Spine 19 (18 Suppl):**S2032–S2036, 1994
11. Deyo RA, Diehr P, Patrick DL: Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. **Control Clin Trials 12 (4 Suppl):**S142–S158, 1991
12. Fairbank JC, Couper J, Davies JB, et al: The Oswestry low back pain disability questionnaire. **Physiotherapy 66:**271–273, 1980
13. Franklin GM, Haug J, Heyer NJ, et al: Outcome of lumbar fusion in Washington State workers' compensation. **Spine 19:** 1897–1904, 1994
14. Fritzell P, Hagg O, Wessberg P, et al: 2001 Volvo Award Winner in Clinical Studies: Lumbar fusion versus nonsurgical treatment for chronic low back pain: a multicenter randomized controlled trial from the Swedish Lumbar Spine Study Group. **Spine 26:**2521–2534, 2001
15. Fujiwara A, Kobayashi N, Saiki K, et al: Association of the Japanese Orthopaedic Association score with the Oswestry Dis-

ability Index, Roland-Morris Disability Questionnaire, and Short-Form 36. **Spine 28:**1601–1607, 2003
16. Greenough CG, Fraser RD: Assessment of outcome in patients with low-back pain. **Spine 17:**36–41, 1992
17. Greenough CG, Peterson MD, Hadlow S, et al: Instrumented posterolateral lumbar fusion. Results and comparison with anterior interbody fusion. **Spine 23:**479–486, 1998
18. Hadley MN, Walters BC, Grabb PA: Guidelines for the management of acute cervical spine and spinal cord injuries. **Neurosurgery 50 (Suppl):**S2–S6, 2002
19. Harper AC, Harper DA, Lambert LJ, et al: Development and validation of the Curtin Back Screening Questionnaire (CBSQ): a discriminative disability measure. **Pain 60:**73–81, 1995
20. Katz JN: Lumbar spinal fusion. Surgical rates, costs, and complications. **Spine 20 (24 Suppl):**S78S–S83, 1995
21. Kopec JA, Esdaile J: Functional disability scales for back pain. **Spine 20:**1943–1949, 1995
22. Kopec JA, Esdaile J, Abrahamowicz M, et al: The Quebec Back Pain Disability Scale. Measurement properties. **Spine 20:**341–352, 1995
23. Lawlis GF, Cuencas R, Selby D, et al: The development of the Dallas Pain Questionnaire. An assessment of the impact of spinal pain on behavior. **Spine 14:**511–516, 1989
24. Leclaire R, Blier F, Fortin L, et al: A cross-sectional study comparing the Oswestry and Roland-Morris Functional Disability scales in two populations of patients with low back pain of different levels of severity. **Spine 22:**68–71, 1997
25. Luo X, Lynn George M, Kakouras I, et al: Reliability, validity, and responsiveness of the short form 12-item survey (SF-12) in patients with back pain. **Spine 28:**1739–1745, 2003
26. Manniche C, Asmussen K, Lauritsen B, et al: Low Back Pain Rating scale: validation of a tool for assessment of low back pain. **Pain 57:**317–326, 1994
27. Million R, Hall W, Nilsen KH, et al: Assessment of the progress of the back-pain patient. 1981 Volvo Award in Clinical Sciences. **Spine 7:**204–208, 1982
28. Moffroid MT, Haugh LD, Henry SM, et al: Distinguishable groups of musculoskeletal low back pain patients and asymptomatic control subjects based on physical measures of the NIOSH Low Back Atlas. **Spine 19:**1350–1358, 1994 (Erratum in **Spine 19:**2137, 1994)
29. Moller H, Hedlund R: Surgery versus conservative management in adult isthmic spondylolisthesis—a prospective, randomized study: part 1. **Spine 25:**1711–1715, 2000
30. 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 8:**141–144, 1983
31. Roland M, Morris R: A study of the natural history of low-back pain. Part II: development of guidelines for trials of treatment in primary care. **Spine 8:**145–150, 1983
32. Ruta DA, Garratt AM, Wardlaw D, Russell IT: Developing a valid and reliable measure of health outcome for patients with low back pain. **Spine 19:**1887–1896, 1994
33. Salen BA, Spangfort EV, Nygren AL, et al: The Disability Rating Index: an instrument for the assessment of disability in clinical settings. **J Clin Epidemiol 47:**1423–1435, 1994
34. Slosar PJ, Reynolds JB, Schofferman J, et al: Patient satisfaction after circumferential lumbar fusion. **Spine 25:**722–726, 2000
35. Stucki G, Daltroy L, Liang MH, et al: Measurement properties of a self-administered outcome measure in lumbar spinal stenosis. **Spine 21:**796–803, 1996
36. Turner JA, Ersek M, Herron L, et al: Patient outcomes after lumbar spinal fusions. **JAMA 268:**907–911, 1992
37. Waddell G, Main CJ: Assessment of severity in low-back disorders. **Spine 9:**204–208, 1984
38. Ware JE, Kosinski M, Keller SD: A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. **Med Care 34:**220–233, 1996

Manuscript received December 7, 2004.
Accepted in final form March 22, 2005.
Address reprint requests to: Daniel K. Resnick, M.D., Department of Neurological Surgery, University of Wisconsin Medical School, K4/834 Clinical Science Center, 600 Highland Avenue, Madison, Wisconsin 53792. email: Resnick@neurosurg.wisc.edu.