The quality of spine surgery from the patient's perspective: Part 2. Minimal clinically important difference for improvement and deterioration as measured with the Core Outcome Measures Index

A. F. Mannion · F. Porchet · F. S. Kleinstück ·
F. Lattig · D. Jeszenszky · V. Bartanusz · J. Dvorak ·
D. Grob

Received: 22 December 2008 / Revised: 17 February 2009 / Accepted: 1 March 2009 / Published online: 19 March 2009
© Springer-Verlag 2009

Abstract The Core Outcome Measures Index (COMI) is a reliable and valid instrument for assessing multidimensional outcome in spine surgery. The minimal clinically important score-difference (MCID) for improvement (MCID_{imp}) was determined in one of the original research studies validating the instrument, but has never been confirmed in routine clinical practice. Further, the MCID for deterioration (MCID_{det}) has never been investigated; indeed, this needs very large sample sizes to obtain sufficient cases with worsening. This study examined the MCIDs of the COMI in routine clinical practice. All patients undergoing surgery in our Spine Center since February 2004 were asked to complete the COMI before and 12 months after surgery. The COMI has one question each on back (neck) pain intensity, leg/buttock (arm/shoulder) pain intensity, function, symptom-specific well-being, general quality of life, work disability, and social disability, scored as a 0–10 index. At follow-up, patients also rated the global effectiveness of surgery, on a 5-point Likert scale. This was used as the external criterion (“anchor”) in receiver operating characteristics (ROC) analyses to derive cut-off scores for individual improvement and deterioration. Twelve-month follow-up questionnaires were returned by 3,056 (92%) patients. The group mean COMI score change for patients

declaring that the “operation helped” was a reduction of 3.1 points; the corresponding value for those whom it “did not help” was a reduction of 0.5 points. The group MCID_{imp} was hence 2.6 points reduction; the corresponding group MCID_{det} was 1.2 points increase (0.5 minus −0.7). The area under the ROC curve was 0.88 for MCID_{imp} and 0.89 for MCID_{det} (both $P < 0.0001$), indicating that the COMI had good discriminative ability. The cut-offs for individual improvement and deterioration, respectively, were ≥ 2.2 points decrease (sensitivity 81%, specificity 83%) and ≥ 0.3 points increase (sensitivity 83%, specificity 88%). The MCID_{imp} score of 2.2 points was similar to that reported in the original study (2–3 points, depending on external criterion used). The MCID_{det} suggested that the COMI is less responsive to deterioration than to improvement, a phenomenon also reported for other spine outcome instruments. This needs further investigation in even larger patient groups. The MCIDs provide essential information for both the planning (sample size) and interpretation of the results (clinical relevance) of future clinical studies using the COMI.

Keywords COMI · Outcome · Spine surgery

Introduction

A decade ago, a standardized set of outcome measures for use in patients with back pain was proposed by a multinational group of experts [7]. The domains recommended for inclusion were pain, back-specific function, generic health status (well-being), work disability, social disability, and patient satisfaction [4, 7]. Accordingly, the group proposed a parsimonious set of six questions that would cover each of these domains, yet be brief enough to

A. F. Mannion (✉)
Spine Center Division, Department of Research
and Development, Schulthess Klinik, Lengghalde 2,
8008 Zurich, Switzerland
e-mail: anne.mannion@kws.ch; anne@annefmannion.com

F. Porchet · F. S. Kleinstück · F. Lattig · D. Jeszenszky ·
V. Bartanusz · J. Dvorak · D. Grob
Spine Center, Schulthess Klinik, Zurich, Switzerland

alleviate respondent burden, and hence be practical for routine clinical use and quality management.

As discussed in Part 1 of this series [17], the psychometric properties of an index comprising these questions, now known as the “Core Outcome Measures Index” (COMI), have subsequently been documented by three independent research groups [10, 14, 15, 24]. However, whilst the availability and ease of administration of simple instruments such as the COMI can be expected to encourage clinicians to collaborate with surgical registries [2], interpretation of the scores thus derived requires a knowledge of what constitutes a meaningful (to the patient) score or change score after treatment. The provision of cut-off scores that clinicians can use to decide whether patients have improved or deteriorated is expected to improve the understandability and hence appeal of quality-of-life measures [22].

There is no “gold standard” methodology for estimating the minimal important difference, although most methods are grouped into two categories: distribution-based methods, which rely only on the statistical characteristics (distribution) of the obtained scores, and anchor-based methods, which examine the relationship between change scores on the target instrument (e.g. COMI) and some independent measure of “worthwhile or important change” [5]. The anchor-based methods appear to follow more intuitively meaningful principles, but they depend on the use of a valid external criterion to indicate “improvement.” There exists no “gold standard” for this and in reality there are often many grades of improvement that can be considered to carry clinical relevance. In practice, the patient’s rating of the “global outcome of treatment” is typically chosen as the external criterion [3, 12]. Although this measure does not constitute 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. It has been suggested that, in the absence of a true gold standard, the best one can do is to ensure construct validity of the criterion that is ultimately chosen for use [8], and global rating scales appear to demonstrate evidence of such validity [12, 15, 16]. Further, as previously highlighted [3], most people would be reluctant to label patients as improved or worse contrary to their personal rating of the global effect of treatment.

One of the original research studies documenting the validity of the COMI used an anchor-based approach to examine its minimal clinically important score-difference (MCID) for improvement (MCID_{imp}) [15]. However, the values derived (2–3 points, depending on the exact “anchor” used to indicate success) have never been confirmed in larger groups of patients, or in routine clinical practice. Further, no MCID for deterioration (MCID_{det}) has ever been determined for the COMI; indeed, given the

typical success rates for spinal surgery, relatively large sample sizes are needed to obtain sufficient cases with worsening to allow meaningful estimates of the MCID_{det}. This likely explains why many studies have examined the MCID_{imp} for various other spine outcome questionnaires (reviewed in Ref. [20]) but only two have (to our knowledge) attempted to evaluate the MCID for worsening after spinal surgery [6, 12].

The aim of the present study was to examine the prospectively measured changes in COMI score (from before to 12 months after surgery) in relation to Likert-scale ratings for the “global effectiveness of surgery” at 12 months follow-up in a large group of patients whose data were being collected in connection with an international surgical registry (SSE Spine Tango). In this way, the minimal clinically relevant change score (for both improvement and deterioration) in “everyday practice” could be determined.

Methods

The methods are described in detail in Mannion et al. [17]. Briefly, the study group comprised all German or English speaking patients undergoing spine surgery (both lumbar and cervical) in the Spine Center of our hospital from March 2004. Before and 12 months after surgery, patients were requested to complete a questionnaire containing the multi-dimensional COMI [10, 14, 15]. The COMI comprises a series of questions covering the domains of pain (back and leg pain intensity, each measured separately on a 0–10 numeric graphic rating scale), and function, symptom-specific well-being, general quality of life, social disability, and work disability (each on a 5-point Likert scale). At follow-up, in addition to the COMI questions, patients rated the global effectiveness of surgery (“how much did the operation help your back/neck problem?” 5-point Likert scale from “helped a lot” to “made things worse”).

All questionnaires were completed by the patient at home, ensuring that the information provided was free of any care-provider influence.

Statistical analyses

Descriptive data are presented as means and standard deviations (SD).

The COMI sum score was calculated as described in the original validation paper [15]; briefly, the items scored 1–5 [function, symptom-specific well-being, general QOL, disability (average of social and work disability)] were firstly re-scored on a 0–10 scale [(raw score – 1) × 2.5]. These items and the pain score (the highest value out of leg pain and back pain; already scored 0–10) were then averaged to provide a COMI index score ranging from 0 to 10.

Following the methods of Hagg et al. [12], the group mean $MCID_{imp}$ was given by subtracting the mean change score of the group of patients for whom the operation “did not help” from that of the group for whom the operation “helped.” Similarly, the group mean $MCID_{det}$ was given by subtracting the mean change score of the group of patients for whom the operation “did not help” from that of the group for whom the operation “made things worse.”

The sensitivity and the specificity of individual COMI change scores in “predicting” global effectiveness of surgery were examined using the receiver operating characteristic (ROC) method. 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 [8]. The ROC curve synthesizes information on sensitivity and specificity for detecting improvement (or deterioration) according to some dichotomized, 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 the change score [8]. 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 utilized as follows: for indicating “improvement” versus “no improvement” as a dichotomous variable; “*improvement*” included “helped a lot” and “helped,” and “*no improvement*” included “did not help” and “made things worse” (the category “helped only little” was not used here); for indicating “deterioration” versus “no deterioration” as a dichotomous variable, “*deterioration*” included “made things worse,” and “*no deterioration*” included all the other outcome categories above this (i.e. “did not help” and upwards). This method of dichotomizing for improvement or otherwise is similar to that described by Campbell et al. [6]. The area under the ROC curve (AUC) was interpreted as the probability of correctly discriminating between patients with “improvement” and “no improvement” (or “deterioration” and “no deterioration”), based on the change in COMI score. The AUC can range from 0.5 (no accuracy in discriminating) to 1.0 (perfect accuracy in discriminating). The ROC curve was used to determine the cut-off change score (MCID) that best indicated “improvement” ($MCID_{imp}$) and best indicated “deterioration” ($MCID_{det}$) [9], using the approach of minimizing “errors,” i.e. maximizing the sum of the specificity and sensitivity [1].

The analyses were conducted using Statview 5.0 (SAS Institute Inc., San Francisco, CA, USA) and MedCalc (MedCalc Software, Mariakerke, Belgium) and statistical significance was accepted at the $P < 0.05$ level.

Results

Minimal clinically important difference for group mean improvement and deterioration

The distribution of answers for global treatment effectiveness at the 12-month follow-up in the group of 3,056 patients has been reported in detail in Part 1 of this work [17]. Briefly, the operation “helped a lot” in 46.4% patients, “helped” in 28.1%, “helped only little” in 14.9%, “did not help” in 8.9%, and “made things worse” in 1.7%. The corresponding reductions in the mean COMI score (max possible reduction = 10 points) were 5.4 points (SD 2.5), 3.1 (SD 2.2), 1.3 (SD 1.7), 0.5 (SD 2.2) and -0.7 (SD 2.2), respectively.

Taking the group mean score difference between “operation helped” and “did not help” gave a group $MCID_{imp}$ of 2.6 points (3.1 minus 0.5 points); the corresponding value for the group $MCID_{det}$ was 1.2 points (0.5 minus -0.7).

Minimal clinically important difference for individual improvement and deterioration

The ROC area under the curve for COMI change scores predicting individual improvement was 0.88 (SE 0.008, 95% CI 0.87–0.90, $P < 0.0001$) and for deterioration 0.89 (SE 0.010, 95% CI 0.88–0.90, $P < 0.0001$; Fig. 1a, b), indicating that the COMI showed good discriminative ability in each case.

The cut-off with the greatest sensitivity for indicating improvement was a decrease of 2.2 points or more in COMI score (81% sensitivity and 83% specificity); the cut-off for distinguishing deterioration was an increase in COMI of 0.3 points or more (83% sensitivity and 88% specificity). In other words, a reduction in COMI score of at least 2.2 points would be needed to indicate (with 81% sensitivity and 83% specificity) that an individual had shown clinical improvement, whilst an increase in COMI score of 0.3 points or more would be needed to indicate (with 83% sensitivity and 88% specificity) deterioration.

Discussion

The present study sought to determine the MCID in the COMI score for indicating improvement and deterioration in a large group of patients undergoing spinal surgery for various different indications, mostly degenerative disorders of the spine. Previous studies in various groups of patients have shown that the COMI is responsive to change after spine surgery and that the standardized response mean or effect size for group change is comparable to that reported

Fig 1 Receiver operating characteristics curves used for identifying the minimal clinically important score-difference for improvement (a) and deterioration (b), 12 months after spine surgery. The point at which both optimal sensitivity and specificity are obtained, and which yields the location for the “cut-off” point for improvement (or deterioration), is given by the *open square* on the *solid line* at the point where it is closest to the *top left corner* of the *graph*

for other longer spine outcome questionnaires [10, 15]. Demonstrating that mean post-treatment scores are significantly different from pre-treatment scores and that change scores are greater in a group with a “good” outcome than in one with a “poor” outcome addresses the issue of a scale’s sensitivity to change, but it says nothing about its specificity [3, 9]. However, the concept of specificity to change is also important, since changes without clinical relevance may also occur in scale scores [9]. In order to better quantify the responsiveness of the COMI on an individual basis, the ROC method was used. The area under the ROC curve (AUC) for the ability of the COMI change score to predict improvement was 0.88. This can be interpreted as indicating that, if one takes two individuals at random—one with and without improvement according to the global effectiveness rating—the probability is 88% that the first individual will have a higher change score than the second. The corresponding value (AUC) for predicting deterioration was 0.89. The accuracy (or discriminatory ability) of tests with AUCs 0.50–0.70 is generally considered low, between 0.70 and 0.90, moderate, and over 0.90, high [23]. The AUCs for the COMI were comparable to or even higher than those previously reported in the literature for back-specific instruments such as the Oswestry Disability Index (AUC 0.82–0.85) [6, 13, 16], the Roland–Morris Disability Questionnaire (AUC 0.84) [16], or the pain intensity visual analog scale (AUC 0.88) [16] in patients undergoing spine surgery; they were also slightly higher than the AUC reported for the COMI during its initial validation studies in surgical patients (AUC 0.82) [15].

In the present study, individual “improvement” was predicted (with 81% sensitivity and 83% specificity) by a reduction in COMI score of 2.2 points or more. In other words, a 2.2-point reduction in COMI score would be considered to be the minimal difference for indicating that

an individual had shown clinical improvement. This is comparable to the individual $MCID_{imp}$ reported in the original COMI validation study (1.9–3.0, depending on the external criterion used for indicating “improvement”) [15] and is not dissimilar to the group $MCID_{imp}$ in the present study (2.6 points). In relation to the group mean pre-surgery COMI score of 7.5, an $MCID_{imp}$ of 2.2 points represents a reduction of approximately 30%. This compares favorably with the individual $MCID_{imp}$ reported for other back-specific outcome questionnaires (approximately 20–30% of their mean baseline scores [16, 18–20]). Few studies have examined the individual $MCID_{det}$ in patients undergoing spinal surgery [6], because doing so requires very large sample sizes to obtain sufficient cases with worsening. This is where the benefits of a huge data set, typical of those acquired in registries, can clearly be seen. Although a tiny proportion of patients (<2%) reported at 12 months that the operation “made things worse,” this still equated to 53 patients, which was considered sufficient to perform the ROC analysis and obtain good discrimination (AUC = 0.89) with narrow confidence intervals. Accordingly, individual “deterioration” was predicted (with 83% sensitivity and 88% specificity) by an *increased* COMI score after surgery of 0.3 points or more. This was somewhat lower than the group mean $MCID_{det}$ in the present study, given by the difference between the group mean scores for the categories “did not help” and “made things worse” (1.2 points). Interestingly, and in complete agreement with previous studies on back-specific as well as generic health measures [6, 12], the absolute size of the $MCID_{det}$ —determined either on an individual basis or a group mean basis—was lower than that for improvement. In other words, these outcome instruments appear to be less responsive to deterioration than to improvement. This was also seen when looking at the group effect size (mean change score/SD change score) for deterioration, which

was only 0.3 (−0.7/2.2) compared with 1.4 (3.1/2.2) for what we considered the minimal relevant improvement (operation helped). Statistical explanations for this phenomenon include the possibility of a notable floor effect for baseline values in the group that worsened (indicating that there would be “nowhere to go” if they deteriorated further), but that was also ruled out in the present study; the “worse” group did not show significantly different pre-op COMI scores compared with the rest of the group and they had adequate room for maneuver as regards their potential for demonstrating a large effect size for worsening (results not shown). Other, more clinical interpretations include the notion that it takes more to become improved than to be worse after treatment, or that the absence of improvement is considered to be deterioration [12]. Interestingly, in the present study, the outcome group “no change” was still associated with a mean reduction in COMI score of 0.5 (SD 2.2) points. Perhaps a small reduction in COMI score is unwittingly considered the minimum “reward” for undergoing an invasive intervention per se, and that only when changes exceed this amount do they begin to find a place on the scale of “success.” The corollary of this would be that the COMI change scores associated with a given outcome category are systematically “overestimated” by a small amount, such that after “calibration”—to make a global rating of “no change” equal to 0 for the COMI change score—the balance between responsiveness for improvement and worsening would be somewhat (though not entirely) redressed.

In presenting the MCIDs of the present study, we made no attempt to split up the patient population in relation to specific diagnoses, types of intervention, gender, age-groups, etc. Such analyses should, however, be performed once the register has grown sufficiently to provide adequately sized sub-groups. Future studies should examine whether the instrument is as responsive in these different sub-groups, and should further investigate why spine outcome measures in general are less good at identifying deterioration than improvement. The MCIDs reported here should be of assistance in planning clinical studies “nested” within spine surgery registries (e.g. [11]); however, as highlighted by others [12, 18, 20, 21], it must be borne in mind that the MCID should not be considered as an exact, fixed value, but as an approximate threshold. The MCID can vary depending on the patient group and their initial scores, the treatment under investigation, the method used for its determination, and the choices made in optimizing for sensitivity or specificity [18, 20, 21]. The strategies used to assess an instrument’s responsiveness always depend on some external criterion for rating “improvement,” and to perform ROC analyses this criterion must be dichotomous. However, as mentioned earlier, there exists no “gold standard” for assessing outcome. In the present

study, the 5-category Likert scale for treatment effectiveness (“how much the operation helped”) was used to generate dichotomous outcomes as the external criteria for use in the ROC analyses. We do not assert that this measure constitutes a definitive gold standard for assessing outcome, but it has at least been shown to have adequate construct validity [16] and can be expected to reflect the changes most important to the given patient as a result of the operation.

Finally, a somewhat provocative but nonetheless relevant issue within the general context of MCID-calculation for outcome questionnaires concerns the fact that the cut-off scores are established using the global outcome rating as the external criterion: this raises the question as to why the global outcome rating itself should not simply be used, as a considerably more straightforward, understandable and presumably valid (if serving as the “gold standard”) indicator of success. There are some keen advocates of the single global outcome question [12], and we also strongly support its use as an important complementary measure in outcome assessment (see the papers by Grob and colleagues, Porchet and colleagues, and Lattig and colleagues in this supplement); indeed, we are of the opinion that both types of data (prospectively measured scores and global outcome) have value, and can be used to different effect. Assessments of *current status* at repeated intervals throughout a course of treatment/follow-up are considered to be less influenced by recall bias. The data thus obtained allow many subsequent types of assessment to be made: the determination of baseline status; the examination of mean (and SD) change scores in groups of patients (considered useful for some purposes, e.g. meta-analyses, in comparing and summarizing the results of different trials); the comparison of score changes for groups with different indications or baseline values/disease severity; and perhaps also the establishment of meaningful change in specific outcome domains. In determining the sample size for trials that propose to use such prospectively collected data, it is important to know the score-difference that will be considered “worth finding” (i.e. the MCID), in relation to the variation in scores within the study groups. Similarly, in interpreting the findings of previous studies (that have not necessarily used a measure of global outcome), it is important to know the clinically relevant change score for the specific outcome instrument used. Ideally, this MCID threshold value is best determined using the data from very large groups of patients for whom the statistical model fits well (large AUC in ROC analysis; see “Results”). Being able to express the results not only as mean values, but also as the proportion of patients meeting a given threshold for change, should better satisfy the demand for improved and simplified reporting of clinical trial results [22]; this, in turn, can be expected to facilitate interpretation and hence

implementation of the findings of such trials in everyday clinical practice.

In conclusion, we have determined the MCID_{S_{imp}} and MCID_{det} in a large group of patients undergoing spinal surgery for mixed indications; these values are expected to provide essential information in both the planning (sample size calculations) and interpretation of the results (assessing clinical relevance) of clinical trials using the COMI. Further studies are encouraged, in even larger patient groups drawn from clinical practice or surgical registries, in order to investigate in greater depth the lower sensitivity to deterioration (than improvement) of spine outcome instruments, and to examine whether the MCIDs differ in different indications and sub-groups of spine surgery patients.

Conflict of interest statement None of the authors has any potential conflict of interest.

References

- Altman DG, Bland MJ (1994) Statistics notes: diagnostic tests 2: predictive values. *BMJ* 309:102
- Balagué F, Mannion AF, Pellise F, Cedraschi C (2007) Clinical update: low back pain. *Lancet* 369:726–728. doi:10.1016/S0140-6736(07)60340-7
- Beurskens AJHM, de Vet HCW, Köke AJA (1996) Responsiveness of functional status in low back pain: a comparison of different instruments. *Pain* 65:71–76. doi:10.1016/0304-3959(95)00149-2
- Bombardier C (2000) Outcome assessments in the evaluation of treatment of spinal disorders: summary and general recommendations. *Spine* 25:3100–3103. doi:10.1097/00007632-200012150-00003
- Brozek JL, Guyatt GH, Schunemann HJ (2006) How a well-grounded minimal important difference can enhance transparency of labelling claims and improve interpretation of a patient reported outcome measure. *Health Qual Life Outcomes* 4:69. doi:10.1186/1477-7525-4-69
- Campbell H, Rivero-Arias O, Johnston K, Gray A, Fairbank J, Frost H (2006) Responsiveness of objective, disease-specific, and generic outcome measures in patients with chronic low back pain: an assessment for improving, stable, and deteriorating patients. *Spine* 31:815–822. doi:10.1097/01.brs.0000207257.64215.03
- Deyo RA, Battie M, Beurskens AJHM, Bombardier C, Croft P, Koes B, Malmivaara A, Roland M, Von Korf M, Waddell G (1998) Outcome measures for low back pain research: a proposal for standardized use. *Spine* 23:2003–2013. doi:10.1097/00007632-199809150-00018
- 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. doi:10.1016/0021-9681(86)90038-X
- Deyo RA, Diehr P, Patrick DL (1991) Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. *Control Clin Trials* 12(Suppl):142S–158S. doi:10.1016/S0197-2456(05)80019-4
- Ferrer M, Pellise F, Escudero O, Alvarez L, Pont A, Alonso J, Deyo R (2006) Validation of a minimum outcome core set in the evaluation of patients with back pain. *Spine* 31:1372–1379. doi:10.1097/01.brs.0000218477.53318.bc discussion 1380
- Grob D, Bartanusz V, Jeszenszky D, Kleinstück FS, Lattig F, O'Riordan D, Mannion AF (2009) A prospective cohort study of two lumbar fusion techniques. *JBJS (Br)* (submitted)
- Hagg O, Fritzell P, Nordwall A, Group SLSS (2003) The clinical importance of changes in outcome scores after treatment for chronic low back pain. *Eur Spine J* 12:12–20
- Hashimoto H, Komagata M, Nakai O, Morishita M, Tokuhashi Y, Sano S, Nohara Y, Okajima Y (2006) Discriminative validity and responsiveness of the Oswestry Disability Index among Japanese outpatients with lumbar conditions. *Eur Spine J* 15:1645–1650. doi:10.1007/s00586-005-0022-7
- Mannion AF, Elfering A, Staerkle R, Junge A, Grob D, Dvorak J, Jacobshagen N, Semmer NK, Boos N (2007) Predictors of multidimensional outcome after spinal surgery. *Eur Spine J* 16:777–786. doi:10.1007/s00586-006-0255-0
- Mannion AF, Elfering A, Staerkle R, Junge A, Grob D, Semmer NK, Jacobshagen N, Dvorak J, Boos N (2005) Outcome assessment in low back pain: how low can you go? *Eur Spine J* 14:1014–1026. doi:10.1007/s00586-005-0911-9
- Mannion AF, Junge A, Grob D, Dvorak J, Fairbank JC (2006) Development of a German version of the Oswestry Disability Index. Part 2: sensitivity to change after spinal surgery. *Eur Spine J* 15:66–73. doi:10.1007/s00586-004-0816-z
- Mannion AF, Porchet F, Kleinstück F, Lattig F, Jeszenszky D, Bartanusz V, Dvorak J, Grob D (2009) The quality of spine surgery from the patient's perspective. Part 1. The Core Outcome Measures Index (COMI) in routine practice. *Eur Spine J*. doi:10.1007/s00586-009-0942-8
- Ostelo RW, de Vet HC (2005) Clinically important outcomes in low back pain. *Best Pract Res Clin Rheumatol* 19:593–607. doi:10.1016/j.berh.2005.03.003
- Ostelo RW, de Vet HC, Knol DL, van den Brandt PA (2004) 24-item Roland-Morris Disability Questionnaire was preferred out of six functional status questionnaires for post-lumbar disc surgery. *J Clin Epidemiol* 57:268–276. doi:10.1016/j.jclinepi.2003.09.005
- Ostelo RW, Deyo RA, Stratford P, Waddell G, Croft P, Von Korf M, Bouter LM, de Vet HC (2008) Interpreting change scores for pain and functional status in low back pain: towards international consensus regarding minimal important change. *Spine* 33:90–94
- Revicki DA, Cella D, Hays RD, Sloan JA, Lenderking WR, Aaronson NK (2006) Responsiveness and minimal important differences for patient reported outcomes. *Health Qual Life Outcomes* 4:70. doi:10.1186/1477-7525-4-70
- Schunemann HJ, Akl EA, Guyatt GH (2006) Interpreting the results of patient reported outcome measures in clinical trials: the clinician's perspective. *Health Qual Life Outcomes* 4:62. doi:10.1186/1477-7525-4-62
- Streiner DL, Cairney J (2007) What's under the ROC? An introduction to receiver operating characteristics curves. *Can J Psychiatry* 52:121–128
- White P, Lewith G, Prescott P (2004) The core outcomes for neck pain: validation of a new outcome measure. *Spine* 29:1923–1930. doi:10.1097/01.brs.0000137066.50291.da