

Interpreting Change Scores for Pain and Functional Status in Low Back Pain

Towards International Consensus Regarding Minimal Important Change

Raymond W. J. G. Ostelo, PhD,*† Rick A. Deyo, PhD,‡ P. Stratford, PhD,§
Gordon Waddell, MSc, MD,¶|| Peter Croft, PhD,|| Michael Von Korff, PhD,**
Lex M. Bouter, PhD,*†† and Henrica C. de Vet, PhD*

Study Design. Literature review, expert panel, and a workshop during the “VIII International Forum on Primary Care Research on Low Back Pain” (Amsterdam, June 2006).

Objective. To develop practical guidance regarding the minimal important change (MIC) on frequently used measures of pain and functional status for low back pain.

Summary of Background Data. Empirical studies have tried to determine meaningful changes for back pain, using different methodologies. This has led to confusion about what change is clinically important for commonly used back pain outcome measures.

Methods. This study covered the Visual Analogue Scale (0–100) and the Numerical Rating Scale (0–10) for pain and for function, the Roland Disability Questionnaire (0–24), the Oswestry Disability Index (0–100), and the Quebec Back Pain Disability Questionnaire (0–100). The literature was reviewed for empirical evidence. Additionally, experts and participants of the VIII International Forum on Primary Care Research on Low Back Pain were consulted to develop international consensus on clinical interpretation.

Results. There was wide variation in study design and the methods used to estimate MICs, and in values found for MIC, where MIC is the improvement in clinical status of an individual patient. However, after discussion among experts and workshop participants a reasonable consensus was achieved. Proposed MIC values are: 15 for the Visual Analogue Scale, 2 for the Numerical Rating Scale, 5 for the Roland Disability Questionnaire, 10 for the Oswestry Disability Index, and 20 for the QBDO. When the baseline score is taken into account, a 30% improvement was considered a useful threshold for identifying clinically meaningful improvement on each of these measures.

Conclusion. For a range of commonly used back pain outcome measures, a 30% change from baseline may be considered clinically meaningful improvement when comparing before and after measures for individual patients. It is hoped that these proposals facilitate the use of these measures in clinical practice and the comparability of future studies. The proposed MIC values are not the final answer but offer a common starting point for future research.

Key words: outcome measures, low back pain, minimal important change. **Spine 2008;33:90–94**

Patient-reported outcomes are well established and there are now many self-reported measures used for low back pain. To facilitate comparison of results between studies and to enable the pooling of data in systematic reviews, an international group of investigators recommended a standardized “core” set of measures in 1998¹ which was revised in 2000.² They suggested 5 domains: pain, back specific function, work disability, generic health status, and patient satisfaction.² This article focuses on what are arguably the 2 most fundamental clinical outcomes: pain and back specific function.

The measurement properties of commonly used measures (*i.e.*, questionnaires) in these 2 domains are well established^{3,4} but the challenge remains: what constitutes an important change? Statistical significance does not necessarily mean the change is clinically important.⁵ For some clinical outcomes such as blood pressure, empirical research, and clinical experience may produce a general feeling whether a change is important or not. But the importance of changes on many questionnaires is less intuitively apparent.⁶

Several empirical studies have tried to determine important changes on these questionnaires, using different methodologies. Some use a distribution-based, whereas others use an anchor-based approach. Distribution-based methods express the observed change in a standardized metric. Examples are the effect size and the standardized response mean, where the numerators of both parameters represent the mean change and the denominators are the standard deviation at baseline and the standard deviation of change, respectively. Another is the standard error of measurement, which relates the reliability of the measurement instrument to the standard deviation of the population.⁷ Effect size and standardized

From the *EMGO Institute, VU University Medical Centre; †Institute for Health Sciences, VU University, Amsterdam, The Netherlands; ‡Department of Medicine, University of Washington, Seattle, WA; §Department of Clinical Epidemiology and Biostatistics, School of Rehabilitation Science, McMaster University, Hamilton, Ontario, Canada; ¶UnumProvident Centre for Psychosocial and Disability Research, Cardiff University, Cardiff, UK; ||Primary Care Musculoskeletal Research Centre, Keele University, UK; **Centre for Health Studies, Group Health Cooperative, Seattle, WA; and ††Executive Board of VU University, Amsterdam, The Netherlands.

Acknowledgment date: April 4, 2007. Revision date: April 26, 2007. Acceptance date: May 10, 2007.

The manuscript submitted does not contain information about medical device(s)/drug(s).

No funds were received in support of this work. No benefits in any form have been or will be received from a commercial party related directly or indirectly to the subject of this manuscript.

Address correspondence and reprint requests to Raymond W.J.G. Ostelo, PhD, EMGO Institute, VU University Medical Center, Van der Boechorststraat 7, 1081 BT Amsterdam, The Netherlands; E-mail: r.ostelo@vumc.nl, www.emgo.nl

response mean are relative representations of change (*i.e.*, without units), whereas standard error of measurement provides a number in the same units as the original measurement. The major disadvantage of all distribution-based methods is that they are purely statistical measures of the magnitude of the change: they do not say anything about its clinical importance.

Anchor-based methods use some external criterion to operationalize the importance of the observed change: *i.e.*, the concept of “minimal importance” is explicitly defined and incorporated in the method. Many different anchors have been used: *e.g.*, comparing change with achieving (or not) predefined treatment goals or with retrospective global rating of improvement. Another problem is the definition of “importance”: *e.g.*, some authors classify “minimal improvement” (or “slightly better”) on a global rating as “clinically important” whereas others classify it as or “no important change” or “stable.” Because anchor-based approaches do not take account of measurement precision, they do not necessarily imply statistical significance and consequently that must be calculated separately.

These methodologic problems have led to confusion about how to interpret change scores and about deciding what change is clinically important, even though this is fundamental to using these measures. The aim of this article is to develop practical guidance: What is the minimal change that can be considered important on frequently used questionnaires of pain and functional status in low back pain? This is partly a matter of empirical evidence, and partly of developing some international consensus on its clinical interpretation.

■ Materials and Methods

The study covered the core set recommendations and most commonly used measures: for pain, the Numerical Rating Scale (scoring range 0–10)⁴ and the Visual Analogue Scale (scoring range 0–100)⁴; and for function, the Roland Morris Disability Questionnaire (scoring range 0–24)⁸ and the Oswestry Disability Index (ODI: scoring range 0–100).⁹ We also included the Quebec Back Pain Disability Questionnaire (scoring range 0–100),¹⁰ as it is frequently used in randomized controlled trials.

Empirical Evidence. The literature was reviewed for studies estimating the minimal important change (MIC) for the above questionnaires. MEDLINE was searched using a combination of Medical Subject Heading (MeSH) terms “back pain” and “low back pain”; the specific names of the questionnaires; and any of the following terms—“responsiveness,” “minimal(ly) clinically important change,” “minimal(ly) clinically important difference,” “minimum clinically important difference,” “minimum detectable change,” “smallest detectable change” and “questionnaires.” Studies were included that reported on the importance of the change scores. The findings were extracted and tabulated by 2 of the authors (RO, HdV). This overview table included the study population, study design, method used to estimate MIC, and cutoff value(s) for the MIC.

The Expert Panel. Experts involved in recommending the original core set, or in the update on the relevant measures, were invited. Additionally, authors of recently published relevant papers on the measures at issue were invited. In total, 6 of 8 invitees agreed to participate. Members of the expert panel were sent the overview table of the empirical evidence and asked the following questions: (1) Based on this overview, what is the most appropriate value for the MIC for the included questionnaires? (2) Is this also your personal opinion? (3) Should the MIC also be described in percentages from baseline score?

Workshop. A workshop was organized during the “VIII International Forum on Primary Care Research on Low Back Pain” in Amsterdam (June 2006) to get the input from a wider range of low back pain researchers. Thirty-six participants were presented with the overview table, but were not given the expert panel answers to avoid influencing them. Participants were then divided into 5 groups, each of which discussed the above 3 questions and the conclusions were reported back to a plenary session.

Synthesis of Recommendations. Answers from the expert panel and the workshop were synthesized (by RO and HdV) into provisional proposals for MIC values. Members of the expert panel reviewed the provisional proposals and issues raised in the workshop discussion, and their comments were incorporated into the final recommendations.

■ Results

Empirical Evidence

Three studies on the Visual Analogue Scale, 5 on the Numerical Rating Scale, 17 on the Roland Disability Questionnaire, 5 on the ODI, and 4 on the QBDQ were identified and included. As expected, there was wide variation in study design and the methods used to estimate MICs. For example, the included studies used different time intervals for the test-retest (ranging from a 1-day interval to a 1-year interval), different external criteria to define important were used and many different statistical techniques were used to calculate MIC. Although these methodologic issues are important and closely connected with the MIC as estimated in each study, little (or no) theoretical or empirical justification was provided for the study design, anchor or method used for estimating MICs in the identified studies. Within the framework of this consensus procedure we decided that the main focus should be on the actual values for MIC, not the methodology. MICs were generally presented either as an absolute value for change (intended for use anywhere in the range of the scale) or as values dependent on initial scores (*e.g.*, as a percentage). Table 1 presents the range of MIC values for each questionnaire based on the empirical evidence.

Expert Panel and Workshop

Both the expert panel and the workshop participants experienced difficulty in answering the first question about the most appropriate MIC cutoffs. This was mainly because of the heterogeneity of the studies, the disparate results, and the lack of any clear rationale for integrating them. Nevertheless, discussion led to reason-

Table 1. Ranges for MIC Values Based on the Empirical Evidence

Questionnaire	Scoring Range	Range of MIC Values (Absolute)	Range of MIC Values (% Improvement From Baseline)
VAS ^{11–13}	0–100	2.0–29.0 points	No empirical evidence
NRS ^{12,14–16}	0–10	1.0–4.5 points	30
RDQ ^{3,10–12,17–28}	0–24	2.0–8.6 points*	30
ODI ^{3,11–13,21}	0–100	4.0–15.0 points	No empirical evidence
QBPO ^{10,14,21}	0–100	8.5–32.9 points	No empirical evidence

Absolute values presented are intended for use anywhere in the range of the scale.

*11–13 points for high baseline scores, when these were taken into account. VAS indicates Visual Analogue Scale; NRS, Numerical Rating Scale; RDQ, Roland Morris Disability Questionnaire; ODI, Oswestry Disability Index; QBPO, Quebec Back Pain Disability Questionnaire.

able consensus. On the second question, personal opinions did not deviate significantly from the answers to question 1. Apparently, personal opinions were already included in reaching consensus on the empirical evidence. On the third question, some experts and workshop participants felt that 1 simple (absolute) value for MIC for each questionnaire is easier to produce from the available evidence. Furthermore, such a uniform value is more likely to be used in clinical practice. Others felt that was an oversimplification as there is evidence that MIC is baseline dependent, so initial values should be taken into account *e.g.*, as percentage improvement from baseline. It was therefore decided to work toward consensus on both issues.

The discussions in the expert group and workshop raised also several other issues. Debate remains about the meaning and definition of a “clinically important change.” For example, some participants regarded “slightly improved” as clinically important whereas others considered this within the range of natural fluctuation. The latter reasoned that an “important” improvement should be greater than these (unimportant) natural fluctuations. Furthermore, patients may easily say that they are slightly improved just to please their physician or therapist. Better methodology will not resolve this; rather, these are clinical judgments that then determine the methodology used.

There was also debate about whether different MIC values should be used for acute, subacute, and chronic low back pain. After discussion it was agreed on that there is insufficient empirical evidence to set different MICs for these different types of low back pain, though that may already be reflected in baseline scores and would in any event be difficult to operationalize. Table 2 presents the range of MIC values for each questionnaire after the first expert panel round and the workshop.

Data Synthesis

The final proposals for the MIC values on each measure are presented in Table 3. During discussion it was suggested that the definition of MIC should be simple and generalizable to different outcome measures. When base-

Table 2. Ranges for MIC Values After First Expert Panel and Workshop

Questionnaire	Scoring Range	Range of MIC Values (Absolute)	Range of MIC Values (% Improvement From Baseline)
VAS	0–100	15.0–20.0 points	20–30
NRS	0–10	1.0–2.0 points	20–30
RDQ	0–24	3.0–6.0 points	20–30
ODI	0–100	10.0–12.0 points	20–30
QBPO	0–100	20.0 points	20–30

Absolute values presented are intended for use anywhere in the range of the scale.

VAS indicates Visual Analogue Scale; NRS, Numerical Rating Scale; RDQ, Roland Morris Disability Questionnaire; ODI, Oswestry Disability Index; QBPO, Quebec Back Pain Disability Questionnaire.

line is taken into account, a 30% improvement was considered a generally useful guide.

Discussion

This is a first attempt to develop recommendations on MICs for commonly used measures of pain and function in low back pain, and the findings must be viewed in light of the methodologic limitations of this study. Firstly, the search strategy was not optimal, though as the experts and participants included many leaders in this field it is unlikely that important articles were missed. Moreover, the literature review was not the primary purpose of the exercise but rather the starting point for discussion and consensus. The included studies were so heterogeneous that any additional studies would probably have compounded rather than resolved the problem. Second, the empirical evidence is limited and heterogeneous and there are no agreed scientific grounds or empirical evidence to determine the optimum method of estimating the MIC. Therefore, the results are variable and difficult to integrate. Finally, consensus was constrained by the experts and participants who contributed, with a particular focus on primary care. Nevertheless, MIC values depend not only on empirical evidence but also on clinical interpretation and judgment, so there is a good argument for combining empirical evidence and consensus procedures to come to a reasonable and parsimonious choice of MIC values.

Table 3. Proposed Cutoff Values for MIC

Questionnaire	Scoring Range	MIC (Absolute Cutoff)	MIC (% Improvement From Baseline)
VAS	0–100	15	30
NRS	0–10	2	30
RDQ	0–24	5	30
ODI	0–100	10	30
QBPO	0–100	20	30

Absolute values presented are intended for use anywhere in the range of the scale.

VAS indicates Visual Analogue Scale; NRS, Numerical Rating Scale; RDQ, Roland Morris Disability Questionnaire; ODI, Oswestry Disability Index; QBPO, Quebec Back Pain Disability Questionnaire.

There was debate about whether MICs should be expressed as a single value or as a range that includes all reasonable values. Ranges, however, require the user to know when to use the larger or smaller values. Many may be tempted to use the smallest MIC in order to demonstrate more improvement, but that may not be most appropriate to the patient group or intervention. There was insufficient empirical evidence to set different MICs for acute or chronic low back pain, though that may already be reflected in baseline scores and would in any event be difficult to operationalize. Nevertheless, different MICs may be more appropriate for different patients or contexts, *e.g.*, children or surgical patients. Again, a smaller MIC may be appropriate to a simple, cheap, and safe intervention, whereas a larger MIC may be more appropriate to an expensive, risky procedure. Indeed, an ODI MIC of 15 points has been suggested for surgical interventions,³ compared with the 10 points proposed here (Table 3). Types of patients and treatments were not specifically taken into account in these proposals. Thus, the proposed values should be taken as generic lower limits for the MICs which can (and should) be modified when necessary.

Many participants stressed that the proposed MIC values were for individual rather than group changes. Randomized controlled trials typically analyze group differences between treatment and control interventions, and investigators and clinicians assume that whether the difference in the means is less than the MIC, the treatment effect is unimportant. However, it is entirely possible that individual patients in the trial do show clinically important improvement.²⁹ Therefore, Guyatt *et al*²⁹ have proposed a method for estimating the proportion of patients who benefit from a treatment when the outcome measure is a continuous variable. Recent Food and Drug Administration Guidance states that there may be situations where it is more reasonable to look at individual rather than group responses, provided the definition of responders is based on prespecified criteria backed by empirical evidence.³⁰ The MIC values proposed in this study can be used for labeling individuals as responders to treatment.

It is hoped that these proposals will facilitate the use of these measures of pain and functional limitation in clinical practice. The proposed MIC values are not the final answer but guidance, which may offer a common starting point for future research. They should also improve the comparability of future studies, pooling, and the clinical interpretation of results. Future research may yield new evidence that necessitates modification of this guidance.

■ Key Points

- Empirical studies have tried to determine meaningful changes for back pain, using different methodologies. This has led to confusion about what change is clinically important for commonly used back pain outcome measures.

- This article provides practical guidance regarding the MIC for a range of commonly used back pain outcome measures for pain and functional status.
- It is hoped that these proposals facilitate the use of these measures in clinical practice and the comparability of future studies. The proposed MIC values are not the final answer but offer a common starting point for future research, which may yield new evidence that necessitates modification of this guidance.

Acknowledgments

The authors gratefully acknowledge the contribution of all who participated in this workshop during Low Back Pain Forum VIII, June 2006, in Amsterdam, the Netherlands.

References

1. Deyo RA, Battie M, Beurskens AJ, et al. Outcome measures for low back pain research. A proposal for standardized use. *Spine* 1998;23:2003–13.
2. Bombardier C. Outcome assessments in the evaluation of treatment of spinal disorders: summary and general recommendations. *Spine* 2000;25:3100–3.
3. Roland M, Fairbank J. The Roland-Morris Disability Questionnaire and the Oswestry Disability Questionnaire. *Spine* 2000;25:3115–24.
4. von Korff M, Jensen MP, Karoly P. Assessing global pain severity by self-report in clinical and health services research. *Spine* 2000;25:3140–51.
5. Wright JG. The minimal important difference: who's to say what is important? *J Clin Epidemiol* 1996;49:1221–2.
6. Juniper EF, Guyatt GH, Willan A, et al. Determining a minimal important change in a disease-specific Quality of Life Questionnaire. *J Clin Epidemiol* 1994;47:81–7.
7. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. *J Clin Epidemiol* 2003;56:395–407.
8. Roland M, Morris R. A study of the natural history of back pain. I. Development of a reliable and sensitive measure of disability in low-back pain. *Spine* 1983;8:141–4.
9. Fairbank JC, Couper J, Davies JB, et al. The Oswestry low back pain disability questionnaire. *Physiotherapy* 1980;66:271–3.
10. Kopec JA, Esdaile JM, Abrahamowicz M, et al. The Quebec Back Pain Disability Scale. Measurement properties. *Spine* 1995;20:341–52.
11. Beurskens AJ, de Vet HC, KÖke AJ, et al. Measuring the functional status of patients with low back pain. Assessment of the quality of four disease-specific questionnaires. *Spine* 1995;20:1017–28.
12. Grotle M, Brox JI, Vollestad NK. Concurrent comparison of responsiveness in pain and functional status measurements used for patients with low back pain. *Spine* 2004;29:E492–E501.
13. Hagg O, Fritzell P, Nordwall A. The clinical importance of changes in outcome scores after treatment for chronic low back pain. *Eur Spine J* 2003;12:12–20.
14. van der Roer N, Ostelo RW, Bekkering GE, et al. Minimal clinically important change for pain intensity, functional status, and general health status in patients with nonspecific low back pain. *Spine* 2006;31:578–82.
15. Farrar JT, Young JP Jr, LaMoreaux L, et al. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. *Pain* 2001;94:149–58.
16. Childs JD, Piva SR, Fritz JM. Responsiveness of the numeric pain rating scale in patients with low back pain. *Spine* 2005;30:1331–4.
17. Stratford PW, Binkley J, Solomon P, et al. Defining the minimum level of detectable change for the Roland-Morris Questionnaire. *Phys Ther* 1996;76:359–65.
18. Stratford PW, Binkley JM. Measurement properties of the RM-18. A modified version of the Roland-Morris Disability Scale. *Spine* 1997;22:2416–21.
19. Riddle DL, Stratford PW, Binkley JM. Sensitivity to change of the Roland-Morris Back Pain Questionnaire: part 2. *Phys Ther* 1998;78:1197–207.
20. Stratford PW, Binkley JM, Riddle DL, et al. Sensitivity to change of the Roland-Morris Back Pain Questionnaire: part 1. *Phys Ther* 1998;78:1186–96.
21. Davidson M, Keating JL. A comparison of five low back disability questionnaires: reliability and responsiveness. *Phys Ther* 2002;82:8–24.

22. Wiesinger GF, Nuhr M, Quittan M, et al. Cross-cultural adaptation of the Roland-Morris Questionnaire for German-speaking patients with low back pain. *Spine* 1999;24:1099–103.
23. Johansson E, Lindberg P. Subacute and chronic low back pain. Reliability and validity of a Swedish version of the Roland and Morris Disability Questionnaire. *Scand J Rehabil Med* 1998;30:139–43.
24. Nusbaum L, Natour J, Ferraz MB, et al. Translation, adaptation and validation of the Roland-Morris Questionnaire–Brazil Roland-Morris. *Braz J Med Biol Res* 2001;34:203–10.
25. Stratford PW, Binkley JM. A comparison study of the back pain functional scale and Roland Morris Questionnaire. North American Orthopaedic Rehabilitation Research Network. *J Rheumatol* 2000;27:1928–36.
26. Stratford PW, Binkley JM, Riddle DL. Development and initial validation of the back pain functional scale. *Spine* 2000;25:2095–102.
27. Garratt AM, Klaber Moffett J, Farrin AJ. Responsiveness of generic and specific measures of health outcome in low back pain. *Spine* 2001;26:71–7.
28. Jordan K, Dunn KM, Lewis M, et al. A minimal clinically important difference was derived for the Roland-Morris Disability Questionnaire for low back pain. *J Clin Epidemiol* 2006;59:45–52.
29. Guyatt GH, Juniper EF, Walter SD, et al. Interpreting treatment effects in randomised trials. *BMJ* 1998;316:690–3.
30. FDA Guidance on PRO. *Guidance for Industry Patient-Reported Outcome Measures. Use in Medical Product Development to Support Labeling Claims*. Rockville, MD: Food and Drug Administration; 2006.