Eur Spine J (2006) 15: 66–73
DOI 10.1007/s00586-004-0816-z

ORIGINAL ARTICLE

A. F. Mannion
A. Junge
D. Grob
J. Dvorak
J. C. T. Fairbank

Development of a German version of the Oswestry Disability Index. Part 2: sensitivity to change after spinal surgery

Received: 5 May 2004
Accepted: 1 August 2004
Published online: 26 April 2005

Part 1 of this article can be found at http://dx.doi.org/10.1007/s00586-004-0815-0

A. F. Mannion · A. Junge
D. Grob · J. Dvorak
Schulthess Klinik, Lengghalde 2,
8008 Zürich, Switzerland

A. F. Mannion (✉)
Department Rheumatology and Institute of Physical Medicine, University Hospital, Zürich, Switzerland
E-mail: anne.mannion@kws.ch
Tel.: +41-44-3857584
Fax: +41-44-3857590

J. C. T. Fairbank
Nuffield Orthopaedic Centre,
Oxford, UK

Abstract When functional scales are to be used as treatment outcome measures, it is essential to know how responsive they are to clinical change. This information is essential not only for clinical decision-making, but also for the determination of sample size in clinical trials. The present study examined the responsiveness of a German version of the Oswestry Disability Index version 2.1 (ODI) after surgical treatment for low back pain. Before spine surgery 63 patients completed a questionnaire booklet containing the ODI, along with a 0–10 pain visual analogue scale (VAS), the Roland Morris disability questionnaire, and Likert scales for disability, medication intake and pain frequency. Six months after surgery, 57 (90%) patients completed the same questionnaire booklet and also answered Likert-scale questions on the global result of surgery, and on improvements in pain and disability. Both the effect size for the ODI change score 6 months after surgery (0.87) and the area under the receiver operating characteristics (ROC) curve for the relative improvement in ODI score in relation to global outcome 6 months after surgery (0.90) indicated that the ODI showed good responsiveness. The ROC method revealed that a minimum reduction of the baseline (pre-surgery) ODI score by 18% (equal to a mean 8-point reduction in this patient group) represented the cut-off for indicating a "good" individual outcome 6 months after surgery (sensitivity 91.4% and specificity 82.4%). The German version of the ODI is a sensitive instrument for detecting clinical change after spinal surgery. Individual improvements after surgery of at least an 18% reduction on baseline values are associated with a good outcome. This figure can be used as a reliable guide for the determination of sample size in future clinical trials of spinal surgery.

Keywords Low back pain · Spine surgery · Condition-specific questionnaires · Sensitivity to change · Responsiveness

Introduction

In recent years patient-oriented, self-administered questionnaires have been used with increasing frequency in the assessment of outcome after treatment for low back pain [11]. For assessing "back-specific function", most state-of-the-art reviews [6, 11] recommend either the Oswestry Disability Index (ODI [15, 16] or the Roland Morris Questionnaire (RM [28]). A number of studies have been carried out to examine the psychometric characteristics of these instruments, especially when validating various non-English language versions, but

most of these investigations have only been concerned with the reliability (internal consistency and test–retest reliability) and validity of the given questionnaires (e.g. [7, 19]). Good reliability and validity are prerequisites of any instrument, especially when it is to be used to discriminate between subjects or predict prognosis [3, 24, 29]. However, the requirements for successful cross-sectional discrimination are not necessarily the same as those for successful longitudinal evaluation [24], and when functional scales are to be used as treatment outcome measures, it is essential to know how well they can detect small but important clinical changes, i.e. how "responsive" they are [13]. This information is essential not only for clinical decision-making, but also for the determination of sample size in clinical trials, to ensure that they are adequately powered to detect a difference between treatments if one is present.

Previous studies have used "effect sizes" to examine the responsiveness of the Oswestry Disability Index to surgical treatment [22]. However, the effect size, i.e. the mean change-score for a group of patients divided by the standard deviation of all the change-scores, predominantly depicts the overall group response; a more complete picture of the responsiveness of an outcome measure on an individual basis is obtained with the use of receiver operating characteristics (ROC). The ROC approach assesses how successfully a given change-score can discriminate between patients who improved and those who did not improve as a result of any given treatment [13]. In this way, both sensitivity *and* specificity to change for a range of possible cut-off change-scores can be calculated.

The present study examined the responsiveness of a German version of the Oswestry disability index, as compared with that of the Roland Morris Disability Score [14, 28] and the visual analogue scale for pain intensity, in a group of Swiss patients undergoing spine surgery.

Materials and methods

The Oswestry disability index

The ODI version 2.1 (the English version of which is reprinted in full in [27]) is a self-administered questionnaire, which comprises ten items to assess the extent of the patient's back pain and difficulty in carrying out nine different activities of daily life: personal care, lifting, walking, sitting, standing, sleeping, sex life, social life, and travelling. The questionnaire is completed in reference to the patient's functional status "today". Each item is scored from 0 to 5, with higher values representing greater disability. The total score is multiplied by 2, and normally expressed as a percentage (in the present study this percentage will simply be referred to as "the ODI score" and discussed in terms of points (0–100), to avoid confusion when discussing percentage changes in the score (as a mathematical expression) following surgery).

The cross-cultural adaptation, reliability and validity of the German version of the ODI version 2.1 are described in detail in Mannion et al. [25].

Patients

Sixty-eight patients with low back pain (LBP) agreed to take part in the study. All had been referred to the hospital's Spine Unit for surgery in connection with spinal stenosis, herniated disc, failed back, spondylolisthesis, or degenerative disease with chronic LBP. The patients completed a baseline questionnaire (see below), sent to them by post approximately 2–3 weeks before their operation. Sixty-three underwent the planned surgery (mainly decompression, fusion, metal removal, or a combination of these), and 57 of these (90%) completed a second questionnaire 6 months after the operation. There were 31 women and 26 men, with a mean age of 53.2 (14.6) years.

Questionnaires

The patients completed a questionnaire booklet containing the German version of the ODI [25], 0–10 visual analogue scales for back/leg pain intensity in the last week (VAS_{pain}) and for general health (VAS_{health}), and a German version of the Roland Morris (RM) disability questionnaire (validated by Exner and Keel [14]). The RM enquires as to whether back pain hinders the performance of 24 activities of daily living (today), each with possible responses of "yes" and "no"; the RM score ranges from 0 to 24 points. At follow-up, the questionnaire booklet also contained the following items: two Likert scale questions enquiring how the patient's (1) back/leg pain and (2) disability in everyday activities had changed compared with the time before the operation (in each case, 6 categories from "now free of complaints/problems" to "now worse"); a question about how much the operation had helped (5 categories from "helped a lot" to "made things worse"); and a question enquiring as to whether, with his/her current knowledge of the result, the patient would make the same decision to undergo surgery if he/she found himself in the same situation as before the operation ("yes"/"no").

The study was approved by the local ethics committee.

Statistical analysis

Paired *t*-tests were used to examine the significance of the change in group mean scores for each instrument,

from pre-surgery to 6 months post-surgery. The effect size for each instrument was calculated by taking the mean of the individual change scores and dividing this by the corresponding standard deviation of these change scores [4]. The effect size was also calculated for each instrument in relation to the five categories of the global outcome question, "did the operation help?" Examination of the correlation between the instrument change-scores and the (ordinal) global outcome scale gave a further indication of responsiveness [30]. The sensitivity and specificity of each instrument, relative to patient global outcome, was examined using the receiver operating characteristic (ROC) method [12]. It has been suggested that instrument responsiveness can be considered analogous to evaluating a diagnostic test, in which the instrument is the diagnostic test and the global outcome represents the gold standard [12]. The ROC curve synthesises information on sensitivity and specificity for detecting improvement according to some dichotomised, external criterion. It consists of a plot of "true-positive rate" (sensitivity) versus "false positive rate" (1-specificity) for each of several possible cut-off points in change score [12]. Thus, sensitivity and specificity are calculated for a change score of 1 point, 2 points, and so on. The five global outcome categories for the question "how much did the operation help?" were collapsed to provide a dichotomous outcome variable: "good outcome" (included "helped a lot" and "helped") and "poor outcome" (included "only helped a little", "didn't help", "made things worse"). (As most of the patients were undergoing elective surgery, we felt that the overall result "only helped a little" should be categorised as a poor outcome.) The area under the ROC curve (ROC_{area}) was interpreted as the probability of correctly discriminating between patients with a "good" and a "poor" outcome, based on the change in instrument scores (examined for ODI, RM and VAS_{pain}). The ROC_{area} can range from 0.5 (no accuracy in discriminating) to 1.0 (perfect accuracy in discriminating). The ROC curve was used to indicate the cut-off change-score for distinguishing between "good" and "poor" outcomes [13], using the approach of minimising "errors" (equivalent to maximising the sum of the specificity and sensitivity) [1].

Statistical significance was accepted at the $P < 0.05$ level.

Results

Group mean scores before and 6 months after surgery

The mean scores for ODI, RM, VAS_{pain} and VAS_{health} before and 6 months after surgery are shown in Table 1. Each of the disability scores (ODI and RM) showed a significant reduction of 30–35% 6 months after surgery ($P < 0.001$), and the changes in scores correlated highly significantly with each other (Fig. 1).The VAS_{pain} showed a reduction of 43% 6 months after surgery ($P < 0.001$) and VAS_{health} improved by about 22% ($P = 0.02$).

Considering the whole group data, the effect sizes were similar for the two disability questionnaires (ODI, 0.84; RM, 0.90) and were both somewhat lower than that of VAS_{pain}(1.07). As expected, the effect size for VAS_{health} (0.29) was considerably smaller than that for any of the condition-specific measures. (The VAS_{health} measures were not considered in any further analyses.)

Global outcome 6 months after surgery

Six months after surgery, 40% of patients reported that the operation "helped a lot", 26% that it "helped", 14% that it "only helped a little", 18% that it "didn't help" and 2% that it "made things worse". There was a highly significant correlation between these "global outcome" ratings and the Likert-scale ratings of perceived improvement in disability (Spearman's $\rho = 0.83$, $P < 0.001$) and perceived improvement in pain (Spearman's $\rho = 0.82$, $P < 0.001$). This indicated that the "global outcome" categories, themselves, had good construct validity in relation to changes in perceived pain and disability.

Correlation between change-scores and outcome category

The change-scores for ODI and RM each showed a significant correlation with the global outcome categories when the latter were expressed as ordinal data (scale of 1–5): ODI Spearman's $\rho = 0.69$, $P < 0.001$; RM Spearman's $\rho = 0.67$, $P < 0.001$.

Table 1 Questionnaire scores before and 6 months after spinal surgery ($n = 57$)

Variable	Before surgery mean (SD)	Six months after surgery mean (SD)	Comparison before versus 6 months after surgery *P*-value	Effect size
ODI	45.0 (15.6)	29.5 (21.0)	< 0.001	0.84
Roland Morris	15.0 (4.4)	9.7 (6.4)	< 0.001	0.90
$VAS_{pain\ intensity}$	7.0 (2.0)	4.0 (3.0)	< 0.001	1.07
$VAS_{general\ health}$	4.7 (2.6)	5.8 (2.9)	0.020	0.29

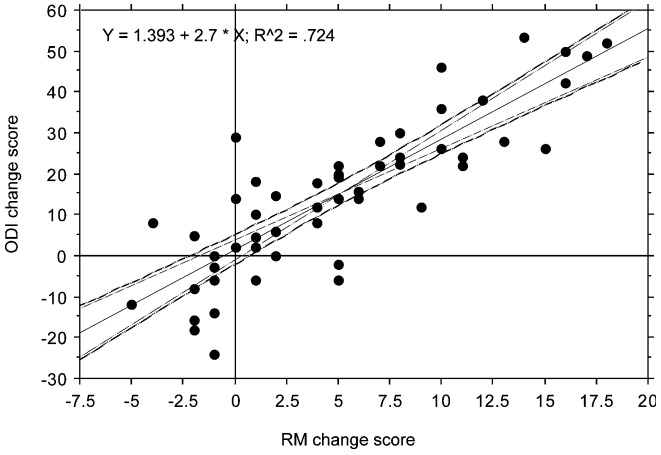

Fig. 1 Relationship between changes in RM and ODI (before surgery to 6 months after surgery) (with 95%CI for the mean and slope of the regression equation)

Change-scores in each outcome category

Table 2 shows the mean change-scores and effect sizes (pre-surgery to 6 months after surgery) for the ODI, RM and VAS_{pain} for each of the five global outcome categories. For each instrument, the mean change-scores differed between the outcome categories, though not always statistically significantly (NB the group sizes for the poor outcome categories were generally quite small). A significant difference in the score-change between patients who reported that the operation "helped a lot" and all less favourable outcomes was observed for each instrument.

The difference in the mean score-change between the categories "didn't help" and "helped" was 10 points for the ODI, 3.6 points for the RM and 2.1 points for the 0–10 VAS_{pain}.

When the five-category global outcome ratings were dichotomised (see Statistical analysis above), the majority of patients (66%) reported a "good" outcome ("poor" outcome, 34%). As expected, the effect size statistics in the "good' group were significantly greater than those in the "poor" group for each outcome measure (Table 2). Thus, all three instruments showed good *sensitivity* to change. For the two disability questionnaires, ODI and RM, the effect size statistics for the "good" outcome group were similar (both around 1.3), and both were somewhat lower than that of VAS_{pain}(1.6). The difference in the mean ODI change score between the "good" and "poor" categories was approximately 20 points.

Both disability questionnaires showed good *specificity*, i.e. the effect size for the patients in the "poor" global outcome group was minimal (Table 2). In contrast, VAS_{pain} showed a moderate effect size of 0.50 for the "poor" outcome group; even for the sub-category "operation didn't help", the effect size for $VAS_{pain}$was 0.53. This indicates that some patients who had not improved according to their global outcome category had still shown a moderate improvement in relation to pain intensity, suggesting that the VAS_{pain} is less specific to change than the two disability questionnaires.

Receiver operating characteristics: area under the ROC curve

Using the dichotomised global outcome as the "external criterion", the ROC curves for the change-scores for ODI, RM and VAS_{pain} were each far to the left above the diagonal, indicating that each had some discriminative ability. The ROC_{areas} for ODI, RM and VAS_{pain} were 0.85 (SEM 0.06), 0.84 (SEM 0.05) and 0.88 (SEM 0.05), respectively.

When the change-scores were expressed as a percentage of their baseline value, the areas under the ROC curves were even higher [0.90 (SEM 0.04), 0.86 (SEM 0.05), 0.92 (SEM 0.04) for ODI, RM and VAS_{pain}, respectively] (Fig. 2).

Very similar results were obtained when, instead of using the "global outcome rating", dichotomous categories formed by collapsing the 6-category Likert scales of the degree of improvement in pain and in disability were used (data not shown).

Table 2 Mean ODI, RM and pain scores in relation to the global rating of the success of surgery 6 months postoperatively ($n = 57$)

Global rating of outcome	Proportion of patients in each category (%)	Change in ODI score points	Effect size ODI	Change in RM score points	Effect size RM	Change in VAS_{pain}	Effect size VAS_{pain}
Operation helped a lot	40	30.0* (15.5)	1.94	9.9* (5.7)	1.74	5.0* (2.5)	2.00
Operation helped	26	11.1 (14.6)	0.76	4.0 (3.8)	1.05	2.9** (2.5)	1.16
Operation helped only a little	14	3.4 (13.2)	0.26	1.9 (4.4)	0.43	0.8 (1.3)	0.62
Operation didn't help	18	1.1 (11.0)	0.10	0.4 (1.8)	0.22	0.8 (1.5)	0.53
Operation made things worse	2	0.0 (–)		−1.0 (–)		−0.5 (–)	
Global outcome "good" (1, 2)	66	22.4* (17.6)	1.27	7.5* (5.7)	1.32	4.2* (2.6)	1.61
Global outcome "poor" (3, 4, 5)	34	1.9 (11.7)	0.16	0.9 (3.1)	0.29	0.7 (1.4)	0.50

*Significantly different from all other outcomes categories $P < 0.05$;, **Significantly different from "operation didn't help" $P < 0.05$

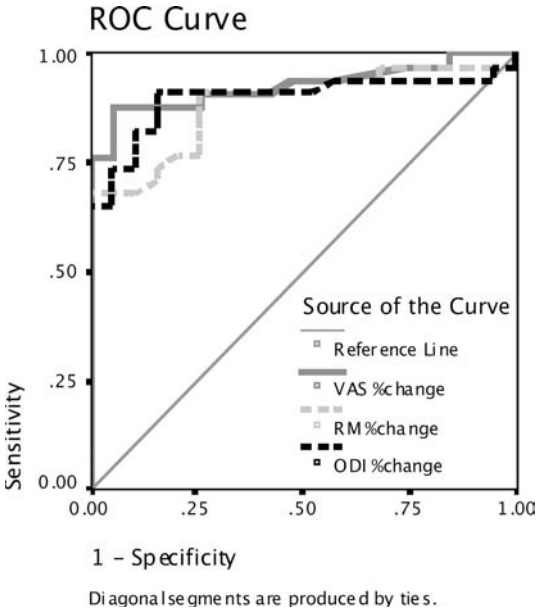

Fig. 2 ROC curves of the percentage change-scores for ODI, RM and VAS_{pain} using the global outcome rating (''good'' versus ''poor'') as the external criterion

Receiver operating characteristics: cut-off change-scores for predicting outcome

Assuming equivalent importance for false-positive and false-negative errors, the absolute change-scores with the best cut-off points for predicting global outcome (''good''/''poor'') were calculated. Different cut-offs sometimes gave the same optimised product of sensitivity and specificity, and in these instances the range is given. The cut-offs were approximately 11 points for ODI (83.8% sensitivity, 84.2% specificity), 1.5 points for RM (83.8% sensitivity, 73.7%specificity) and 1.5–2.8 points for VAS_{pain}(76.3–81.6% sensitivity, 73.7–89.5% specificity). The corresponding cut-offs for the percentage score reduction from baseline were 18% for ODI (91.7% sensitivity, 84.2% specificity), 8% for RM (88.9% sensitivity, 73.7% specificity) and 32% for VAS_{pain} (86.5% sensitivity, 94.7% specificity). These change-scores can be considered to represent the minimal clinically significant change, at the level of the individual patient.

Discussion

The German version of the ODI used in the present study was developed in accordance with established recommendations [2, 20] and has been found to be a reliable and valid instrument [25].

In the present study, the responsiveness of the ODI, determined using the various recommended statistical methods [30], was confirmed in a group of LBP patients undergoing spinal surgery. The difference in the mean ODI change score between the global outcome categories ''good'' and ''poor'' was approximately 20 points. This is higher than the score of 10 points previously reported by Hagg et al. [22] for the difference in ODI change-score between patients who showed ''improvement'' and those who showed ''no relevant change'' after surgery. However, in the present study, the global category ''good'' included not only those patients for whom the operation ''helped'', but also those who reported that the operation ''helped a lot'' (i.e. more than just ''improved''). When the difference between the narrower categories ''operation helped'' and ''operation didn't help'' was examined (analogous to the analysis carried out by Hagg et al. [22]), then a similar mean ODI change-score (10 points) to that of Hagg et al. [22] was obtained. In the present study, no minimal clinically relevant difference for ''worsening of the condition'' could be calculated, as too few patients declared that the operation ''made things worse''.

Demonstrating that post-treatment scores are significantly different from pre-treatment scores and that the change-scores are greater in an ''improved'' group than in a ''no change'' group addresses the sensitivity to change of the scale, but not its specificity [4, 12]. The concept specificity to change is also important, since changes without clinical relevance may occur in function scale scores [12]. For example, in the present study, the change-score for the VAS_{pain} was very high in the ''good'' outcome group (4.2 points; effect size 1.6) but was also moderately high in the ''poor'' outcome group (0.7 points; effect size 0.50), indicating that a number of patients who were not improved according to the global outcome criterion still decreased appreciably in their pain score.

In order to better quantify the responsiveness of the ODI, the ROC method was used. The area under the ROC was 0.85 for ODI and 0.84 for RM. These values are generally somewhat higher than those previously reported in the literature for acute or chronic LBP patients undergoing conservative treatment (ODI: 0.76 [4], 0.78 [29], 0.94 [17], 0.78 [8]; RM: 0.79 [29], 0.93 [4], 0.77 [8]). Slight differences between studies may be the result of the questionnaire version used (e.g. Beurskens et al. [4] used an older version of the ODI), or the differing LBP populations and treatment strategies investigated (e.g. acute versus chronic LBP; conservative versus surgical patients).

The patients showed quite wide-ranging disability scores at baseline, and we therefore considered it of interest to examine whether the responsiveness of the instruments improved when, instead of *absolute change* scores, *relative change* scores (i.e. before surgery score-6 month score/before surgery score) were used as the ''discriminating variable'' in the ROC analysis. For each

of the three instruments (ODI, RM and VAS_{pain}), the areas under the ROC curve were even higher (0.90, 0.86, and 0.92, respectively) when the relative scores were used. Thus, we tentatively suggest that it may be more appropriate to discuss the cut-off scores for indicating "improvement" (as determined from ROC curves) in terms of the percentage change-score from baseline. Although percentages of change scores are not recommended for use in the statistical analysis of outcome in clinical trials [31], they may be of some practical use for the calculation of sample size for such trials, especially when populations with differing baseline scores are being investigated: using the percentage of change value, one can calculate the corresponding absolute score-change required to be considered as "improvement" in relation to the expected baseline scores for the given population. This absolute value can then be used in the subsequent power calculations. For the ODI, a "good" global outcome was predicted (with 92% sensitivity and 84% specificity) by a change in ODI score greater than or equal to an 18% reduction from baseline values. In relation to the mean pre-surgery ODI value in the present study (45 points), this is equivalent to an approximate 8-point reduction. The ROC analysis done using the *absolute* change-score revealed a cut-off for a good outcome of 11 points. Interestingly, both of these cut-off values are somewhat higher than the previously reported values for conservatively treated acute or chronic LBP patients of 4–6 points [4] and 6 points [17]. The precise value may depend on the patient group and treatment under investigation: in less disabled patients, changes of up to 6 points may represent a similar *percentage* reduction from baseline to that reported for the patients in the present study.

An individual change-score of 8–11 points lies relatively close to minimal detectable change ($MDC_{95\%}$) for the ODI (9 points; Mannion et al. [25]). This is the value required to detect (with 95% confidence) real *individual* change over and above measurement error [23]. Nonetheless, in clinical practice, the 95% confidence level may be too strict for governing the presence of real individual change: with a standard error of measurement (SEM) of 3.4 points for the ODI [25], a score-change of approximately 7 points (2×SEM) could still be considered "real change" with a 92% confidence level, or of 5 points (1.5×SEM) with an 86% confidence level [25].

Clinically relevant *group mean* changes in an outcome instrument appear to be somewhat more difficult to define, and are not (directly) determined by the same factors as those governing clinically relevant *individual* change [18, 30]. As regards the ODI, the clinically relevant *group mean* change is likely to be considerably lower than 10 points; indeed, previous studies have suggested that differences in group mean scores as low as 4 points can carry clinical significance [26]. Perhaps power calculations for clinical trials in low back pain research should be based on the proportion of individuals who are expected to achieve a clinically relevant change-score rather than on the expected (and difficult to ascertain) clinically relevant group mean change; this might lead to more relevant findings, although the necessary sample sizes for the trials would undoubtedly increase [5, 10].

It is important to point out that both strategies used to assess an instrument's responsiveness (effect sizes and the ROC method) depend on some external criterion for rating "improvement"; further, to perform the ROC analysis, this criterion must be dichotomous. However, there exists no "gold standard" for assessing outcome and, in reality, there are often more than two grades of improvement that can be considered to carry clinical relevance. In the present study, the five-category Likert scale for "how much the operation helped" was collapsed into a dichotomous variable for "good" and "poor" outcome to provide the external criterion for use in the ROC analyses. Although we do not suggest that this measure constitutes a definitive gold standard for assessing outcome, it can at least be expected to reflect the most important changes to the individual patient elicited by the operation. The construct validity of this global outcome scale appeared to be satisfactory: it showed highly significant associations with each of the two Likert scale ratings for improvement in disability and in pain, and when the effect size/ROC analyses were carried out using improvement in disability or pain as the external criteria, the results were largely consistent with those obtained using the global outcome scale. In the absence of a true gold standard, the best one can do is ensure construct validity of the criterion that is ultimately chosen for use [12]. Further, as highlighted by Beurskens et al. [4], most people would be reluctant to label patients as improved or worse contrary to their personal rating of the global effect of treatment. An alternative may have been to use the answer to the question "if you found yourself in the same situation as before the operation, would you make the same decision to undergo surgery, with your current knowledge of the result?" (yes/no). However, although it has been used as a main outcome measure in other retrospective studies (e.g. [9]), our experience with this question has indicated that it is a confusing construct for some patients to understand. Some report no change in (or even a worsening of) symptoms or disability, but still tend to say "yes", as if perhaps interpreting the question to be an enquiry regarding their propensity to think that "everything's worth a try" as opposed to a direct evaluation of their perceived outcome after the intervention received. Further, people sometimes simply don't like to consider that they "made a wrong decision" and therefore answer "yes" regardless of the outcome, to avoid being confronted with feelings of regret or self-blame. As such, and in keeping with the methodology used by

previous authors [21], we consider that collapsing the five-category Likert scale into a dichotomous variable provides the more accurate representation of global outcome.

Conclusion

Our studies on the responsiveness of the German version of the ODI version 2.1 are the first to address both the sensitivity and specificity of ODI change-scores in categorising outcome after spinal surgery and to provide cut-off scores for interpreting meaningful clinical change. A good global outcome was predicted (with 92% sensitivity and 84% specificity) by a change in ODI score greater than or equal to an 18% reduction of the individual's baseline value.

Acknowledgements The authors would like to thank Gordana Balaban, Simon Smit and Katrin Knecht for the administration of the questionnaires. The study was funded by the Schulthess Klinik Research Funds.

References

1. Altman DG, Bland JM (1994) Statistics notes: diagnostic tests 3: receiver operating characteristic plots. BMJ 309:188
2. Beaton DE, Bombardier C, Guillemin F, Ferraz MB (2000) Guidelines for the process of cross-cultural adaptation of self-report measures. Spine 25:3186–3191
3. Beurskens AJ, de Vet HC, Köke AJ, van der Heijden GJ, Knipschild PG (1995) Measuring the functional status of patients with low back pain. Assessment of the quality of four disease-specific questionnaires. Spine 20:1017–1028
4. Beurskens AJHM, de Vet HCW, Köke (1996) Responsiveness of functional status in low back pain: a comparison of different instruments. Pain 65:71–76
5. Bhandari M, Lochner H, Tornetta P 3rd (2002) Effect of continuous versus dichotomous outcome variables on study power when sample size of orthopaedic randomised trials are small. Arch Orthop Trauma Surg 122:96–98
6. Bombardier C (2000) Outcome Assessments in the evaluation of treatment of spinal disorders. Summary and general recommendations. Spine 25:3100–3103
7. Boscainos PJ, Sapkas G, Stilianessi E, Prouskas K, Papadakis SA (2003) Greek versions of the Oswestry and Roland-Morris disability questionnaires. Clin Orthop 411:40–53
8. Davidson M, Keating JL (2002) A comparison of five low back disability questionnaires: reliability and responsiveness. Phys Ther 82:8–24
9. Davis TT, Delamarter RB, Sra P, Goldstein TB (2004) The IDET procedure for chronic discogenic low back pain. Spine 29:752–756
10. Deyi BA, Kosinski AS, Snapinn SM (1998) Power considerations when a continuous outcome variable is dichotomised. J Biopharm Stat 8:337–352
11. Deyo RA, Battie M, Beurskens AJHM, Bombardier C, Croft P, Koes B, Malmivaara A, Roland M, Von Korff M, Waddell G (1998) Outcome measures for low back pain research. A proposal for standardized use. Spine 23:2003–2013
12. Deyo RA, Centor RM (1986) Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. J Chronic Dis 39:897–906
13. Deyo RA, Diehr P, Patrick DL (1991) Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Control Clin Trials 12(Suppl):142–158
14. Exner V, Keel P (2000) Erfassung der Behinderung bei Patienten mit chronischen Rückenschmerzen. Schmerz 14:392–400
15. Fairbank JC, Couper J, Davies JB, O'Brien JP (1980) The Oswestry low back pain questionnaire. Physiotherapy 66:271–273
16. Fairbank JC, Pynsent PB (2000) The Oswestry disability index. Spine 25:2940–2952
17. Fritz JM, Irrgang JJ (2001) A comparison of a modified Oswestry low back pain disability questionnaire and the Quebec back pain disability scale. Phys Ther 81:776–788
18. Goldsmith C, Boers M, Bombardier C, Tugwell P (1993) Criteria for clinically important changes in outcomes: development, scoring and evaluation of rheumatoid arthritis patient and trial profiles. OMERACT Committee. J Rheumatol 20:561–565
19. Grotle M, Brox JI, Vollestad NK (2003) Cross-cultural adaptation of the Norwegian versions of the Roland-Morris disability questionnaire and the Oswestry disablity index. J Rehabil Med 35:241–247
20. Guillemin F, Bombardier C, Beaton D (1993) Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. J Clin Epidemiol 46:1417–1432
21. Hagg A, Fritzell P, Oden A, Nordwall A (2002) Simplifying outcome measurement: evaluation of instruments for measuring outcome after fusion surgery for chronic low back pain. Spine 27:1213–1222
22. Hagg O, Fritzell P, Nordwall A (2003) The clinical importance of changes in outcome scores after treatment for chronic low back pain. Eur Spine J 12:12–20
23. Hopkins WG (2000) Measures of reliability in sports medicine and science. Sports Med 30:1–15
24. Kirschner, Guyatt A (1985) A methodological framework for assessing health indices. J Chronic Dis 38:27–36
25. Mannion AF, Junge A, Fairbank JCT, Dvorak J, Grob D (2005) Development of a German version of the Oswestry Disability Index. Part 1: cross-cultural adaptation, reability, and validity. Eur Spine J. DOI 10.1007/s00586-004-0815-0
26. Meade T, Browne W, Mellows S et al. (1986) Comparison of chiropractic and outpatient management of low back pain: a feasibility study. J Epidemiol Commun Health 40:12–17
27. Roland M, Fairbank J (2000) The Roland-Morris disability questionnaire and the Oswestry disability questionnaire. Spine 25:3115–3124
28. Roland M, Morris R (1983) A study of the natural history of back pain. Part 1: Development of a reliable and sensitive measure of disability in low-back pain. Spine 8:141–144

29. Stratford PW, Binkley J, Solomon P, Gill C, FInch E (1994) Assessing change over time in patients with low back pain. Phys Ther 74:528–533

30. Stratford PW, Spadoni G, Kennedy D, Westaway MD, Alcock GK (2002) Seven points to consider when investigating a measure's ability to detect change. Physiother Can Winter 2002:16–24

31. Vickers AJ (2001) The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study. BMC Med Res Methodo 1(1):1–6 (Epub: www.biomedcentral.com/1471-2288/1471/1476