OPEN

SPINE An International Journal for the study of the spine, Publish Ahead of Print

DOI: 10.1097/BRS.0000000000003107

The NIH minimal dataset for chronic low back pain: responsiveness and minimal clinically important change

Alisa L. Dutmer[1], MSc, Michiel F. Reneman[1], PhD, Henrica R. Schiphorst Preuper[1,2], MD, PhD, André P. Wolff[2,3], MD, PhD, Bert L. Speijer[2], MPA, Remko Soer[2,4], PhD;

1. University of Groningen, University Medical Center Groningen, Department of Rehabilitation, Groningen, The Netherlands
2. University of Groningen, University Medical Center Groningen, Groningen Spine Center, Groningen, The Netherlands.
3. University of Groningen, University Medical Center Groningen, Department of Anaesthesiology, Pain Center, Groningen, The Netherlands
4. Saxion University of Applied Sciences, Expertise Center of Health and Movement, Enschede, The Netherlands.

Contact information :

Alisa L. Dutmer, MSc
Center for Rehabilitation,
University Medical Center Groningen,
P.O. Box 30002
9750 RA Haren The Netherlands
Tel: 0031503617200
Email: a.l.dutmer@umcg.nl

The manuscript submitted does not contain information about medical device(s)/drug(s).
No funds were received in support of this work.
No relevant financial activities outside the submitted work.

Abstract

Study Design: prospective cohort study.

Objective: To analyze responsiveness and minimal clinically important change (MCIC) of the US National Institutes of Health (NIH) minimal dataset for chronic low back pain (CLBP).

Summary of Background Data: The NIH minimal dataset is a 40-item questionnaire developed to increase use of standardized definitions and measures for CLBP. Longitudinal validity of the total minimal dataset and the subscale Impact Stratification are unknown.

Methods: Total outcome scores on the NIH minimal dataset, Dutch Language Version, were calculated ranging from 0-100 points with higher scores representing worse functioning. Responsiveness and MCIC were determined with an anchor based method, calculating the area under the receiver operating characteristics (ROC) curve (AUC) and by determining the optimal cut-off point. Smallest detectable change (SDC) was calculated as a parameter of measurement error.

Results: In total 223 patients with CLBP were included. Mean total score on the NIH minimal dataset was 44±14 points at baseline. The total outcome score was responsive to change with an AUC of 0.84. MCIC was 14 points with a sensitivity of 72% and specificity 82%, and SDC was 23 points. Mean total score on Impact Stratification (scale 8-50) was 34.4±7.4 points at baseline, with an AUC of 0.91, an MCIC of 7.5 with a sensitivity 96% of and specificity of 78%, and an SDC of 14 points.

Conclusion: The longitudinal validity of the NIH minimal dataset is adequate. An improvement of 14 points in total outcome score and 7.5 points in Impact Stratification can be interpreted as clinically important in individual patients. However, MCIC depends on

baseline values and the method that is chosen to determine the optimal cut-off point. Furthermore, measurement error is larger than the MCIC. This means that individual change scores should be interpreted with caution.

Key Words: Questionnaire, Self-report, Longitudinal validity, Measurement error, Smallest detectable change, Functioning, Disability, Quality of life

Level of Evidence: 4

Key points

1. The longitudinal validity of the NIH minimal dataset is adequate
2. An improvement of 14 points on the total outcome score and 7.5 points on Impact Stratification can be interpreted as clinically important in patients with CLBP.
3. The measurement error is larger than the minimal clinically important change. This means that individual change scores should be interpreted with caution.
4. Generally, for more severe patients, higher change scores should be obtained to be considered clinically important.

Introduction

In 2014, The US National Institute of Health (NIH) introduced a minimal dataset for chronic low back pain (CLBP) to increase use of standardized definitions and measures and to facilitate comparison in clinical and epidemiological studies[1-3]. This self-report questionnaire has been translated and adapted to Canadian French[4], Farsi[5], and Dutch[6]. The NIH minimal dataset includes items related to medical history and self-report measures of physical function, psychosocial functioning, sleep disturbance, pain intensity and pain interference[3]. A Dutch validation study of the NIH minimal dataset revealed sufficient to good measurement properties and a good fit in a 7-factor model[6]. Longitudinal validity of the NIH minimal dataset, however, has not been tested in any language version.

To compare longitudinal measurements and to allow better comparison across studies, an outcome score needs to be constructed. The NIH task force proposed an outcome score called the Impact Stratification. This score consists of three domains only: pain intensity, pain interference, and physical function. To compare full biopsychosocial characteristics and effects after interventions for patients with CLBP, an outcome score should also include the remaining domains of the questionnaire such as depression and sleep disturbance.

Detecting change in health status over time (responsiveness), and being able to interpret change scores are important aspects of patient reported outcome measures[7,8]. Change scores can be interpreted with the minimal clinically important change (MCIC) and the measurement error, expressed as smallest detectable change (SDC). Both MCIC and SDC are expressed on the actual scale of measurement and are therefore advantageous for clinical interpretation. When the MCIC is larger than the SDC, an outcome measure is able to distinguish clinically important change from measurement error[7,9]. The MCIC can also be used in responder analyses, where a proportion of patients is identified that improved by more than the MCIC[10].

Many commonly used patient reported outcome measures in LBP have previously been studied on responsiveness. However, these outcome measures are predominantly unidimensional, measuring a single construct (e.g. pain or back specific function)[11,12]. Whereas the NIH minimal dataset is a multidimensional instrument that combines multiple constructs that are relevant in LBP research, such as pain interference, physical and psychosocial functioning, sleep, and depression. Therefore, studying responsiveness and MCIC of the NIH minimal dataset is deemed important. The objectives of this study were to: construct outcome of the NIH minimal dataset for CLBP, analyze responsiveness, and interpret change scores by determining MCIC and SDC. Secondary analyses were performed to explore whether clinically important change depends on baseline-score.

Materials and methods

Procedures

Data were collected from July 2015 to September 2018 in the Groningen Spine Center, a university-based multispecialty tertiary care center in the north of the Netherlands. Baseline (T0) and 12-months follow-up data (T1) were extracted from a longitudinal cohort. Patients digitally filled out a set of questionnaires, including the NIH minimal dataset Dutch Language Version, the Pain Disability Index (PDI), the EuroQol-5D (EQ5D), a single item on work status, and a Global Perceived Effect (GPE) scale. The Medical Ethical Committee of the University Medical Center Groningen, the Netherlands provided a waiver for this study with respect to medical ethical permission, because the study was performed within care as usual. All patients signed informed consent. The handling of the data was done in accordance with the guidelines for Good Research Practice[13].

Patients

All patients 18-65 years old who reported pain in their lower back and/or leg for more than 12 weeks were included. Patients with no Internet access, insufficient Dutch reading skills,

and who did not respond to the follow-up questionnaire were excluded. The items in the NIH minimal dataset were specifically chosen by the NIH research task force for their importance to a wide range of patients with chronic LBP with or without specific pathoanatomic diagnoses[1]. Therefore, patients with specific or multifactorial (often referred to as non-specific) LBP, were both included. Interventions were chosen based on indication and patient preference. Possible treatment options were multidisciplinary rehabilitation for patients with multifactorial LBP, surgery or conservative therapy for patients with specific complaints such as herniated disks or stenosis, and anesthesiology for patients with clear sensitization patterns in well-described dermatomes.

Measures

NIH Minimal Dataset for CLBP

The NIH minimal dataset includes 40 items related to demographics, medical history, and self-reported symptoms and functioning[1]. Seventeen of these items are derived from the 29-item Patient Reported Outcomes Measurement Information System (PROMIS) short form[14]. An outcome score, Impact Stratification, was created with 9 of the items (1 item on pain intensity, 4 items on pain interference, and 4 items on physical function). For each item a score of 1 is least severe and 5 most severe, with the exception of the single item on pain intensity, which ranges from 0 (no pain) to 10 (worst possible pain). Total scores range from 8 (least impact) to 50 (most impact). Impact Stratification showed moderate and strong correlation with the Roland-Morris Disability Questionnaire (RMDQ) (R_S = 0.66) and Oswestry Disability Index (ODI) (R_S = 0.81) and demonstrated higher responsiveness compared to the RMDQ at 3 months follow-up[3].

Exploratory factor analyses led to a 7-factor model for the NIH minimal dataset with 29 of the original 40 items remaining (Appendix 1)[6]. Two items on ethnicity and race were removed beforehand for their lack of clinical relevance and 9 items were removed during the

analyses for having insufficient variance or significant factor loadings. The 7 factors that were identified are 1: pain intensity and interference (6 items), 2: pain history (7 items), 3: medical interventions (5 items), 4: depression and catastrophizing (6 items), 5: physical function (3 items), 6: sleep disturbance (4 items), and 7: lifestyle (2 items). Factor 3 and one item of factor 2 ("How long has low-back pain been an ongoing problem for you?") are not used in follow-up measurements because their levels are fixed in a cohort of patients with a chronic condition. Each individual factor showed a fair to good correlation with the PDI and EQ5D. Two-week test-retest reliability per factor was moderate to good (ICC = 0.71; range = 0.52-0.82) and showed substantial agreement per item ($\kappa\omega = 0.65$)[6].

Work status

Patients were asked whether they were currently employed. If yes, a question about the status of their employment (working, partial sick leave, sick leave) followed.

Pain Disability Index

The PDI measures self-reported pain interference in 7 categories of daily life activities[15]. The questionnaire is constructed on an 11-point numeric rating scale in which 0 means "no disability" and 10 "maximum disability". Total scores range from 0 to 70 where a higher score means a greater disability due to pain. The Dutch language version of the PDI was used. Two-week test-retest reliability is good[16].

EuroQol-5D

The EQ5D is a 5-item (representing 5 dimensions) questionnaire that measures quality of life[17]. Each item has 3 levels: no problems, some problems, and extreme problems. The dimensions measured are mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. The Dutch utility index was used to calculate a total score between -0.33 and 1.00[18]. Higher scores represent a better quality of life. Criterion validity of the Dutch language version is moderate and responsiveness moderate to good in patients with LBP[19].

Global Perceived Effect Scale

A Global Perceived Effect Scale (GPE) was used as an external criterion[20]. At twelve months follow-up (T1) patients answered the question: "How much did your treated complaints change compared with pretreatment level?" Possible answers ranged from 1 to 7 on a 7-point Likert scale (1, "extremely worsened"; 2, "much worsened"; 3, "little worsened"; 4, "unchanged"; 5, "little improved"; 6, "much improved"; an 7, "completely improved"). Next, patient scores were divided into two categories: improved (much improved and completely improved) and not improved (all others). Studies have reported strong correlations between GPE scores and changes in pain and disability[21,22]. Test-retest reliability of 11-point GPEs are excellent[23]. Overall, the use of a 7 to 11 points scale is recommended when taking into account patient preference, adequate discriminative ability, and test-retest reliability[24].

Data analyses

Constructing outcome scores

All records containing more than 3 missing items were excluded from the study. If less than 3 items were missing, scores were imputed based on the mean score of the item. The response "don't know" for the item on leg pain was defined as a missing value. The raw 29 item scores of the 7-factor model were recoded to scores between 0 and 1, where 0 represents highest levels of functioning and 1 lowest level of functioning. Factor scores were calculated by taking the mean of the corresponding item scores and multiplying them by 100. This led to factor scores ranging from 0-100 points with lower scores representing higher functioning, giving equal weight to each item. A total outcome score was calculated by taking the average of all factor scores (0-100 points). Floor or ceiling effects were considered present when >15% of patients achieved the lowest or highest possible score[25].

Responsiveness and minimal clinically important change

Responsiveness and minimal clinically important chance (MCIC) were calculated according to the Consensus-based Standards for the selection of health Measurement Instruments criteria (COSMIN)[26,27]. To differentiate between improved and unimproved patients the group with the GPE score "improved" was compared with the group "unimproved". Area Under the Receiver Operator Curve (AUC) were calculated for the total NIH minimal dataset, each individual factor and for Impact Stratification. An AUC higher than 0.70 was considered responsive. The MCIC was measured by determining the optimal cut-off point (OCP) of the AUC. This is the cut-point closest to the top-left corner of the ROC curve, where the sum of squares of 1-sensitivity and 1-specificity is minimized (equation 1)[28].

$$\text{OCP} = \min\{(1 - sens)^2 + (1 - spec)^2\} \quad (1)$$

To study the effect of baseline score on MCIC (i.e. patients with low functional problems may have lower MCICs), secondary analyses were performed where responsiveness and MCICs were calculated for different baseline-score groups. For the total outcome score, three equally sized subgroups were created based on tertile baseline scores. For the Impact Stratification, three subgroups were created: mild impact 8-27 points; moderate impact 28-35; severe impact ≥35 points[3].

Measurement error

To determine the measurement error, standard error or measurement (SEM) and smallest detectable change (SDC) were calculated. The SEM was based on the variability between T0 and T1 plus the variability caused by random error (equation 2) in patients who reported to be "unchanged" on the GPE[9,29]. The variance components for the SEM formula were retrieved with the VARCOMP command in SPSS (Version 23; IBM Corp., Armonk, N.Y., USA).

$$\text{SEM} = \sqrt{\sigma^2_{time} + \sigma^2_{error}} \quad (2)$$

The SDC is the smallest change in score that a patient must show to ensure that the observed change is real and not attributed to measurement error. The SDC can be determined in individual patients and at group level with the following equations, where 1.96 relates to a confidence level of 95% and √2 represents a correction for repeated measurements[7,9]:

$$\mathrm{SDC}_{\mathrm{individual}} = 1.96 \times \sqrt{2} \times \mathrm{SEM} \quad (3)$$

$$\mathrm{SDC}_{\mathrm{group}} = \frac{\mathrm{SDC}_{\mathrm{individual}}}{\sqrt{n}} \quad (4)$$

Results

Patients

Baseline and one year follow-up data was available for 223 patients. The majority of patients was diagnosed with multifactorial LBP (78%), and 22% were diagnosed with specific spinal pathology (e.g. fractures, radiculopathy, malignancy, rheumatoid arthritis). No patients were excluded due to having more than 3 missing items on the NIH minimal dataset. Only for the item on the presence of leg pain were there missing values, due to 19 patients answering "don't know". Demographic and clinical variables are presented in Table 1. Mean age was 49.7 ± 11.9 years and 58% of patients were female. A majority (56%) experienced LBP for over five years.

Responsiveness and minimal clinically important change

Total Sample

Scores on the NIH minimal dataset on T0 and T1, mean changes, 95% confidence intervals and responsiveness and MCIC parameters (AUC, OCP, sensitivity and specificity) are presented in Table 2. Relevant floor effects were found at T0 for factor 7 (lifestyle; 81%) and at T1 for factor 4 (depression and catastrophizing; 24%), factor 5 (physical function; 17%), and factor 7 (lifestyle; 78%). Mean change of the total outcome score and Impact

Stratification were respectively 20.5±13.7 and 16.3±7.9 points for improved (n=50) and 3.6±11.1 and 3.5±6.4 points for unimproved (n=173) patients. The total outcome score and Impact Stratification showed good responsiveness with an AUC of respectively 0.84 (0.78-0.91) and 0.91 (0.86-0.96). The MCIC was 14.3 points for the total outcome score with a sensitivity and specificity of 0.72 and 0.82 and 7.5 points for the Impact Stratification with a sensitivity and specificity of 0.86 and 0.96. Factor 1 (pain intensity and interference) and factor 5 (physical function) also showed good responsiveness with AUCs≥0.70. The lower bound of the AUC confidence interval was <0.70 for factor 2 (95% CI = 0.67-0.83) and factor 4 (95% CI = 0.62-0.79), whereas factor 6 (sleep disturbance) and factor 7 (lifestyle) showed insufficient responsiveness (AUC≤0.70).

Baseline-score groups

Responsiveness and MCIC parameters for the different baseline-score groups are presented in Table 3. Total outcome score groups were equal in size, but for Impact Stratification group sizes differed with the severely impacted group being the largest baseline-score subgroup. Scores between T0 and T1 improved significantly for all subgroups except for both lowest scoring (thus highest functioning) subgroups at baseline, i.e. tertile 1 of the total outcome score and subgroup mild from Impact Stratification. Adversely, both groups proportionally had most improved patients according to the GPE (Tertile 1, 26%; Mild, 32%). MCICs for the total outcome score were 6.9 points for tertile 1, 19.7 points for tertile 2, and 17.1 points for tertile 3. For the Impact Stratification MCICs were 7.5 points for mildly impacted patients, 11.5 points for moderately, and 12.5 points for severely impacted patients.

Measurement Error

The SEM was 8.3 points for the total outcome score with an $SDC_{individual}$ of 22.9 and an SDC_{group} of 1.8 points. The SEM for the Impact Stratification was 5.2 points with an $SDC_{individual}$ of 14.4 and an SDC_{group} of 1.1 points.

Discussion

This is the first study to calculate a total outcome score for the NIH minimal dataset, using a 7-factor model from a previous Dutch validation study[6]. Results show that the total outcome score for the NIH minimal dataset is responsive. A change score of 14 points on the total outcome score (0-100) and 7.5 points on Impact Stratification (8-50) can be considered clinically important. However, individual change scores up to 23 points for the total outcome score and 14 points for Impact Stratification should be interpreted with caution, because of a greater than 5% risk of measurement error.

The total outcome score of the NIH minimal dataset, Impact Stratification, and separate factors related to pain and functioning showed good responsiveness. These findings correspond with other studies in patients with musculoskeletal pain[30] and lumbar spinal surgery[31]. MCICs in our study are similar to proposed MCICs for commonly used pain and disability measures in LBP and also vary between 10-20% of a total score[32]. Furthermore, as a rule of thumb a 30% change from baseline is considered a useful threshold for identifying clinically meaningful improvement[32]. Overall, responsiveness for the NIH total outcome score and pain and functioning domains is established and clinically relevant change scores match recommendations from literature.

The SDC individual for the total outcome score (22.9 points) and Impact Stratification (14.4) both exceeded the MCICs of the corresponding measures. This is also observed in other studies on back pain[16,33-35]. Individual change scores larger than the MCIC but smaller

than the SDC need to be interpreted with caution, because there is a risk of falsely labeling patients as improved while their scores fall within the measurement error. As the SDC_{group} was considerably smaller than the $SDC_{individual}$, both outcome scores are better at detecting changes at a group level. However, we recommend reporting the percentage of improved patients (responders; determined with the MCIC) instead of comparing change scores on a group level.

It is also important to take baseline scores into account when interpreting individual change scores[36]. Higher MCICs apply for higher baseline values (more severely impacted), since there is more potential for improvement[34,37]. MCICs for the total outcome score baseline-score groups were less proportionally distributed with estimates of roughly 7, 20, and 17 points in order from lowest to highest scoring tertile. Change scores for the lowest scoring tertile (< 29 points at baseline) appear more difficult to interpret compared to other baseline-score groups due to partially insufficient responsiveness and lower sensitivity and specificity for the MCIC. A larger MCIC for the second tertile compared to the highest tertile seems counterintuitive, but has been observed before in a study on pain and disability instruments in patients with LBP[38] . The authors hypothesized that the more disabled patients at baseline possibly learned not to have too high expectations to the treatment outcome.

Responsiveness was insufficient for the factors depression and catastrophizing, sleep disturbance and lifestyle. Floor effects were observed for the factor depression and catastrophizing at follow-up (24%) and for lifestyle at baseline (81%) and follow-up (78%), indicating that a decrease in score could possibly be underestimated, or that, for lifestyle in particular, there was little effect of LBP on these domains. It should also be noted that the concept of recovery can be complex and that we do not know what patients take into account as they rate their perceived overall change[24,39]. Patients may perceive a “change in treated complaints” (phrasing of the GPE item), in terms of reduced pain and disability and higher

levels of functioning instead of improvement in domains such as sleep and depression. However, several studies on the patient perspective on successful treatment for chronic pain do indicate the relevance of outcomes such as enjoyment of life, emotional well-being, fatigue and weakness, and sleep-related problems[40-42]. For that reason, the NIH minimal dataset could still provide useful information into important domains other than pain and functioning.

Methodological considerations

A variety of methods exist to measure responsiveness and determine the MCIC and measurement error of patient reported outcome measures, and results may vary significantly depending on which method is used. We determined MCICs with an anchor-based method because distribution-based methods do not take into account patients' perspective and are inappropriate to use when the magnitude of the effect of an intervention is unknown[26,43]. A second consideration was how to determine the optimal cut-off point of the ROC curve in order to best estimate the MCIC. Given that sensitivity and specificity in chronic conditions such as CLBP are often valued equally, the cut-point to the top-left corner of the ROC curve represents the optimal cut-off point for the MCIC. One method of estimation that is often used is by determining where the sum of 1-sensitivity and 1-specificity is at its smallest[37,44,45]. However, the most efficient way to choose a cut-point closest to the top-left corner of the ROC curve is by first squaring the 1-sensitivity and 1-specificity terms[28]. Had we chosen to utilize the first method, we would have found a similar MCIC for the Impact stratification but a larger MCIC for the total outcome score; 17 instead of 14 points. SEM was based on the variability between time points and variability caused by random error in the "unchanged" patients in our cohort[9,46]. SEM can also be calculated with $SD\sqrt{(1-ICC)}$, where the ICC can be obtained from a test-retest study with a similar sample. While it can have a significant impact on the magnitude of the SEM and SDC, studies often differ in

which standard deviation they use in their calculations. By using baseline SD[47,48] we would have found a SEM and SDC of approximately 14 and 21 points for the total outcome score and 7 and 10 points for the Impact Stratification. A pooled SD, obtained from an ANOVA analysis with the "unchanged" patients[49], would produce a substantially lower SEM and SDC of approximately 8 and 12 points for the total outcome score and 3 and 7 points for the Impact Stratification.

A limitation of this study is the generalizability of the results in terms of CLBP severity and level of care. Also, patients were recruited from a single clinical research facility, which could further limit external validity. Our patient sample scores similar on pain (NRS: 6.7±1.8) and disability (PDI: 38.0±14.1) compared to Dutch patients with chronic pain referred to pain rehabilitation[50], but score much higher compared to Dutch workers with chronic musculoskeletal pain who do not seek specialty care (NRS: 4.6±2.1; PDI: 19.1±11.1)[51]. However, the baseline value analyses in the present study do provide us with more insight into responsiveness and MCIC for different CLBP severity. Future research should further explore longitudinal validity of the NIH minimal dataset in patient who receive primary and secondary level LBP care.

The GPE consists of one item only and may not be very representative for individual differences in what is perceived as a change in treated complaints. Furthermore, with a follow-up period of 12 months there is a fairly high risk of recall bias[52] or response-shift, where rating results might be influenced by functional status at discharge[39,53]. We do not know to what extent and in what direction this effect might have influenced the results of this study. As far as we know, no better alternative exists for an external criterion that is known to correlate with pain, disability, and quality of life measures.

Conclusion

The NIH minimal dataset is responsive in patients with CLBP seeking tertiary multispecialty care. A change of 14 points on the total outcome score and 7.5 points on Impact Stratification can be considered clinically important. MCIC depends among others on baseline values and the method that is chosen to determine the optimal cut-off point. Furthermore, individual change scores have to be interpreted with caution due to a risk of measurement error.

References

1. Deyo RA, Dworkin SF, Amtmann D, et al. Report of the NIH task force on research standards for chronic low back pain. *Spine J*. 2014;14(8):1375-1391.
2. Deyo RA, Dworkin SF, Amtmann D, et al. Focus article: Report of the NIH task force on research standards for chronic low back pain. *Eur Spine J*. 2014;23(10):2028-2045.
3. Deyo RA, Dworkin SF, Amtmann D, et al. Report of the NIH task force on research standards for chronic low back pain. *Spine (Phila Pa 1976)*. 2014;39(14):1128-1143.
4. Lacasse A, Roy JS, Parent AJ, et al. The canadian minimum dataset for chronic low back pain research: A cross-cultural adaptation of the national institutes of health task force research standards. *CMAJ Open*. 2017;5(1):E237-E248.
5. Noormohammadpour P, Tavana B, Mansournia MA, et al. Translation, cross-cultural adaptation and validation of the farsi version of NIH task force's recommended multidimensional minimal dataset for research on chronic low back pain. *Spine (Phila Pa 1976)*. 2018;43(9):E537-E544.
6. Boer A, Dutmer AL, Schiphorst Preuper HR, et al. Measurement properties of the NIH-minimal dataset dutch language version in patients with chronic low back pain. *Spine (Phila Pa 1976)*. 2017;42(19):1472-1477.
7. van Kampen DA, Willems WJ, van Beers LW, et al. Determination and comparison of the smallest detectable change (SDC) and the minimal important change (MIC) of four-shoulder patient-reported outcome measures (PROMs). *J Orthop Surg Res*. 2013;8:40-799X-8-40.
8. Mokkink LB, Terwee CB, Knol DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. *BMC Med Res Methodol*. 2010;10:22-2288-10-22.

9. Terwee CB, Roorda LD, Knol DL, et al. Linking measurement error to minimal important change of patient-reported outcomes. *J Clin Epidemiol*. 2009;62(10):1062-1067.
10. Schunemann HJ, Akl EA, Guyatt GH. Interpreting the results of patient reported outcome measures in clinical trials: The clinician's perspective. *Health Qual Life Outcomes*. 2006;4:62-7525-4-62.
11. Ostelo RW, Deyo RA, Stratford P, et al. Interpreting change scores for pain and functional status in low back pain: Towards international consensus regarding minimal important change. *Spine (Phila Pa 1976)*. 2008;33(1):90-94.
12. Bombardier C, Hayden J, Beaton DE. Minimal clinically important difference. low back pain: Outcome measures. *J Rheumatol*. 2001;28(2):431-438.
13. *Internationaal richtsnoer voor 'good clinical practice' voor het onderzoek met geneesmiddelen; vertaling naar de Nederlandse praktijk.* rev version ed. Den Haag: GCP Begeleidingscommissie; 2003.
14. Cella D, Riley W, Stone A, et al. The patient-reported outcomes measurement information system (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. *J Clin Epidemiol*. 2010;63(11):1179-1194.
15. Tait RC, Chibnall JT, Krause S. The pain disability index: Psychometric properties. *Elsevier*. 1990;40:171-182.
16. Soer R, Koke AJ, Vroomen PC, et al. Extensive validation of the pain disability index in 3 groups of patients with musculoskeletal pain. *Spine (Phila Pa 1976)*. 2013;38(9):E562-8.
17. EuroQol Group. EuroQol--a new facility for the measurement of health-related quality of life. the EuroQol group. *Health Policy*. 1990;16(3):199-208.

18. Lamers LM, McDonnell J, Stalmeier PF, et al. The dutch tariff: Results and arguments for an effective design for national EQ-5D valuation studies. *Health Econ*. 2006;15(10):1121-1132.
19. Soer R, Reneman MF, Speijer BL, et al. Clinimetric properties of the EuroQol-5D in patients with chronic low back pain. *Spine J*. 2012;12(11):1035-1039.
20. Soer R, Reneman MF, Vroomen PC, et al. Responsiveness and minimal clinically important change of the pain disability index in patients with chronic back pain. *Spine (Phila Pa 1976)*. 2012;37(8):711-715.
21. Pengel LH, Refshauge KM, Maher CG. Responsiveness of pain, disability, and physical impairment outcomes in patients with low back pain. *Spine (Phila Pa 1976)*. 2004;29(8):879-883.
22. Stewart M, Maher CG, Refshauge KM, et al. Responsiveness of pain and disability measures for chronic whiplash. *Spine (Phila Pa 1976)*. 2007;32(5):580-585.
23. Costa LO, Maher CG, Latimer J, et al. Clinimetric testing of three self-report outcome measures for low back pain patients in brazil: Which one is the best? *Spine (Phila Pa 1976)*. 2008;33(22):2459-2463.
24. Kamper SJ, Maher CG, Mackay G. Global rating of change scales: A review of strengths and weaknesses and considerations for design. *J Man Manip Ther*. 2009;17(3):163-170.
25. Terwee CB, Bot SD, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol*. 2007;60(1):34-42.
26. Mokkink LB, Terwee CB, Knol DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. *BMC Med Res Methodol*. 2010;10:22.

27. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol*. 2010;63(7):737-745.
28. Froud R, Abel G. Using ROC curves to choose minimally important change thresholds when sensitivity and specificity are valued equally: The forgotten lesson of pythagoras. theoretical considerations and an example application of change in health status. *PLoS One*. 2014;9(12):e114468.
29. de Vet HC, Terwee CB, Knol DL, et al. When to use agreement versus reliability measures. *J Clin Epidemiol*. 2006;59(10):1033-1039.
30. Deyo RA, Katrina R, Buckley DI, et al. Performance of a patient reported outcomes measurement information system (PROMIS) short form in older adults with chronic musculoskeletal pain. *Pain Med*. 2016;17(2):314-324.
31. Purvis TE, Neuman BJ, Riley LH 3rd, et al. Discriminant ability, concurrent validity, and responsiveness of PROMIS health domains among patients with lumbar degenerative disease undergoing decompression with or without arthrodesis. *Spine (Phila Pa 1976)*. 2018.
32. Ostelo RW, Deyo RA, Stratford P, et al. Interpreting change scores for pain and functional status in low back pain: Towards international consensus regarding minimal important change. *Spine (Phila Pa 1976)*. 2008;33(1):90-94.
33. 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 (Phila Pa 1976)*. 2006;31(5):578-582.

34. Demoulin C, Ostelo R, Knottnerus JA, et al. Quebec back pain disability scale was responsive and showed reasonable interpretability after a multidisciplinary treatment. *J Clin Epidemiol*. 2010;63(11):1249-1255.

35. Ostelo RW, Swinkels-Meewisse IJ, Knol DL, et al. Assessing pain and pain-related fear in acute low back pain: What is the smallest detectable change? *Int J Behav Med*. 2007;14(4):242-248.

36. Frahm Olsen M, Bjerre E, Hansen MD, et al. Minimum clinically important differences in chronic pain vary considerably by baseline pain and methodological factors: Systematic review of empirical studies. *J Clin Epidemiol*. 2018;101:87-106.e2.

37. de Vet HC, Ostelo RW, Terwee CB, et al. Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. *Qual Life Res*. 2007;16(1):131-142.

38. Lauridsen HH, Hartvigsen J, Manniche C, et al. Responsiveness and minimal clinically important difference for pain and disability instruments in low back pain patients. *BMC Musculoskelet Disord*. 2006;7:82-2474-7-82.

39. Wang YC, Sindhu BS, Kapellusch J, et al. Global rating of change: Perspectives of patients with lumbar impairments and of their physical therapists. *Physiother Theory Pract*. 2018:1-9.

40. Turk DC, Dworkin RH, Revicki D, et al. Identifying important outcome domains for chronic pain clinical trials: An IMMPACT survey of people with pain. *Pain*. 2008;137(2):276-285.

41. O'Brien EM, Staud RM, Hassinger AD, et al. Patient-centered perspective on treatment outcomes in chronic pain. *Pain Med*. 2010;11(1):6-15.

42. Robinson ME, Brown JL, George SZ, et al. Multidimensional success criteria and expectations for treatment of chronic pain: The patient perspective. *Pain Med*. 2005;6(5):336-345.
43. Musoro ZJ, Hamel JF, Ediebah DE, et al. Establishing anchor-based minimally important differences (MID) with the EORTC quality-of-life measures: A meta-analysis protocol. *BMJ Open*. 2018;8(1):e019117-2017-019117.
44. de Vet HC, Terluin B, Knol DL, et al. Three ways to quantify uncertainty in individually applied "minimally important change" values. *J Clin Epidemiol*. 2010;63(1):37-45.
45. Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3(1):32-35.
46. de Vet HC, Terwee CB, Knol DL, et al. When to use agreement versus reliability measures. *J Clin Epidemiol*. 2006;59(10):1033-1039.
47. Soer R, Koke AJ, Vroomen PC, et al. Extensive validation of the pain disability index in 3 groups of patients with musculoskeletal pain. *Spine (Phila Pa 1976)*. 2013;38(9):E562-8.
48. Furlan L, Sterr A. The applicability of standard error of measurement and minimal detectable change to motor learning research-A behavioral study. *Front Hum Neurosci*. 2018;12:95.
49. Beemster T, van Bennekom C, van Velzen J, et al. The interpretation of change score of the pain disability index after vocational rehabilitation is baseline dependent. *Health Qual Life Outcomes*. 2018;16(1):182-018-1000-1.
50. Koke AJ, Smeets RJ, Schreurs KM, et al. Dutch dataset pain rehabilitation in daily practice: Content, patient characteristics and reference data. *Eur J Pain*. 2017;21(3):434-444.

51. de Vries HJ, Reneman MF, Groothoff JW, et al. Self-reported work ability and work performance in workers with chronic nonspecific musculoskeletal pain. *J Occup Rehabil*. 2013;23(1):1-10.

52. Herrmann D. Reporting current, past, and changed health status. what we know about distortion. *Med Care*. 1995;33(4 Suppl):AS89-94.

53. Schwartz CE, Sprangers MA. Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. *Soc Sci Med*. 1999;48(11):1531-1548.

Table 1: Patient Characteristics

Characteristic	Patients (n = 223)
Age, mean years ± SD	49.7 ± 11.9
Sex, n female (%)	130 (58)
Duration LBP, n (%)	
12 weeks - 1 year	34 (15)
1 year – 5 years	65 (29)
> 5 years	124 (56)
Education level, n (%)	
No education	2 (1)
Low	81 (36)
Middle	78 (35)
High	45 (20)
Other/unknown	17 (8)
Work status, n (%)	
Working	65 (29)
Partial sick leave	34 (15)
Sick leave	39 (18)
No job	85 (38)
Baseline measures of pain, disability, and quality of life	
NRS pain (0-10), mean ± SD	6.6 ± 1.6
NIH total outcome Score (0-100), mean ± SD	44.3 ± 13.9
NIH Impact Stratification (8-50), mean ± SD	34.4 ± 7.4
PDI score (0-70), mean ± SD	36.4 ± 14.1
EQ5D score (-0.33-1.00), mean ± SD	0.48 ± 0.29

N=number of patients; SD=standard deviation; LBP=low back pain; NRS=numeric rating scale; NIH=National Institutes of Health; PDI=Pain Disability Index; EQ5D=EuroQol-5D

Table 2: Responsiveness and Minimal Clinically Important Change of the NIH Minimal Dataset (n=223)

	Total Outcome Score	Factor 1: Pain intensity and interference	Factor 2: Pain history	Factor 4: Depression and catastrophizing	Factor 5: Physical function	Factor 6: Sleep disturbance	Factor 7: Lifestyle	Impact Stratification
Improved patients, n (%)	50 (29)							
Scores								
Score T0, mean ± SD	44.3 ± 13.9	68.6 ± 17.8	53.2 ± 19.6	35.6 ± 26.2	49.8 ± 26.4	50.5 ± 22.4	8.1 ± 19.5	34.4 ± 7.4
Min-max	7.7 – 75.1	15 – 98.3	0 – 100	0 – 100	0 - 100	6.25 - 100	0 – 100	14 - 49
Score T1, mean ± SD	36.9 ± 16.9	50.6 ± 23.7	45.2 ± 23.0	28.0 ± 24.9	40.0 ± 27.5	46.4 ± 22.8	10.8 ± 23.4	28.1 ± 9.7
Min-max	4.1 – 80.4	0 – 96.7	0 – 100	0 – 100	0 – 100	0 – 100	0 – 100	8 – 48
Mean change ± SD	7.4 ± 13.6**	18.0 ± 21.9**	8.0 ± 19.4**	7.6 ± 23.8**	9.8 ± 25.9**	4.1 ± 19.7*	-2.8 ± 18.9**	6.3 ± 8.6**
95% CI of mean change	5.6 – 9.2	15.1 – 20.8	5.4 – 10.5	4.5 – 10.7	6.3 – 13.2	1.5 – 6.7	-5.3 – 0.3	5.2 – 7.5
Change (%)	16.7	26.2	15.0	21.3	19.7	8.1	-34.6	18.3
Responsiveness								
AUC	0.84	0.91	0.75	0.70	0.78	0.65	0.49	0.91
95% CI	0.78 – 0.91	0.86 – 0.95	0.67 – 0.83	0.62 – 0.79	0.70 – 0.86	0.57 – 0.73	0.40 – 0.59	0.86 – 0.96
MCIC								
OCP	14.2	25.4	16.7	16.8	16.7	3.1	-8.3	7.5
Sensitivity (%); specificity (%)	72;82	88;84	66;78	58;72	70;76	72;56	84;13	96;78

Total Outcome Score indicates total outcome score of the NIH minimal dataset (scale 0-100 points); all factors (scale 0-100); Impact Stratification (scale 8-50 points); n=number of patients; SD=standard deviation; CI=confidence interval; AUC=area under the receiver operating characteristic (ROC) curve; MCIC=minimal clinically important change; OCP=optimal cut-off point of the ROC curve.

Factor 3 not included due to the fact that the corresponding items are only administered at baseline (T0).

*** significant change between T0 and T1 ($p<0.01$).*

** significant change between T0 and T1 ($p<0.05$).*

Table 3: Responsiveness and Minimal Clinically Important Change of the NIH Minimal Dataset Total Outcome Score and Impact Stratification per Baseline-score Group

	Total Outcome Score			Impact Stratification		
	Baseline Tertile 1	Baseline Tertile 2	Baseline Tertile 3	Baseline Mild (8-27)	Baseline Moderate (28-34)	Baseline Severe (≥35)
Patients, n	74	74	75	37	73	113
Improved patients, n (%)	19 (26)	18 (24)	13 (17)	12 (32)	19 (26)	19 (17)
Scores						
Score T0, mean ± SD	28.8 ± 7.6	44.3 ± 3.7	59.5 ± 6.0	22.7 ± 3.6	31.2 ± 2.0	40.3 ± 3.8
Min-max	7.7 – 38.0	38.1 – 50.1	50.1 – 75.1	14 – 27	28 - 34	35 - 49
Score T1, mean ± SD	26.3 ± 12.6	35.4 ± 15.3	48.8 ± 14.6	20.8 ± 8.9	25.1 ± 7.5	32.3 ± 8.9
Min-max	4.9 – 59.0	4.1 – 64.5	9.0 – 80.4	8 – 37	10 – 40	8 – 48
Mean change ± SD	2.6 ± 11.3	9.0 ± 14.0**	10.7 ± 14.2**	1.9 ± 8.3	6.1 ± 7.8**	8.0 ± 8.7**
95% CI of mean change	-0.1 – 5.2	5.7 – 12.2	7.5 – 14.0	-0.9 – 4.7	4.2 – 7.9	6.3 – 9.6
Change (%)	9.0	20.3	18.0	8.4	19.6	19.9
Responsiveness						
AUC	0.76	0.94	0.91	0.91	0.92	0.97
95% CI	0.64 – 0.88	0.88 – 1.00	0.82 – 0.99	0.77 – 1.00	0.85 – 0.99	0.93 – 1.00
MCIC						
OCP	6.9	19.7	17.1	7.5	11.5	12.5
Sensitivity (%); specificity (%)	74;71	94;91	85;86	92;96	84;91	95;88

Total Outcome Score indicates total outcome score of the NIH minimal dataset (scale 0-100 points); Impact Stratification (scale 8-50 points); n=number of patients; SD=standard deviation; CI=confidence interval; AUC=area under the receiver operating characteristic (ROC) curve; MCIC=minimal clinically important change; OCP=optimal cut-off point of the ROC curve.

*** significant change between T0 and T1 ($p<0.01$).*