The Use of the Percentage Change in Oswestry Disability Index Score As an Outcome Measure in Lumbar Spinal Surgery

David G. Little, MBBS, and Donald MacDonald, MB, ChB, FRCS, FRACS

Study Design. A retrospective analysis of the change in the Oswestry Low Back Pain Disability Questionnaire in a heterogeneous group of 144 operatively managed patients was undertaken to examine the change in disability index as an outcome measure.

Objectives. To establish the percent change in disability index as an outcome measure able to identify risk factors for poor results in lumbar spinal surgery.

Methods. Epidemiologic, diagnostic, and surgical variables were examined as risk factors using step-wise multiple linear regression analysis at both follow-up times (6 months and 2 years), with percent change in disability index used as the outcome measure.

Results. At 6-months follow-up, previous surgery, female gender workers' compensation, a lower initial disability index score, increasing age, and spinal fusion alone as an operative procedure were independently and significantly negatively correlated with outcome. Further analysis revealed that for patients with spinal canal stenosis, the magnitude of the initial disability index did not correlate with outcome, whereas patients with low back pain or a prolapsed intervertebral disc fared better if they had high initial disability scores. Patients at 2-years follow-up were less accurate. However, previous surgery on the spine and low initial disability score were significant negative predictors of outcome at 6-month and 2-year follow-up.

Conclusions. The findings indicate that the absolute value and change in these scores after surgery vary from patient to patient, but that their percentage change is likely to be the best marker of outcome when such subjective scoring systems are used. (Key words: back pain, disability, spinal surgery) Spine 1994;19:2139-2143

The Oswestry Low Back Pain Questionnaire has been proposed as a reliable means of scoring the disability of patients with low back pain. It concentrates on scoring (0 to 5) 10 aspects of the patient's home and work life and analgesic requirements. The disability index (DI) score then is calculated as a percentage, with a high score indicating a high level of disability. The questionnaire has been shown to have an excellent re-test reliability on consecutive days, and has been shown to improve as a group of patients improves.

This study sought to assess whether absolute change in the DI (ΔDI) or the percentage change (ΔADI) can be linked with the level of patients' subjective improvement (SI). The change in index then is examined as an outcome measure in short-term follow-up to identify possible predictors of poor outcomes.

Methods

The Oswestry Low Back Pain Questionnaire was introduced by the senior author (DMcD) in 1984 as an adjunct to decision making in surgical treatment of the degenerate lumbar spine. By 1987, it had become generally adopted as a potential monitor of success (or failure) after such surgery. All patients were asked to complete it in the waiting room on initial assessment, at 6-months postoperatively, and at all subsequent assessments until discharge. Patients were excluded from this retrospective investigation if a preoperative or postoperative DI was not recorded. Patients with more than one operation were included again. Outcomes were assessed at two mean follow-up times: 6 months (range, 5 to 10 months) and 2 years (range, 18 to 36 months). Patients were not aware of the scoring of the questionnaire, nor did they see their previous questionnaires on subsequent visits. All operations were performed by the senior author (DMcD). Of 222 surgically managed patients over 4 years, all but one was seen at minimum 6-month follow-up. Preoperative and postoperative DI scores were available for 144. At a mean 2-year follow-up, preoperative and postoperative DI scores were available for 74 operations. Fifty-six operations had data at both 6-month and 2-year mean follow-up.

Patients' SI was obtained by asking them to rank their improvement from their preoperative state on a scale of 0 to 100. Patients who indicated no improvement were entered as 0. In this series, no patient stated they were worse.

The group examined was heterogeneous, consisting of patients who had undergone combined decompressive laminectomy and fusions for spinal canal stenosis, laminectomy and discectomy for prolapsed intervertebral discs, spinal fusions...
for low back pain, and revisions of pseudarthroses. Indication for surgery was grouped as low back pain, prolapsed intervertebral disc, lateral canal stenosis, or central canal stenosis. Operations were classified as laminectomy/discectomy, spinal fusion alone, or decompressive laminectomy and fusion.

The absolute change in the DI, or ΔDI, was obtained by subtracting the postoperative from the preoperative score. The percentage improvement in DI (ADI) was calculated simply as:

\[
\text{ADI} = \left( \frac{\text{Preoperative DI} - \text{Postoperative DI}}{\text{Preoperative DI}} \right) \times 100
\]

Thus, deterioration would be shown as a negative ΔDI or ADI. These indices then were subject to linear regression analysis and their correlation with SI was assessed.

Various patient factors then were analyzed for the group using ADI as an outcome measure.

All data were recorded on SPSS for Windows 5.0 computer program (SPSS Inc., Chicago, IL) and statistical analyses were performed with that software. Differences between mean scores were analyzed with an unpaired two-tailed test, and step-wise linear multiple regression analysis was used to analyze the relationship of certain variables with outcome, as measured by ΔDI.

### Results

Epidemiologic data on the patient groups at 6 months and 2 years are given in Table 1. The group of 144 patients in follow-up for 6 months had improvement of their DI scores from a mean of 45.5 to 27.0 (P < 0.001). For the 56 patients who had been in follow-up a mean of both 6 months and 2 years, the raw DI scores improved significantly from a mean of 42.4 to 27.3 (P < 0.001) at 6-month follow-up, and were stable at a mean of 26.3 at the later follow-up (P = 0.58).

The correlation of the absolute (ΔDI) and percentage (ADI) changes in indices with SI at a mean of 6 months and 2 years are shown in Table 2. Figures 1 and 2 compare these relationships in graphic form at 6-month follow-up.

Having established that the percentage change in DI score was correlated best with the patients’ subjective improvement, we used this as an outcome measure to identify whether any preoperative factors were associated with poor outcomes.

Table 3 gives the mean and standard deviation of ΔDI, allowing for the presence or absence of workers’ compensation or previous surgery to the spine, and for each diagnostic and surgical category. Note in Table 1 that at 2 years there are a limited number of patients in each subgroup, warranting caution in interpretation.

Step-wise linear multiple regression analysis revealed significant negative correlation with outcome for certain factors at mean 6-month follow-up. Table 4 lists the factors included in the analysis—those above the line are significantly negatively correlated with ΔDI. The most significant factor was previous surgery on the spine, followed by female gender, workers’ compensation, low initial DI score, increasing age, and spinal fusion alone as an operative procedure. Although the magnitude of the initial DI was correlated with a good outcome for the whole group, patients with spinal canal stenosis did not exhibit this relationship using simple linear regression. This does not reflect a failure to improve, because their mean ΔDI approached that of the whole group. However, a high initial disability score did correlate with a good outcome for patients with low back pain or prolapsed intervertebral disc (Table 5). Likewise, patients with spinal fusion operations were more likely to

---

**Table 1. Patient Characteristics**

<table>
<thead>
<tr>
<th></th>
<th>6 Months</th>
<th>2 Years</th>
</tr>
</thead>
<tbody>
<tr>
<td>n = 144</td>
<td>n = 74</td>
<td></td>
</tr>
<tr>
<td>Mean age (yrs)</td>
<td>50.6</td>
<td>50.4</td>
</tr>
<tr>
<td>F:M ratio</td>
<td>76:68</td>
<td>44:30</td>
</tr>
<tr>
<td>Heavy work</td>
<td>49</td>
<td>50</td>
</tr>
<tr>
<td>Workers’ comp.</td>
<td>33</td>
<td>34</td>
</tr>
<tr>
<td>Previous surgery</td>
<td>56</td>
<td>35</td>
</tr>
<tr>
<td>Diagnosis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Central stenosis</td>
<td>50</td>
<td>40</td>
</tr>
<tr>
<td>Lateral stenosis</td>
<td>34</td>
<td>13</td>
</tr>
<tr>
<td>Low back pain</td>
<td>37</td>
<td>13</td>
</tr>
<tr>
<td>Prolapsed disc</td>
<td>13</td>
<td>9</td>
</tr>
<tr>
<td>Procedure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Laminectomy/discectomy</td>
<td>26</td>
<td>18%</td>
</tr>
<tr>
<td>Decompression and fusion</td>
<td>87</td>
<td>60%</td>
</tr>
<tr>
<td>Spinal fusion</td>
<td>31</td>
<td>22%</td>
</tr>
</tbody>
</table>

**Table 2. Correlation of Absolute Change in DI (ΔDI) and Percentage Change in DI (ADI) With Subjective Improvement at 6-month and 2-year Follow-up**

<table>
<thead>
<tr>
<th></th>
<th>6-Months</th>
<th>2 Years</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>r</td>
<td>n</td>
</tr>
<tr>
<td>ΔDI</td>
<td>0.54</td>
<td>74</td>
</tr>
<tr>
<td>ADI</td>
<td>0.61</td>
<td>74</td>
</tr>
</tbody>
</table>

All correlations are significant at P < 0.001.

Figure 1. Correlation of ADI after surgery versus patients’ SI.
have a good outcome with high initial DI scores, whereas this was not so for other surgical categories. At mean 2-year follow-up, previous surgery and the magnitude of the initial DI score still significantly correlated with outcome by ΔDI (Table 6).

**Discussion**

It is useful to be able to show with a score such as the DI that patients have improved with surgery, because these scores are almost totally independent of the surgeon. Series often are reported with gradings rated by the surgeon or his or her assistant, which are subject to marked observer bias and which make meaningful interpretation difficult. Radiologic criteria for success in areas such as spinal fusion also are open to large observer error, and radiologic success may not correlate with clinical success. Physical findings may not change despite patients reporting improvement or deterioration.

We found that the ΔDI (percentage change) scores we examined correlated more closely with the patients' own ratings of their improvement (SI) than did ΔDI (absolute change) scores (Table 1). We believe that the absolute values of these scores are not necessarily comparable between patients, because different people interpret their conditions differently. However, we assume they will do so to a similar degree on each occasion they complete the questionnaires, and thus the percentage change may be a more comparable guide between patients. This is borne out by the higher correlations found when the percentage change is examined.

The patients’ ratings of their improvement (SI) has been used as a reference point in this article, but this measure has its own intrinsic problems. Although the patients completed this as part of their questionnaires, we assume that those who were worse were not comfortable saying so, creating a floor effect (Figures 1 and 2).

The patient population in this investigation was not specifically enrolled for the study. They represent the experience of one surgeon in private practice. Patients often were discharged once they had improved and thus were not included in follow-up for more than 18 months. Likewise, some patients were lost to follow-up, which always creates the possibility of bias. This study
cannot be taken as an outcome study for the results of spinal fusion or any other procedure, because the follow-up, even at 2 years, is inadequate to make such conclusions. However, we were able to identify some factors relating to short-term (6-month) outcome as measured by ΔDI.

One interesting finding was that although the initial magnitude of the DI score correlated with its percentage improvement overall, this was not so for all diagnostic categories. Patients with low back pain or prolapsed intervertebral discs who reported higher initial disability did better than those who did not, whereas there was no such relationship for spinal canal stenosis. Patients undergoing fusion operations who did well reported higher initial disability scores. Because outcome was measured by a percentage change in index, ΔDI, a high initial DI in itself cannot explain the tendency toward improvement. The data are not sufficient to shed further light on these findings, but the differences in age, diagnosis, and possible comorbidities each may play a role.

Female patients also fared worse, as did those involved in workers' compensation, older patients, and those with previous spinal surgery. Participation in heavy labor did not correlate with short-term outcome.

Many methods have been put forward to assess the improvement in pain, disability, and personality. These methods have not been assessed, to our knowledge, with respect to validity as measured by the patient's rating of SI at the end of treatment. When applied to a group of 25 patients with their first episode of back pain (a group assumed to have a "strong likelihood of recovery"), the Oswestry group had a statistically significant improvement in DI over 3 weeks. They also showed a correlation of \( r = 0.99 \) when scores were assessed on two consecutive days. More recently, in an article on the Minnesota Multiphasic Personality Index (MMPI), the DI was shown to have improved significantly in a group of anterior lumbar fusion patients. This paralleled an improvement in physical examination. The MMPI was not predictive of outcome in itself cannot explain the tendency toward improvement. The questionnaire also appears to be sufficiently simple to allow accurate self-administration in most patients. We caution against using the initial value of a DI score as an indication for surgery, although it appears to be predictive of improvement in some patients. A prospective investigation with long-term follow-up using ΔDI as an outcome measure may shed more light on the possible negative associations with outcome identified in the present study.

Table 6. Epidemiologic, Diagnostic, and Surgical Factors Negatively Correlated With ΔDI at a Mean of 2 Years (Step-wise Linear Multiple Regression)

<table>
<thead>
<tr>
<th></th>
<th>( r )</th>
<th>Change in ( r )</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Previous spinal surgery</td>
<td>0.39</td>
<td>—</td>
<td>0.001</td>
</tr>
<tr>
<td>Low initial DI score</td>
<td>0.46</td>
<td>0.07</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Other factors not significant: \( P > 0.05 \).

In an investigation of patients in lumbar supports, Million used slightly different subjective indices and compared these with objective indices. Subjective indices were shown to improve with the support, but objective indices did not, and it was concluded that subjective indices were more valuable. In an article promoting a functional rating scale very similar to the DI but published later, Evans showed in a group of patients in a pain clinic that this index improved significantly over 7 months. Again, it was an entirely subjective assessment by the same questionnaire given over time. In none of these investigations were the patients asked at final follow-up to rate their improvement after treatment. Thus, correlation between these statistically improved scores and patient satisfaction was not obtained.

Initial trials by the senior author (DMcD) were predicated on DI scores acting prospectively in a predictive way (for instance, that a high initial score would constitute a valid indication for surgery, and by implication predict those cases where surgery would result in improvement). This study has modified these assumptions. They may be valid for patients with low back pain and prolapsed intervertebral discs, but not for those with spinal canal stenosis. The absolute change in score (ΔDI) did not correlate as well with the patients' ratings of SI as did the percentage change in the scores (ΔDI).

It was not the primary purpose of this study to analyze the results of a particular procedure, and the surgical groupings in this investigation are somewhat heterogeneous in themselves. The study population was not large enough to examine the many surgical subgroups and indications for surgery at long-term follow-up. However, we have established that ΔDI can be used as an outcome measure across diagnostic and surgical boundaries in patients undergoing lumbar spinal surgery.

In conclusion, in assessing the outcome of surgery in the lumbar spine, the percentage change in Oswestry DI (ΔDI) is reliable, independent of surgeon bias, and correlates well with the patients' subjective assessments of improvement. The questionnaire also appears to be sufficiently simple to allow accurate self-administration in most patients. We caution against using the initial value of a DI score as an indication for surgery, although it appears to be predictive of improvement in some patients. A prospective investigation with long-term follow-up using ΔDI as an outcome measure may shed more light on the possible negative associations with outcome identified in the present study.

References


Address reprint requests to
Dr. David Little
155 Flood Street
Leichhardt, NSW, 2040
Australia

Point of View

Jeffrey N. Katz, MD, MS
Boston, Massachusetts

This article reinforces a central paradigm in orthopedic outcomes research: Because the goal of most elective spine surgery is to enhance the quality of patients’ lives, the most valid standard for judging surgical outcome is evidence of improvement in quality of life. Thus, the moderately high correlations between changes in questionnaire scores and patients’ ratings of subjective improvement support the validity of functional status questionnaires as surgical outcome measures.

Nevertheless, this article should be interpreted cautiously by clinicians and policy makers alike. As the authors state candidly, the follow-up rates are low. Extensive loss to follow-up may create bias in either direction (those who fail to return may have better—or worse—outcomes than the reported sample), rendering results difficult to interpret. In addition, the authors were unable to measure psychologic status and the extent of medical comorbidity, which may confound the reported effects of gender and age. Thus, the provocative findings of this study must be viewed cautiously, as hypothesis generating observations rather than established associations.

The authors report that the percentage change in disability score correlates better than the absolute change with subjective improvement, but they do not discuss possible explanations for this phenomenon. It may reflect an underlying biologic truth—improvement truly may be perceived in proportional rather than absolute terms. In fact, many familiar processes have proportional rather than linear kinetics, such as the growth of bacteria in culture and of money in an interest bearing account. Alternatively, the observation may arise from a structural limitation in the Oswestry instrument, referred to as a ceiling effect. Let’s assume that Patient A has a low disability score preoperatively (say 15 on a scale from 0 to 100) and patient B a higher score (say 75). Each patient reports identical, substantial improvement on a 0–100 subjective index (say 80% better).

Because Patient A started with a low score and cannot get much lower, the absolute change in score will be minimal (say from 15 to 10). Patient B has ample room for improvement on the scale and will record a large absolute change (say from 80 to 40). Patient B has eightfold greater absolute improvement, and 1.5 fold greater percentage change than A, yet according to the external standard, each improved precisely the same amount. What happened? Both percentage and especially absolute change were attenuated by ceiling effects.

This seemingly pedantic quandary may be crucial to health care policy. The authors show that patients with the best preoperative disability scores have the worst outcomes—that those with the most to gain gain the most. Research such as this is used increasingly to establish clinical policies. Although the movement toward evidence-based guidelines is laudable, poor policies may emerge if the research is not done carefully. For example, a health care payer might consider denying surgery to patients with the most favorable preoperative disability scores because research indicates that such patients improve the least. As we have seen, this result simply may arise from a ceiling effect of the instrument rather than a true association. Methodologic problems such as ceiling effects can be identified and resolved, but only with painstakingly detailed data analyses. As health services research begins to shape health care policy, we should demand no less.