Eur Spine J (2004) 13 : 301–313
DOI 10.1007/s00586-003-0665-1

REVIEW

U. Müller
M. S. Duetz
C. Roeder
C. G. Greenough

Condition-specific outcome measures for low back pain

Part I: Validation

Received: 1 December 2003
Accepted: 5 December 2003
Published online: 17 March 2004
© Springer-Verlag 2004

U. Müller (✉) · C. Roeder
Institute for Evaluative Research
in Orthopaedic Surgery,
University of Bern, Murtenstrasse 35,
P.O. Box 8354, 3001 Bern, Switzerland
e-mail: umueller@memced.org

M. S. Duetz
Institute for Social
and Preventive Medicine,
University of Bern, Bern, Switzerland

C. G. Greenough
Middlesbrough General Hospital,
Middlesbrough, UK

Abstract A literature review of the nine most widely used, condition-specific, self-administered assessment questionnaires for low back pain has been undertaken. General and historic aspects, reliability, responsiveness and minimum clinically important difference, external validity, floor and ceiling effects and available languages were analysed for the nine most-used outcome tools. When considering which condition-specific measure to employ, the present overview on assessment tools should provide the necessary information to define the technical aspects of the nine questionnaires. These criteria, however, are only part of the consideration. In part II the construction of these scales in relationship to the measurement domains will be evaluated.

Keywords Spine · Outcome · Assessment · Review · Low back pain

Introduction

The evaluation of therapies for low back pain requires consideration of a number of variables. A full evaluation is recommended to include a condition-specific disability measure, a general health measure (e.g. EQ-5D [22], WHODAS II [65], SF36 [67]), a pain measure (e.g. VAS [34]), a satisfaction measure and a measure of employment [7]. The present review is concerned with condition-specific measures.

Among a broad range of available tools, only a limited number of measurement instruments are generally known and frequently used. The application of a frequently used tool allows comparisons to be made between the study group and other populations. All currently available measures have flaws or restrictions regarding their construction, validation or application. In the absence of an ideal instrument [8], the choice of a commonly used measurement tool may be considered reasonable.

Apart from its currency of application, however, the availability, validity and responsiveness of the outcome scale are important criteria in choosing the appropriate measuring tool [45, 52, 58]. Other important, but sometimes underestimated criteria, are the characteristics of the individual questions and answers in the questionnaire [50, 58, 62]. Does the question focus on one domain or several domains? Are the sentences and questions unambiguous? Do they address performance or capacity? Are the offered answers precise and clear? Does the scoring system allow separate assessment of subscales?

The goal of this review is to compare the most common condition-specific assessment instruments for low back pain and to analyse each of these instruments separately, addressing these issues. This first paper addresses internal consistency, reliability, external validation, floor and ceiling effects, responsiveness and availability. In part II, the questionnaires will be discussed in relation to their face and construct validity and domains of measurement. Following this analysis, we aim to provide a basis for choice

source: https://doi.org/10.7892/boris.117744 | downloaded: 12.10.2024

when considering a self-administered assessment tool in the field of low back problems.

Methods

Using the Internet, various medical search sites were addressed in order to find condition-specific, self-administered outcome measurements used in spine surgery (http://www.BioMedNet.com, http://www.PubMed.com, http://www.aaos.org, http://www.outcomes-trust.org, http://www.qlmed.org, http://www.hoi-stratishealth.org, http://www.isoquol.org, http://www.update-software.com/Cochrane, http://www.mapi-research-inst.com). Out of 82 condition-specific outcome scales, the nine most commonly used in the literature were selected (Table 1).

To quantify the use of each of these outcome tools, the number of studies evaluating or using it were determined. Measures used ten times or less in the literature were excluded from further analysis. General measures such as the SF36 [67] were not included in the search. Isolated pain measures were also excluded.

The nine scales examined were: (1) the Oswestry Disability Index (ODI), (2) the Roland–Morris Disability Questionnaire (RMDQ), (3) the Low Back Outcome Score (LBOS), (4) the Quebec Back Pain Disability Scale (QBPDS), (5) the Million Visual Analogue Scale (MVAS), (6) the Aberdeen Low Back Disability Scale (ALBDS), (7) the NASS Lumbar Spine Outcome Assessment Instrument (NASS LSO), (8) the Low Back Pain Rating Scale (LBPRS), and (9) the Waddell Disability Index (WDI).

The construction of outcome scales requires a number of considerations [58]. The measures were examined for general characteristics, reliability, internal consistency, responsiveness, correlation with other measures, floor and ceiling effects and available languages.

General characteristics

The population from which the score was developed, number of items, items scored, whether the measure produces one single score or is divided into subscales and a brief description of the domains are provided.

Reliability

There are a variety of ways of examining the reproducibility of a measure administered on different occasions. Test–retest reliability is the most important. It is measured best by using tests of agreement such as the kappa test [6, 45, 58]. The Pearson correlation coefficient [6] is a measure of correlation and, although commonly used, is a less reliable measure. Pearson correlation values should exceed 0.8 and kappa values should exceed 0.5. Another measure is the Bland–Altman plot [6]. This describes the spread of the score values within the same individuals between the test and the retest examination and provides a 95% confidence interval.

Internal consistency

Measures of internal consistency are based on a single administration of the outcome measure. If the outcome measure has a relatively large number of items addressing the same dimension, such as measures of physical function, it is reasonable to expect that scores on each item would be correlated with scores on all other items. Thus, if the internal consistency is low, the different items should not be summed, because they measure different domains. Internal consistency is predominantly measured by Cronbach's alpha correlations [14]. Values above 0.8 are acceptable.

Responsiveness to changes

The minimum clinically important difference (MCID) is the value of the change in the score which equates to the smallest change in the condition of interest the patient can detect. Responsiveness can also be evaluated using the receiver operating characteristic (ROC) curve which is constructed by calculating the sensitivity (true positive rate) and specificity (true negative rate) of the cut-off point for each of the possible score values [58]. An index of the "goodness" of the questionnaire is the area under this curve (AUC), which is usually abbreviated as D'. A poorly discriminating questionnaire has an area of 0.5 and a perfect test has an AUC of 1.0 [58].

External validation

Comparison of a new score with existing scores allows assessment of its performance against known measures, particularly in selection of measurement domains, responsiveness and floor and ceiling effects.

Floor and ceiling effect

Floor and ceiling effects describe the percentage of subjects which have maximal or minimal points in the score [7, 58]. Here the measure is inefficient in discriminating between subjects. A similar problem occurs when the results are skewed in a certain region. Floor and ceiling effects may be observed if a measure developed in one population, e.g. severely disabled subjects in a pain clinic, is used in a very different population, e.g. attenders in primary care.

The nine questionnaires

The Oswestry Disability Index (ODI)

General characteristics

The score was initiated in 1976 in a specialist referral clinic with a large number of chronic low back pain pa-

Table 1 Characteristics of the nine chosen condition-specific questionnaires

Characteristic	ODI	RMDQ	LBOS	QBPDS	MVAS	ALBDS	NASS LSO	LBPRS	WDI
Existing since	1980	1982	1992	1995	1982	1994	1996	1995	1984
Items	10	24	13	20	15	19	62	21	9
Completion time	5	10	5	10	10	10	21	15	5
Med Line used	117	103	23	30	29	71	21	14	64

Table 2 Content and question-and-answer characteristics of the chosen evaluation tools using the ICF classification

Characteristic	ODI	RMDQ	LBOS	QBPDS	MVAS	ALBDS	NASS LSO	LBPRS	WDI
Assessment of									
Pain	1a	1a	1a	–	1a	1a	1a	1a	–
Sleep	1a	1a	1a	1a	1a	1a	1a	1a	1a
Self-care	2	2	–	2	–	2	–	–	–
Walking	2	2	2	2	2	2	2	2	2
Sitting	2	2	2	2	2	2	2	2	2
Standing	2	2	–	2	2	2	2	–	2
Lifting	2	2	–	2	–	–	2	2	2
Sex life	2	–	2	–	–	–	2	–	2
Travelling	2	–	2	–	–	–	2	–	2
Social life	3	–	–	–	3	–	3	3	3
Work	–	2	2	–	2	–	–	2	–
Dressing	–	2	2	–	–	–	2	2	2
Sport	–	–	2	2	–	2	–	–	–
Stairs	–	2	–	2	–	–	–	2	–
Housework	–	2	2	2	–	2	–	2	–
Resting	–	2	2	–	–	2	–	–	–
Appetite	–	1a	–	–	–	–	–	–	–
Need of help	–	2	–	–	–	–	–	–	–
Psychological factors	–	–	–	–	–	–	–	1a	–
Need of treatment	–	–	X.9	–	–	–	–	–	–
Need of medicaments	–	–	X.8	–	X.8	X.8	–	–	–
Car driving	–	–	–	2	–	–	–	2	–
Throwing	–	–	–	2	–	–	–	–	–
Stiffness	–	–	–	–	1a	–	–	–	–
Twisting	–	–	–	–	1a	–	–	–	–
Bending	–	–	–	–	–	1a	–	–	–
Loss of feelings	–	–	–	–	–	1b	1b	–	–
Leg weakness	–	–	–	–	–	1b	1b	–	–
Special features	–	–	A VAS is used for pain, but not for the rest	–	–	–	SF36 and health survey item 18–28 are included	VAS for pain	–
Question's targets	1 clear target	1–2 targets	1 target	1–2 targets	1–2 targets	1–2 targets	1–2 targets	1–2 targets	1–2 targets
Answer levels	2 or more	Yes/no	1–2	1	1	1 (some more)	1–2	1	1
Answer type	Text	Yes/no	Text	Scale	Scale	Text	Text	Scale	Yes/no
Answer scale	Scaled text	Yes/no	Scaled text	0–5	0–11	Scaled text	Scaled text	0–3	Yes/no
Scoring points	0–50	0–24	0–75	100	150	0–100	0–102	0–90	0–9

Table 3 Technical data of the nine condition-specific questionnaires in comparison

	ODI	RMDQ	LBOS	QBPDS	MVAS	ALBDS	NASS LSO	LBPRS	WDI
Floor effect	Yes [49]	Unknown	0% [27]	Unknown	Unknown	Unknown	Unknown	Unknown	16.1% [27]
Ceiling effect	Unknown	Present [59]	4.6% [27]	Unknown	Unknown	Unknown	Unknown	Unknown	1/0% [27]
Retest Pearson correlation	0.83 [29]	0.91 [55]	0.92 [32]	0.67 [26]	0.92 [46]	0.94 [56]	0.85–0.99 [15]	–	0.73–0.9
	0.78 [26]	0.88 [35]	–	–	–	–	–	–	–
	0.91 [38]	0.83 [17]	–	–	–	–	–	–	–
Kappa values	–	–	0.51–0.86 [32]	–	–	–	0.85–0.97 [15]	–	0.27–1.0
Bland–Altman plot	–	–	11.6 [32]	–	–	26.6 [56]	–	–	–
ICC	0.91/0.9 [26]	0.53 [16]	–	0.92/0.55 [26]	–	–	–	–	0.74 [16]
	0.84 [16]	0.86 [59]	–	0.84 [16]	–	–	–	–	–
	0.89 [1]	0.91 [38]	–	0.92 [38]	–	–	–	–	–
Cronbach's alpha	0.76 [25]	0.89/0.92 [33]	0.85 [32]	0.96 [38]	0.93 [46]	0.8 [56]	0.88 [15]	0.98 [44]	–
	0.81 [38]	0.91 [61]	–	–	–	–	–	–	–
	0.77/93 [33]	–	–	–	–	–	–	–	–
Spearman rank analysis	0.59 [60]	0.6 [60]	–	–	–	–	–	–	–
ROC with an AUC value of	0.76 [4]	0.78/0.74	–	0.87/0.74	–	–	–	–	0.76 [16]
	0.94 [26]	0.77 [16]	–	0.74 [16]	–	–	–	–	0.87 [26]
	0.78 [16]	–	–	0.93 [38]	–	–	–	–	–
Norman–Steiner sensitivity analysis	0.15 [38]	0.2 [38]	–	0.26 [38]	–	–	–	–	–
Standardized response mean	0.36 [59]	0.5 [59]	–	–	0.49 [16]	–	–	–	0.35 [16]
	0.52 [16]	0.55 [16]	–	–	–	–	–	–	–
MCID	16 [24]	5.2 [53]	7.5 [32]	14 [38]	–	–	–	–	2.8 [16]
	15 [16]	8.6–9.5 [16]	–	19 [16]	–	–	–	–	–
	6 [26]	4 [59]	–	15 [26]	–	–	–	–	–
Compared with	RMDQ [23, 25, 45]	ODI [16, 33, 42]	ODI [27, 63]	RMDQ [16, 38]	ODI [46, 70]	SF36 [56]	SF36 [15]	ODI [11]	RMDQ [16]
	QBPDS [16, 26, 38]	QBPDS [16, 38]	WDI [27]	ODI [16, 26, 38]	RMDQ [16]	–	–	–	QVPDS [16]
	WDI [16, 63]	WDI [16]	SF36 [63]	SF36 [16, 38]	WDI [16]	–	–	–	ODI [16]
	LBOS [27]	SF36 [16]	–	–	–	–	–	–	LBOS [27]
	SF36 [16]	–	–	–	–	–	–	–	SF36 [16]
	MVAS [70]	–	–	–	–	–	–	–	MVAS [16]
	LBPRS [11]	–	–	–	–	–	–	–	–

tients. The index was designed as a measure for both assessment and outcome. Version one was published and validated in 1980 [24] using a sample of 25 patients suffering from acute low back pain. A larger validation was published in 1994 by Stratford on a population with musculoskeletal LBP [60]. The ten items can be completed in approximately 5 min and scored in less than 1 min. No sub-scoring is reported. The administration is easy.

The ODI was further developed and validated and is now available in version 2.0 [23]. A modified ODI is used in the NASS [15] questionnaire (see below) and another modification was used in England (computer interview). Version 2.0 is now recommended [1, 53] and no permission for its use is required. The questionnaire focuses on abilities (personal care, lifting, walking, sitting, standing, sleeping, sex life, social life and travelling in combination with pain) and on pain level. However, important items considering the ability to work, need for help or items about environmental factors are not included (Table 2).

Reliability

Reliability has been tested in the literature (Table 3) using only test–retest correlations (Pearson correlation) in a small population (n=22) [26, 30, 38]. Kappa values or Bland–Altman plots are not available.

Internal consistency

Version 2.0 of the ODI shows an acceptable internal consistency, with Cronbach's alpha between 0.76 and 0.8716 [25, 33, 38].

Responsiveness

The ROC was shown to be between 0.76 and 0.78 [4, 16, 60], which is acceptable (but not as good as the Roland–Morris score [16, 60]). Several studies show a good correlation between the severity of pain and disability on the one hand and patients satisfaction on the other [26, 60, 63], even in patients with cervical problems [70]. The ODI is able to detect clinical change 3 weeks after surgery [56] and the MCID varies between 6 and 16 points [4, 16, 26, 60]. Using mean values, Beurkens [4] adopted the method advocated by Cohen [12] to determine the effect size and obtained values of 0.8 for patients who had improved and 0.04 for patients whose condition had not changed. However, Taylor found that the ODI was more sensitive to patients who had improved and less sensitive for patients whose condition had remained unchanged [63].

External validation

The ODI was compared with the RMDQ [16, 33, 42], the LBOS [27], the QBPDS [16, 26, 37, 38], the LBPRS [16], the MVAS [70] and the SF36 [16] (see below).

Floor and ceiling effects

Floor effects have been demonstrated in non-surgical patients [49], demonstrating that the ODI is less sensitive in less disabled patients.

Languages

The ODI is validated in English [24], German [2], French [21], Finnish [29] and Greek [9] (Table 4). Translations in several other languages do not appear to be validated.

The Roland–Morris Disability Questionnaire (RMDQ)

General characteristics

The RMDQ was derived from the Sickness Impact Profile, of which 24 out of 136 items were selected. The questionnaire was designed as a self-reporting measure for both assessment and outcome, and was published in 1983 [54]. First validation was done using a LBP population in a general practice treated with pain medication. In the first version a six-point pain rating scale was included. Recently, the authors recommended the use of the pain scale of the SF36 [53]. Despite several published modification proposals, [20, 49, 61, 64]the original version of the RMDQ is favoured by an international expert group [18].

The questions in the RMDQ focus consistently on disabilities related to the back and the answers are dichotomous: yes/no. This might cause subtle changes in the functional abilities of patients to remain undetected. The questionnaire can be completed in a maximum of 5 min and an un-weighted score can be calculated in less than 1 min. No sub-scoring is reported and administration is very easy. The questions deal with body functions (pain, sleeping and appetite) as well as activities (self care, walking, sitting, standing, lifting, work, dressing, stairs, housework and resting), but no environmental questions are asked (social life, need of help etc.) (Table 2).

Reliability

The test–retest reliability using Pearson correlation lies between 0.81, 0.88 and 0.91 [17, 35, 54]. Kappa values or Bland–Altman plots are not available.

Table 4 Available languages for all the nine condition-specific outcome scores

	ODI	RMDQ	LBOS	QBPDS	MVAS	ALBDS	NASS LSO	LBPRS	WDI
Original language	English [24]	English [54]	English [27]	English [38]	English [46]	English [56]	English [15]	Danish [44]	English [66]
Validated languages	Finnish [29, 30]	French [13]	–	Dutch [57]	–	Chinese [43]	German [51]	–	–
	French [21]	German [68]	–	French [38]	–	–	Italian [48]	–	–
	German [2]	Greek [10]	–	–	–	–	–	–	–
	Greek [9]	Portuguese [47]	–	–	–	–	–	–	–
	–	Spanish [40]	–	–	–	–	–	–	–
	–	Swedish [35]	–	–	–	–	–	–	–
	–	Turkish [41]	–	–	–	–	–	–	–
Unvalidated languages	Danish	Flemish	German	–	–	–	–	–	French
	Dutch	Czech	Spanish	–	–	–	–	–	–
	Norwegian	Italian	–	–	–	–	–	–	–
	Spanish	Polish	–	–	–	–	–	–	–
	Swedish	Romanian	–	–	–	–	–	–	–

Internal consistency

The RMDQ seems to be a reliable evaluation tool with an internal consistency between 0.77 and 0.93, [33, 61] and with equal [42, 52] or even slightly superior [33] consistency in comparison with the Oswestry questionnaire (Table 3).

Responsiveness

The RMDQ seems to detect changes over time slightly more sensitively than the Oswestry scale [16, 42], provided that the initial score is in the range between 4 and 20 [59]. Beaton suggests that a change of 5 points in the RMDQ can be interpreted as a significant improvement or decline in patients' outcome [3]. This value comes close to the figure of Davidson [16], who defined the MCID to be 5.2 points.

In the French version, Dionne finds a poor relation between the RMDQ and the work status of patients [19]. When comparing the ODI with the RMDQ, Baker et al. [1] found that the RMDQ is more sensitive in patients with mild disabilities and the ODI discriminates better in more severe disabilities. This means that the RMDQ and the ODI are not linearly related.

In contrast, Yang [69] and Leclaire [42] found a linear but moderate Pearson correlation coefficient of r=0.72 and 0.64, respectively. Hsieh [33] compared the two measures in four treatment groups and found a poor correlation (r=0.51–0.61) at the initial visit and a high correlation after treatment (r=0.85–0.92). The latter suspected that both instruments may have a different performance at different levels of pain. Davidson [16] found similar responsiveness when comparing the RMDQ with the SF36, the ODI, the QBPDS and the WDI – but the RMDQ showed an insufficient reliability (Table 3).

Floor and ceiling effects

A ceiling effect is reported [59].

Languages

The RMDQ is translated into several languages (Table 4). Many of these translations do not appear to be validated.

The Low Back Outcome Score (LBOS)

General characteristics

The questionnaire was first published in 1992 and tested using 274 consecutive patients with low back pain [27].

The score was designed as a self-reporting measure for both assessment and outcome. The 13 items can be completed in approximately 5 min and scored in about 1 min. The answering possibilities of each item are scaled. For pain, an eleven-point scale ranging from "no pain" to "maximum pain possible" is used. The other items use a four-point scaled text. The total score gives different weights to different questions. For example, normal abilities in sport score 9 points, normal abilities in sex life score 6 points and normal abilities in dressing score 3 points. Although this weighting has to be conducted with care, the administration is easy. All aspects of the ICF classification [28] are considered, which is the strength of this questionnaire (for the questionnaire's content see Table 2). Slight extensions to the questionnaire have been reported [32].

Reliability

The test–retest reliability is high, with a Pearson correlation of 0.92 and kappa values between 0.51 and 0.8645. The Bland–Altman plot shows that test–retest scores can change by up to 11.6 points [32]. This scale is the only one to have been subjected to this more stringent criterion.

Internal consistency

Holt reported a good internal reliability with a Cronbach's alpha of 0.85 in an English population and 0.79 in an Australian population [32].

Responsiveness

The MCID is reported to be 7.5 points [36].

External validation

The LBOS correlates well with the ODI (r=0.87), the WDI (r=0.74) and the Waddell physical impairment rating (r=0.63) [32]. Taylor et al. [63] made a comparison between the LBOS, the SF36 and the ODI, and found a Spearman rank correlation of 0.64 for the ODI, 0.62 for the LBOS, 0.56 for the SF36 PF and 0.54 for the SF36 PCS. They demonstrated that the condition-specific questionnaires (ODI and LBOS) showed a greater overall responsiveness, but the SF36 PF was more sensitive to changes than both of the specific questionnaires.

Floor and ceiling effects

Values for floor and ceiling effects have been defined in a non-surgical population by Greenough [27], and are 0%/0.8% respectively in a compensated population and 0%/8.3% in a non-compensated low back pain population.

Languages

The LBOS has been validated in English [27, 32] and translated into German and Spanish.

The Quebec Back Pain Disability Scale (QBPDS)

General characteristics

The QBPDS was validated on a back pain population and published in 1995 [38]. The questions were designed using a conceptual model. Item selection was done using factor analysis for 46 disability items. Twenty items were selected and tested for reliability (Table 2).

The QBPDS measures only functional disability (self care, walking, sitting, standing, lifting, sport, stairs and housework) and sleep. Pain has to be evaluated with other tools. Items about social life, sex life and need for help are not included. Nevertheless, the items give a comprehensive view of the patient's disabilities because questions about easy as well as more difficult functional abilities are asked. The questionnaire can be filled out in about 5–10 min and scored in about 2 min. Sub-scores are not reported and administration is easy.

Reliability

Test–retest reliability was analysed using an intraclass correlation coefficient (ICC) (0.92) [38] (Table 3) but kappa values or Bland–Altman plots are not available.

Internal consistency

Internal consistency was measured using the Cronbach's alpha (0.96) [38] and is satisfactory.

Responsiveness

The MCID lies between 14 and 19 points [16, 37]. The Norman-Steiner sensitivity coefficient [38] was 0.19 after a six month period, and the sensitivity in relation to a self-rated change of disability was good.

External validation

Kopec [38] correlated the questionnaire in French and English with the RMDQ (r=0.81 and 0.72), ODI (r= 0.83

and 0.77), and SF36 physical functioning (r=0.77 and 0.67), and against a single-item pain scale with a seven-point Likert scale (r=0.54 and 0.51). Fitz [26] found that the modified ODI showed higher levels of test–retest reliability and responsiveness compared with the QBPDS. The MCID was approximately 15 points for the QBPDS (and approximately 6 points for the modified ODI). Davidson [16] found similar responsiveness when comparing the QBPDS with the SF36, the ODI, the RMDQ and the WDI, but the QBPDS showed a better reliability than the RMDQ and the WDI (Table 3).

Floor and ceiling effects

Values for floor and ceiling effects are not available in the literature.

Languages

The QBPDS is validated in English, French [38] and Dutch [57] (Table 4).

The Million Visual Analogue Scale (MVAS)

General characteristics

The questionnaire was first published in 1982 and validated on patients with chronic low back pain [46]. The 15 items focus on body functions (pain, sleep, stiffness and twisting), on activities (walking, sitting, standing and work) and on social life. Questions on self care, sex life lifting, housework etc. are not included. The questionnaire offers a 100 mm visual analogue scale as the answering method. Its administration is easy, the completion time is around 5–10 min and scoring requires 2–3 min.

Reliability

The Pearson correlation is 0.84–0.94 [46] (Table 3). Kappa values or a Bland–Altman plot are not available.

Internal consistency

Internal consistency shows a high Cronbach's alpha of 0.93 [46].

Responsiveness

The MCID or ROC are not available.

External validation

Zoega found a good correlation and a similar reliability between the MVAS and the ODI [70].

Floor and ceiling effects

Values for floor and ceiling effects are not available in the literature.

Languages

The MVAS is validated in English [46] (Table 4).

The Aberdeen Low Back Disability Scale (ALBDS)

General characteristics

The ALBDS was first published by Ruta in 1994 [56] and validated using a random sample of the general population. The 19 questions focus on body functions (pain, sleep and bending) and activities (self care, walking, sitting, standing, sport, housework and resting), and include questions about body structures (loss of feeling, leg weakness). Answer categories vary between three and six categories. Questions on environmental factors, lifting, sex life, work, dressing etc. are not included. The questionnaire is easy to administer, can be completed within 5–10 min and scoring can be done within 3 min. Because of the various answering scales, the ALBDS gives variable weights to the different questions.

Reliability

The ALBDS showed an acceptable test–retest correlation of 0.94 (Pearson correlation) [56]. Kappa values or Bland–Altman plots are not available.

Internal consistency

Internal consistency is acceptable with a Cronbach's alpha value of 0.8 (Table 3) [56].

Responsiveness

The MCID or ROC values are not available. The ALBDS is able to detect clinical change 2 weeks after surgery [56]. Ruta reports that the ALBDS is more sensitive than the SF36 regarding the measurement of the local outcome [56].

External validation

Ruta compared his ALBDS with the SF36 and found a rather low correlation between these two instruments (e.g. r=0.56 for SF36 physical functioning) [56]

Floor and ceiling effects

Values for floor and ceiling effects are not available in the literature.

Languages

The ALBDS is validated in English [56] and Chinese [43]. (Table 4).

The NASS Lumbar Spine Outcome Assessment Instrument (NASS LSO)

General characteristics

The NASS LSO was first published by Daltroy in 1996 [15] and is based on a consensus of the North American Spine Society. Validation was done in a cross-sectional study of a convenience sample of lumbar spine patients (n= 136). The NASS LSO questionnaire is made for assessment and outcome measurements. It considers all aspects of the ICF classification, e.g. demographic data (age, sex, race, education and insurance information), medical history (co-morbidities, past surgeries etc.), body functions (pain, neurogenic symptoms etc.) and employment history.

The questionnaire construct includes the SF36, a modified ODI and a modified employment assessment published by Bigos [5]. For follow-up assessment the questionnaire is slightly modified. It takes 20–25 min to fill in the form and the scoring is complex. To score the SF36, a special algorhythm is used. Sub-scores are extractable (modified ODI, SF36, pain and disability scale, neurogenic symptoms scale, job exertion scale, expectation and satisfaction scale). Although the NASS LSO is not a condition-specific questionnaire in a strict sense, we included this tool in our analysis because it contains condition-specific measures.

Reliability

The reliability tests for the condition-specific parts of the NASS LSO show a test–retest reliability (Pearson correlation) of 0.96 for pain and disability and 0.81 for neurogenic symptoms. The ICC is 0.97/0.85 for the two subscales [15].

Internal consistency

Cronbach's alpha is 0.93 in both measures [15].

Responsiveness

The MCID was not found in the literature.

External validation

The NASS LSO pain and limitation scale correlates with other measures of the same phenomena (pain VAS r=0.84, SF36 pain subscale r= 0.66 and SF36 physical limitation subscale r=0.75). The neurogenic symptoms correlated on a lower level with the three mentioned variables (r=0.60, 0.43 and 0.49 respectively) [15]. Even though the ODI is included, a correlation between the NASS LSO disability and pain scale in comparison with the original ODI could not be found in the literature.

Floor and ceiling effects

Values for floor and ceiling effects are not available in the literature.

Languages

The NASS LSO is validated in English [15], German [51] and Italian [48] (Table 4).

The Low Back Pain Rating Scale (LBPRS)

General characteristics

The LBPRS was first published and validated on different populations by Manniche in 1994 [44]. It consists of a disability index with 15 items (walking, sitting, lifting, work, dressing and car riding), and a pain assessment index with six questions. Questions about self care, lifting, standing, sex life and sport are not included. It can be filled out in about 15 min and scored within 3–5 min. Sub-scores for low back pain, leg pain, disability and physical impairment are extractable.

Reliability

Data on test–retest reliability are not available.

Internal consistency

Internal consistency was calculated using Cronbach's alpha with values between 0.89 and 0.95 for the sub-scores and 0.98 for the entire LBPRS [44].

Responsiveness

The MCID was not found in the literature.

External validation

Christensen et al. [11] could show a correlation with r=0.82 between the ODI and the LBPRS 18 months after therapy. Despite detailed clinical assessment of the patients involved, the clinical results were not correlated with the values of the two outcome measures.

Floor and ceiling effects

Values for floor and ceiling effects are not available in the literature.

Languages

The LBPRS is available in English and validated in Danish [44] (Table 4).

The Waddell Disability Index (WDI)

General characteristics

The WDI was first published in 1984 [66] and validated on a chronic low back pain population. It is a brief nine-item scale focussing on disabilities (walking, sitting, standing, lifting, sex life, travelling and dressing), on body functions (pain, sleep) and on social life. Questions about work, self care and sports are not included. The questionnaire is easy to administer, can be completed in 5 min and scored in about 1 min.

Reliability

The test–retest reliability of the different items show kappa values between 0.27 (unacceptable low) and 1.0 (very high) [27].

Internal consistency

Internal consistency seems to be rather low with an ICC of 0.74 [16], and Cronbach's alpha values are not available.

Responsiveness

The ROC is acceptable [16, 26] and the MCID is 2 [16]. The WDI is able to detect clinical change 4 weeks after surgery [56].

External validation

Davidson analysed the WDI against the RMDQ, ODI and the QBPDS [16] and found a lower ICC in the WDI (0.74) than in the ODI (0.84) but a higher ICC in the WDI than in the RMDQ (0.53). The ROC was similar in all four scores but the standardized response mean was lowest in the WDI (0.35). The MCID was similar in all four scores.

Floor and ceiling effects

Values for floor and ceiling effects have been defined in conservatively treated patients by Greenough [27] and are 2.2%/5.9% in a compensation group and 0% and 27% in a non-compensation group.

Languages

The WDI is validated in English [66] and Spanish. A French translation is available, but has not been validated [31] (Table 4).

Discussion

If a specific outcome questionnaire has to be chosen, five main aspects should be considered: First: Is the questionnaire reliable? Second: Is it responsive? Third: Can the results be compared with the literature? Fourth: Is the questionnaire available in the target language? And fifth: Is the questionnaire easy to complete and to score?

1. Reliability

The reliability of a questionnaire is the most important quality a questionnaire has to fulfil. All the nine mentioned questionnaires went through a validation process. The LBOS [27] was particularly stringently validated and the QBPDS [26, 38] and the ALBDS [56] have also passed a

serious validation process. The ODI and the RMDQ went through several validation processes and have been used for comparison in many studies. Although Davidson found an insufficient ICC for the RMDQ [16], this was in contradiction to other studies [37], and all the nine questionnaires reviewed may be considered to have passed the minimum acceptable validation standard.

2. Responsiveness in the study population

If it is important to know how subjects have improved following an intervention, then this figure is crucial. Figures are available for the ODI [4, 16, 21, 26], the RMDQ [16, 53, 59], the LBOS [32], the QBPDS [16, 26, 38], and the WDI [16]. For the other questionnaires the MCID was not available.

3. Comparability

Comparing the results with other studies is one of the important goals in quality assessment. In an optimal study design, patients are randomised in two groups (treatment group/control group). This procedure is sometimes not possible and the subjects have to be matched with a control group from the literature. In this case, the same questionnaire as used in the literature must be chosen. A broad comparability is assured if the Oswestry [24] or Roland–Morris scale [54, 55] is used. Nevertheless, this comparability should not be overestimated because of differences in culture, patients' collectives etc.. If randomisation within the study group is possible, then comparability of the results with the literature is less crucial. In this situation, the LBOS offers a short and comprehensive measure.

4. Availability

For non English speaking countries the availability of a questionnaire is an important factor. The RMDQ followed by the ODI have the largest number of translations. Target languages are defined in Table 4. A more scientific approach is to define the optimal questionnaire using the criteria 1–3. If the optimal questionnaire is not available in the target language, it should be translated and validated before use.

5. Ease of administration

This issue is of some importance in self-administered questionnaires, to avoid question fatigue and increase compliance. Easy and short questionnaires like the RMDQ, LBOS, QBPDS, MVAS or the WDI, give fewer opportunities for errors and are easier for data analysis.

Conclusion

There is no gold standard for quality or outcome assessment in low back therapies. In order to define the optimal questionnaire, its reliability and responsiveness in the study population must be considered. Only the ODI, the RMDQ, the LBOS, the QBPDS and the WDI provide the crucial data on the MCID.

The present overview on the reliability and responsiveness of condition-specific assessment tools should provide the necessary information to define the optimal questionnaire to be used in any proposed study. However, the circumstances of use, domains of interest, construct validity of the instrument as well as possible score bias are of crucial importance and will be considered in part II.

References

1. Baker D, Pynsent P, Fairbank J (1989) The Oswestry Index revisited. In: Roland M, Jenner J (eds) Back pain: new approaches to rehabilitation and education. Manchester University Press, Manchester, UK, pp 174–86
2. Basler H, Jakle C, Kroner-Herwig B (1997) Incorporation of cognitive behavorial treatment into the medical care of chronic low back patients: a controlled randomized study in German pain treatment centers. Patient Educ Counsel 31:113–124
3. Beaton D (2000) Understanding the relevance of measured change through studies of responsiveness. Spine 25: 3192–3199
4. Beurskens AJ, de Vet HC, Koke AJ (1996) Responsiveness of functional status in low back pain: a comparison of different instruments. Pain 65:71–76
5. Bigos SJ, Battie MC, Spengler DM, Fisher LD, Fordyce WE, Hansson TH, Nachemson AL, Wortley MD (1991) A prospective study of work perceptions and psychosocial factors affecting the report of back injury. Spine 16:1–6
6. Bland J, Altman D (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 8:307–310
7. Bombardier C (2000) Outcome assessment in the evaluation of treatment of spinal disorders. Summary and general recommendations. Spine 25:3100–3103
8. Bombardier C (2000) Spine focus issue introduction. Spine 25:3097–3099
9. Boscainos P, Sapkas G, Stilianesi E (1999) Clinical relevance of specific parameters isolated within the Oswestry and Roland Morris functional disability index. J Bone Joint Surg Br 81 [Suppl]: 239
10. Boscainos PJ, Sapkas G, Stilianessi E, Prouskas K, Papadakis SA (2003) Greek versions of the Oswestry and Roland–Morris disability questionnaires. Clin Orthop 40–53
11. Christensen TH, Bliddal H, Hansen SE, Jensen EM, Jensen H, Jensen R, Bay H (1993) Severe low-back pain. I: Clinical assessment of two weeks conservative therapy. Scand J Rheumatol 22:25–29

12. Cohen J (1977) Statistical power: analysis for the behavioural sciences. Academic, New York, pp 1–27
13. Coste J, Le Parc JM, Berge E, Delecoeuillerie G, Paolaggi JB (1993) [French validation of a disability rating scale for the evaluation of low back pain (EIFEL questionnaire)]. 60:335–341
14. Cronbach L (1951) Coefficient alpha and the internal structure of tests. Psychometrika 16:297–334
15. Daltroy LH, Cats Baril WL, Katz JN, Fossel AH, Liang MH (1996) The North American spine society lumbar spine outcome assessment instrument: reliability and validity tests. Spine 21: 741–749
16. Davidson M, Keating JL (2002) A comparison of five low back disability questionnaires: reliability and responsiveness. Phys Ther 82:8–24
17. Deyo R, Centor R (1986) Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. J Chronic Dis 39:897–906
18. Deyo RA, Battie M, Beurskens AJ, 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
19. Dionne CE, Koepsell TD, Von Korff M, Deyo RA, Barlow WE, Checkoway H (1997) Predicting long-term functional limitations among back pain patients in primary care settings. J Clin Epidemiol 50:31–43
20. Dionne CE, Von Korff M, Koepsell TD, Deyo RA, Barlow WE, Checkoway H (1999) A comparison of pain, functional limitations, and work status indices as outcome measures in back pain research. Spine 24:2339–2345
21. Dropsy R, Marty M (1994) Quality-of-life indexes for assessment of low-back-pain. Rev Rhum 61:S44–48
22. Euroquol Group (1990) Euroquol – a new facility for the measurement of health-related quality of life. Health Pol 16:199–208
23. Fairbank J (1995) Use of Oswestry Disability Index (ODI). Spine 20: 1535–1537
24. Fairbank JC, Couper J, Davies JB, O Brien JP (1980) The Oswestry low back pain disability questionnaire. Physiotherapy 66:271–3
25. Fisher K, Johnson M (1997) Validation of the Oswestry low back pain questionnaire, its sensitivity as a measure of change following treatment and its relationship with other aspects of the chronic pain experience. Physiother Theory Pract 22:61–80
26. 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
27. Greenough CG, Fraser RD (1992) Assessment of outcome in patients with low-back pain. Spine 17:36–41
28. Grimby G, Smedby B (2001) ICF approved as the successor of ICIDH. J Rehabil Med 33:193–194
29. Gronblad M, Hupli M, Wennerstrand P, Jarvinen E, Lukinmaa A, Kouri JP, Karaharju EO (1993) Intercorrelation and test-retest reliability of the Pain Disability Index (PDI) and the Oswestry Disability Questionnaire (ODQ) and their correlation with pain intensity in low back pain patients. Clin J Pain 9:189–195
30. Gronblad M, Jarvinen E, Hurri H, Hupli M, Karaharju EO (1994) Relationship of the Pain Disability Index (PDI) and the Oswestry Disability Questionnaire (ODQ) with three dynamic physical tests in a group of patients with chronic low-back and leg pain. Clin J Pain 10: 197–203
31. Guillemin F, Constant F, Collin JF, Boulange M (1994) Short and long-term effect of spa therapy in chronic low back pain. Br J Rheumatol 33: 148–151
32. Holt AE, Shaw NJ, Shetty A, Greenough CG (2002) The reliability of the Low Back Outcome Score for back pain. Spine 27:206–210
33. Hsieh CY, Phillips RB, Adams AH, Pope MH (1992) Functional outcomes of low back pain: comparison of four treatment groups in a randomized controlled trial. J Manipulative Physiol Ther 15:4–9
34. Huskisson EC (1974) Measurement of pain. Lancet 2:1127–1131
35. Johansson E, Lindberg P (1998) Subacute and chronic low back pain. Reliability and validity of a Swedish version of the Roland and Morris Disability Questionnaire. Scand J Rehabil Med 30:139–143
36. Kahtri H, Greenough CG (2002) Minimum clinically important difference in the low back outcome score. Society of Back Pain Research, Leeds, UK
37. Kopec JA, Esdaile JM (1995) Functional disability scales for back pain. Spine 20:1943–1949
38. Kopec JA, Esdaile JM, Abrahamowicz M, Abenhaim L, Wood Dauphinee S, Lamping DL, Williams JI (1995) The Quebec Back Pain Disability Scale. Measurement properties. Spine 20: 341–352
39. Kosinski M, Zao S, Dedhiya S, Osterhaus J, Ware JJ (2000) Determine minimally important change in generic and disease-specific health-related quality of life questionnaires in clinical trials of rheumatoid arthritis. Arthritis Rheum 43(7):1478–1487
40. Kovacs FM, Llobera J, Gil Del Real MT, Abraira V, Gestoso M, Fernandez C, Primaria Group KA (2002) Validation of the Spanish version of the Roland-Morris questionnaire. Spine 27: 538–542
41. Kucukdeveci AA, Tennant A, Elhan AH, Niyazoglu H (2001) Validation of the Turkish version of the Roland–Morris Disability Questionnaire for use in low back pain. Spine 26:2738–2743
42. Leclaire R, Blier F, Fortin L, Proulx R (1997) 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
43. Leung AS, Lam TH, Hedley AJ, Twomey LT (1999) Use of a subjective health measure on Chinese low back pain patients in Hong Kong. Spine 24: 961–966
44. Manniche C, Asmussen K, Lauritsen B, Vinterberg H, Kreiner S, Jordan A (1994) Low back pain rating scale: validation of a tool for assessment of low back pain. Pain 57:317–326
45. McDowell I, Newell C (1996) The theoretical and technical foundations of health measurement. Measuring health. A guide to rating scales and questionnaires. Oxford University Press, pp 10–42
46. Million R, Hall W, Nilsen KH, Baker RD, Jayson MI (1981) Assessment of the progress of the back-pain patient 1981 Volvo Award in Clinical Science. Spine 7:204–212
47. Nusbaum L, Natour J, Ferraz MB, Goldenberg J (2001) Translation, adaptation and validation of the Roland–Morris questionnaire – Brazil Roland–Morris. Braz J Med Biol Res = Revista brasileira de pesquisas medicas e biologicas / Sociedade Brasileira de Biofisica ... [et al.] 34:203–210
48. Padua R, Padua L, Ceccarelli E, Romanini E, Bondi R, Zanoli G, Campi A (2001) Cross-cultural adaptation of the lumbar North American Spine Society questionnaire for Italian-speaking patients with lumbar spinal disease. Spine 26:E344–347
49. Patrick DL, Deyo RA, Atlas SJ, Singer DE, Chapin A, Keller RB (1995) Assessing health-related quality of life in patients with sciatica. Spine 20:1899–1908; discussion 1909
50. Porst R (2000) Question wording – Zur Formulierung von Fragebogen-Fragen. ZUMA, Mannheim 2:1–11

51. Pose B, Sangha O, Peters A, Wildner M (1999) [Validation of the North American Spine Society Instrument for assessment of health status in patients with chronic backache]. Z Orthop Ihre Grenzgeb 137:437–441
52. Roese I, Kohlmann T, Raspe H (1996) [Measuring functional capacity in backache patients in rehabilitation: a comparison of standardized questionnaires]. Rehabilitation 35:103–108
53. Roland M, Fairbank J (2000) The Roland–Morris Disability Questionnaire and the Oswestry Disability Index. Spine 25:3115–3124
54. Roland M, Morris R (1983) 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
55. Roland M, Morris R (1983) 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
56. Ruta DA, Garratt AM, Wardlaw D, Russell IT (1994) Developing a valid and reliable measure of health outcome for patients with low back pain. Spine 19:1887–1896
57. Schoppink LE, van Tulder MW, Koes BW, Beurskens SA, de Bie RA (1996) Reliability and validity of the Dutch adaptation of the Quebec Back Pain Disability Scale. Phys Ther 76:268–275
58. Steiner D, Normann G (1995) Health measurement scales. A practical guide to their development and use. Oxford Medical, 2nd edn.
59. Stratford PW, Binkley J, Solomon P, Finch E, Gill C, Moreland J (1996) Defining the minimum level of detectable change for the Roland–Morris questionnaire. Phys Ther 76:359–365; discussion 366–368
60. 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
61. Stratford PW, Binkley JM (1997) Measurement properties of the RM-18. A modified version of the Roland–Morris Disability Scale. Spine 22: 2416–2421
62. Sudman S, Bradburn N, Schwarz N (1996) Thinking about answers. The application of cognitive processes to survey methodology. Jossey-Bass, San Francisco
63. Taylor SJ, Taylor AE, Foy MA, Fogg AJ (1999) Responsiveness of common outcome measures for patients with low back pain. Spine 24:1805–1812
64. Tubergen A van, Landewe R, Heuft-Dorenbosch L, Spoorenberg A, van der Heijde D, van der Tempel H, van der Linden S (2003) Assessment of disability with the World Health Organisation Disability Assessment Schedule II in patients with ankylosing spondylitis. Ann Rheum Dis 62:140–145
65. Underwood MR, Barnett AG, Vickers MR (1999) Evaluation of two time-specific back pain outcome measures. Spine 24:1104–1112
66. Waddell G, Main CJ (1984) Assessment of severity in low-back disorders. Spine 9:204–208
67. Ware JEJ, Sherbourne CD (1992) The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 30:473–483
68. Wiesinger GF, Nuhr M, Quittan M, Ebenbichler G, Wolfl G, Fialka Moser V (1999) Cross-cultural adaptation of the Roland–Morris questionnaire for German-speaking patients with low back pain. Spine 24:1099–1103
69. Yang Y, Eaton S, Maxwell M (1983) The relationship between the St. Thomas and Oswestry Disability Scores and the severity of low back pain. J Manipulative Physiol Ther 16:14–28
70. Zoega B, Karrholm J, Lind B (2000) Outcome scores in degenerative cervical disc surgery. Eur Spine J 9:137–143