



ISSLS prize in clinical science 2020: the reliability and interpretability of score change in lumbar spine research

C. Parai^{1,2} · O. Hägg² · B. Lind^{1,2} · H. Brisby^{1,3}

Received: 6 October 2019 / Revised: 6 October 2019 / Accepted: 15 November 2019 / Published online: 23 November 2019
© The Author(s) 2019

Abstract

Purpose A statistically significant score change of a PROM (Patient-Reported Outcome Measure) can be questioned if it does not exceed the clinically Minimal Important Change (MIC) or the SDC (Smallest Detectable Change) of the particular measure. The aim of the study was to define the SDC of three common PROMs in degenerative lumbar spine surgery: Numeric Rating Scale (NRS_{BACK/LEG}), Oswestry Disability Index (ODI) and Euroqol-5-Dimensions (EQ-5D_{INDEX}) and to compare them to their MICs. The transition questions Global Assessment (GA_{BACK/LEG}) were also explored.

Methods Reliability analyses were performed on a test–retest population of 182 symptomatically stable patients, with similar characteristics as the Swespine registry population, who underwent surgery for degenerative lumbar spine conditions 2017–2018. The MIC values were based on the entire registry ($n = 98,732$) using the ROC curve method. The ICC for absolute agreement was calculated in a two-way random-effects single measures model. For categorical variables, weighted kappa and exact agreement were computed.

Results For the NRS, the SDC exceeded the MIC (NRS_{BACK}:3.6 and 2.7; NRS_{LEG}: 3.7 and 3.2, respectively), while they were of an equal size of 18 for the ODI. The gap between the two estimates was remarkable in the EQ-5D_{INDEX}, where SDC was 0.49 and MIC was 0.10. The GA_{BACK/LEG} showed an excellent agreement between the test and the retest occasion.

Conclusion For the tested PROM scores, the changes must be considerable in order to distinguish a true change from random error in degenerative lumbar spine surgery research.

Graphic abstract

These slides can be retrieved under Electronic Supplementary Material.

Parai C, Hägg O, Lind B, Brisby H (2019) ISSLS PRIZE IN CLINICAL SCIENCE 2020: The reliability and interpretability of score change in lumbar spine research. Eur Spine J; Springer

Parai C, Hägg O, Lind B, Brisby H (2019) ISSLS PRIZE IN CLINICAL SCIENCE 2020: The reliability and interpretability of score change in lumbar spine research. Eur Spine J; Springer

Parai C, Hägg O, Lind B, Brisby H (2019) ISSLS PRIZE IN CLINICAL SCIENCE 2020: The reliability and interpretability of score change in lumbar spine research. Eur Spine J; Springer

Keywords Test–retest reliability · Patient-reported outcome · NRS ODI · EQ-5D · Pragmatic clinical study

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s00586-019-06222-8>) contains supplementary material, which is available to authorized users.

✉ C. Parai
catharina.parai@spinecenter.se

Extended author information available on the last page of the article

Introduction

During 2000 to 2009, there was an annual publication of 40 RCTs on lumbar pain, using a Patient-Reported Outcome Measure (PROM). The most common measures were the Oswestry Disability Index (ODI—physical function), the

Numeric Rating Scale (NRS—pain) and the Euroqol-5-Dimensions (EQ-5D—quality of life) [1].

When a PROM is used repeatedly on the same patient, a measurement error will be present because of natural fluctuations in symptoms, variation in the measurement process, or both. A useful way of presenting the measurement error is the Smallest Detectable Change (SDC). It is described by Polit and Yang as a change in score of sufficient magnitude that the probability of it being the result of random error is low [2]. In trials, where a measurement of change is involved, it is practical to refer to a repeatability parameter such as the SDC, which is in the units of the PROM in question.

The SDC is a measure of the reliability of a PROM, based on the measurement error and repeatability of each instrument. Recently published reviews found that studies exploring such measurement properties were few and of inadequate quality [3–5].

A statistically significant change in outcome does not necessarily mean that it is of interest in real life. A person's *opinion* about the smallest score change is named the Minimal Important Change (MIC) [6]. For many years, there has been a conceptual confusion around the many measurements of change parameters defining the cut-off in a PROM score that distinguishes a success from failure [7–12]. Terwee et al. [13] have emphasized the important link between SDC and MIC.

The aim of this study was to define the SDC in the most commonly used outcome measures in degenerative lumbar spine surgery and compare them to the MIC.

Patients and methods

Outcome variables

The Numeric Rating Scale for back and leg pain, respectively, ($NRS_{BACK/LEG}$), the Oswestry Disability Index (ODI), version 2.1a, and the European Quality of life questionnaire (EQ-5D_{INDEX}) are well known and described in detail earlier [1].

The Global Assessment of back and leg pain, respectively, ($GA_{BACK/LEG}$) [14] assesses patients' retrospective perception of treatment effect. The question is worded: "How is your back/leg pain today as compared to before you had your back surgery?" with 6 response options: 0/Had no back/leg pain, 1/Completely pain free, 2/Much Better, 3/Somewhat Better, 4/Unchanged, 5/Worse.

The first question of the Short-Form 36 questionnaire (SF36_{GH}) [15] was added to reveal changes in global health during the retest period. The question is worded: "In general, would you say your health is" with response options: Excellent/Very Good/Good/Fair/Poor.

The MIC population

MIC computations were based on the entire Swespine register [16]. Table 1 presents anthropometrics, baseline data and 1-year follow-up of the degenerative lumbar spine population, operated 1998–2017 ($n=98,732$). Adults, with either of the three degenerative diagnoses, lumbar disk herniation, lumbar spinal stenosis or degenerative disk disease, were included.

The retest population

The study participants were collected consecutively at Stockholm Spine Center and Spine Center Göteborg between November 2017 and May 2019. In order to cover as much of the range of each PROM scale as possible, they were collected from both the waiting list (pre-op group) and from those followed up 1 year after surgery (post-op group). At least 30 individuals from each of the three diagnoses groups were obtained.

Table 1 Baseline data of the retest and Swespine populations, respectively

	Retest population $n=182$	Swespine population $n=98,732$
Female	54%	51%
Mean age	60± 14 years	57± 17 years
Retirement pension	35%	38%
Unemployed	7%	11%
Smoker	0.6%	15%
Duration of back pain > 1 year	67%	66%
Duration of leg pain > 1 year	54%	57%
Previous spine surgery	20%	20%
Pre-op NRS back pain	5.7± 2.4	5.5± 4.8
Pre-op NRS leg pain	6.0± 2.5	6.1± 2.7
Pre-op ODI	38± 14	45± 17
Pre-op EQ-5D index	0.43± 0.31	0.32± 0.33
Post-op NRS back pain	3.2± 2.9	3.1± 2.9
Post-op NRS leg pain	2.7± 2.9	2.9± 3.0
Post-op ODI	18± 17	25± 19
Post-op EQ-5D index	0.72± 0.30	0.64± 0.31

NRS, Numeric Rating Scale (0–10); *ODI*, Oswestry Disability Index (0–100); *EQ-5D*, Euroqol-5 Dimensions (−0.59–1)

Pre-op PROM score means in the retest population are based on the number of individuals in the pre-op group ($n=83$) and post-op means on the number of individuals in the post-op group ($n=99$)

PROM score means in the Swespine population are based on the number of respondents on each questionnaire, which varied between 59,855 and 77,485

The pre-op group filled out the first booklet (T_1) at the clinic on the day they were listed for surgery. The second booklet (T_2) was sent by mail 1 week later, and the respondents were asked to return the form within 5 days. One reminder was sent after 1 week.

In the post-op group, a request for study participation was added to the 1-year Swespine follow-up booklet (T_1). One week after the booklet was registered at the Swespine office, the second questionnaire (T_2) was sent out by mail, with a request to return the form within 5 days. Inclusion to the pre-op group stopped as the total number of participants exceeded 30 in all three diagnoses. For the analyses, the pre-op and the post-op groups, as well as the diagnoses, were merged.

The time interval between the two points of estimation, T_1 and T_2 , was within 10 to 35 days. The difference in PROM score for each participant between T_1 and T_2 was plotted against the time interval and correlated in Spearman rank analyses to check whether the number of days between T_1 and T_2 had an influence on the PROM score or not.

The occurrence of systematic differences between T_1 and T_2 was examined using the Sign test for categorical data (i.e., $GA_{\text{BACK/LEG}}$ and $SF_{36_{\text{GH}}}$) and the Wilcoxon's sign rank test for continuous data (i.e., $NRS_{\text{BACK/LEG}}$, ODI and $EQ-5D_{\text{INDEX}}$).

A maximum of two missing items was accepted for the ODI and zero missing items for the remaining PROMs, according to published score algorithms [17, 18].

The study was conducted according to the COSMIN checklist, boxes B, C, and J [6].

Descriptive data are presented as means (\pm SD) or numbers (%).

MIC

The MIC estimates were previously calculated for the diagnosis groups LDH, LSS, and DDD [19] using the anchor-based ROC curve method [20]. In the current study, MIC values without stratification for diagnosis were added. The measure used as gold standard was the GA, which has been shown to have an acceptable correlation to the instruments at issue [14]. Patients' self-assessments on the GA as either "pain free" or "much better" was considered an important improvement (i.e., equal to, or above the MIC). The ability of each PROM to distinguish between improved and not improved was measured by the Area Under the ROC Curve (AUC), with an acceptable level of 0.70. The cut-off score defining the MIC also represents the level where the sensitivity and specificity of the PROMs are mutually maximized. The probability that a patient reaching the MIC will also express an important improvement on the GA is called the positive predictive value (PPV). The probability that a patient not reaching the MIC will express a non-important

improvement on the GA is called the negative predictive value (NPV) [21].

SDC

The reliability of change scores the Smallest Detectable Change ($SDC = 1.96 \times \sqrt{2} \times SEM$ (Standard Error of Measurement)).

SEM

Agreement between T_1 and T_2 was expressed as the intra-individual standard deviation, also known as the Standard Error of Measurement [13]. The SEM is a standard error in an observed score that obscures the true score and is given in the units of the PROM. The $SEM = \sqrt{\text{intra individual variance}}$ of an ANOVA analysis. The difference between a subject's PROM score and the true value would be expected to be within $\pm 1.96SEM$ for 95% of the individuals. The assumption that the score distribution is unrelated to the magnitude of the measurement (heteroscedasticity) was checked by plotting the individual patient's standard deviations against his or her means.

ICC

The reliability parameter was the Intra-class Correlation Coefficient, ICC. ICC estimates and their 95% CI were calculated using an absolute agreement, two-way random-effects single measures model. Based on the 95% CI of the ICC, estimate values less than 0.40 indicate poor reliability, while estimates between 0.4 and 0.59 indicate fair, 0.6–0.74 good and 0.75–1.00 excellent reliability [22]. The relation of the ICC to the SEM is described as $SEM = SD \sqrt{1-ICC}$.

Kappa

The reliability measure weighted kappa was calculated for the categorical variables (i.e., $GA_{\text{BACK/LEG}}$ and $SF36_{\text{GH}}$). An instrument is reliable when the kappa is above 0.70 [6]. Since these instruments have several ordinal response options, kappa was calculated using the weighting scheme of quadratic weights which is mathematically identical to an ICC of absolute agreement. Further, overall agreement between T_1 and T_2 as well as the proportion of respondents indicating a better outcome at T_1 than at T_2 or vice versa were calculated.

IBM SPSS Statistics for Windows, Version 24.0. Armonk, NY: IBM Corp. was used in all the statistical analyses apart from the MIC computations, where JMP[®], Version 13.1 SAS Institute Inc., Cary, NC, 1989-2019, was used.

Ethical considerations

Informed consent was obtained from all participants in Swespine, and written consent was acquired from the participants in the retest study. This research project was approved by the regional ethical review board.

Results

Descriptives

In total, 248 participants filled out the booklet at T1. Both questionnaires were returned by 182 (74.6%) participants, 83 from the pre-op group and 99 from the post-op group. Table 1 presents demographics and mean PROM scores at baseline and at the 1-year follow-up for the retest group and for the Swespine population.

Timing of measurements

The time interval between T1 and T2 was 20 ± 8 days. The number of days between T1 and T2 did not correlate with the PROM scores (Spearman rank correlation coefficient for ODI: -0.07 ; NRS_{BACK} : 0.06 ; NRS_{LEG} : -0.08 ; $EQ-5D_{INDEX}$: -0.03 ; GA_{BACK} : 0.045 ; GA_{LEG} : -0.144 ; Satisfaction: 0.128 ; $SF36_{GA}$: -0.064). Figure 1 visualizes this pattern in a scatter plot, exemplified by the ODI.

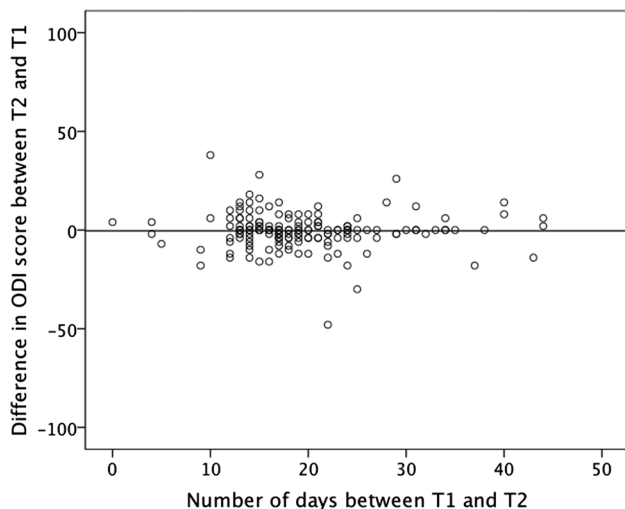


Fig. 1 Scatter plot illustrating a non-correlation between the time interval and ODI score. The horizontal line is the mean score difference between the two occasions of measurement ($T2-T1 = -0.25$), $n = 169$. The same pattern was seen for $NRS_{BACK/LEG}$ and the $EQ-5D_{INDEX}$

Measurement error and score change reliability

There were no statistically significant systematic differences between T1 and T2 as measured by the Wilcoxon sign rank test (NRS , ODI , $EQ-5D_{INDEX}$) and the Sign test (GA , Satisfaction, $SF-36_{GH}$) for any of the PROMs.

The data were not found to be heteroscedastic, meaning that the measurement error appeared to be uniform across scale values. Table 2 presents reliability measures of each PROM demonstrating excellent or good reliability and large SDCs for all prospective PROMs. The influence of random error on the SDCs is illustrated in Fig. 2, with the ODI illustrating the typical pattern.

The MIC calculations were based on the lumbar Swespine register population, stratified for diagnosis [19]. In Table 3, the SDCs are compared to these MIC values. For $NRS_{BACK/LEG}$ and ODI the SDCs exceeded the MICs to some extent. As for the $EQ-5D$, the difference was more remarkable.

In Table 4, the SDCs are compared to the MIC values that were calculated for the entire lumbar Swespine population. The SDC for both NRS scales exceeded the corresponding MICs, while the SDC and MIC were equal for the ODI . The considerable gap between the SDC and MIC for the $EQ-5D_{INDEX}$ remained. The AUCs were all above 0.70. The ODI had the best ability to correctly classify patients as importantly improved according to GA with a sensitivity of 76%. The specificity was similar for all PROMs. NRS_{LEG} reached the highest specificity (83%), indicating the best ability to correctly classify patients as not importantly improved.

The weighted kappa for the categorical variables $GA_{BACK/LEG}$ and $SF-36_{GH}$ were above the level of

Table 2 Measurement properties (reliability and measurement error parameters) of four PROMs

PROM	Parameter	
NRS_{BACK} $N = 172$	ICC	0.8 (0.75;0.86)
	SEM	1.3
	SDC	3.6
NRS_{LEG} $N = 172$	ICC	0.8 (0.77;0.87)
	SEM	1.3
	SDC	3.7
ODI $N = 169$	ICC	0.9 (0.84;0.91)
	SEM	6.5
	SDC	17.9
$EQ-5D_{INDEX}$ $N = 170$	ICC	0.7 (0.62;0.78)
	SEM	0.18
	SDC	0.49

NRS , Numeric Rating Scale for back or leg pain; ODI , Oswestry Disability Index; $EQ-5D$, Euroqol 5-Dimensions; ICC , Intra-Class Correlation Coefficient of absolute agreement; SEM , Standard Error of Measurement; SDC , Smallest Detectable Change = $1.96 \sqrt{2SEM}$

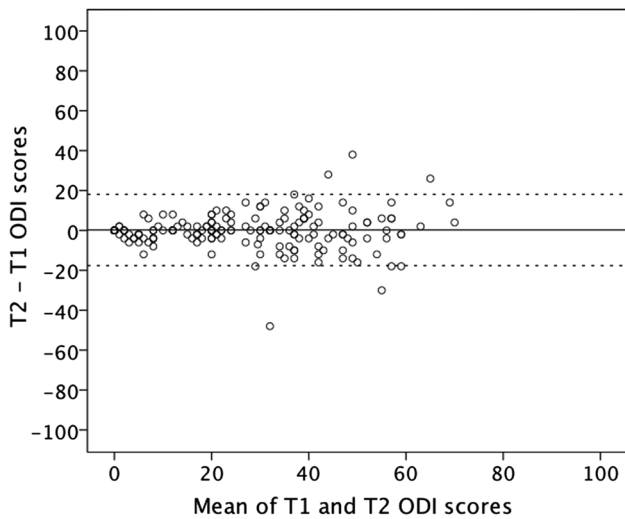


Fig. 2 A Bland–Altman plot of the ODI test–retest scores. The horizontal line close to “0” illustrates the mean difference in score between the test and the retest occasion. The upper dotted line is of interest if the concern is an improvement and the lower line if the research question is about a deterioration. Values within these limits are, with 95% confidence, due to random error, $n = 172$

Table 3 Measurement of change parameters (SDC, MIC) for PROMs in three lumbar spine conditions

Parameter of change	NRS _{BACK}	NRS _{LEG}	ODI	EQ-5D _{INDEX}
SDC	3.6	3.7	18	0.49
LDH group				
MIC	2.0	3.9	22	0.18
LSS group				
MIC	2.8	2.7	14	0.10
DDD group				
MIC	2.9	2.3	16	0.10

The MIC calculations were based on the Swespine population operated for LDH (ODI $n = 8359$; NRS-BACK $n = 9193$; NRS-LEG $n = 9778$; EQ-5D $n = 9214$), LSS (ODI $n = 17,549$; NRS-BACK $n = 16,038$; NRS-LEG $n = 16,083$; EQ-5D $n = 19,252$) and DDD (ODI $n = 5493$; NRS-BACK $n = 5519$; NRS-LEG $n = 4081$; EQ-5D $n = 5930$) 1998–2016 using the anchor-based ROC curve method

acceptance. The percentages of agreement are given in Table 5.

Discussion

This study found large SDCs, frequently exceeding tough MIC cut-off values, for some of the most commonly used PROMs in spine surgery research. The error was mainly due to a large intra-individual variation between the two test

Table 4 The SDC of four PROMs compared to their MIC values based on the entire Swespine lumbar population

Parameter of change	NRS _{BACK}	NRS _{LEG}	ODI	EQ-5D _{INDEX}
SDC	3.6	3.7	18	0.49
MIC	2.7	3.2	18	0.104
AUC	0.80	0.86	0.85	0.74
Se	0.64	0.74	0.76	0.60
Sp	0.81	0.83	0.80	0.80
PPV	0.86	0.88	0.87	0.77
NPV	0.57	0.64	0.65	0.63

The MIC calculations were based on the Swespine population operated for lumbar degenerative conditions during 1998–2016 using the anchor-based ROC-curve method. ODI $n = 29,944$; NRS-BACK $n = 33,001$; NRS-LEG $n = 32,935$; EQ-5D $n = 32,662$

Se, Sensitivity; Sp, Specificity; PPV, Positive Predictive Value; NPV, Negative Predictive Value

Table 5 Reliability of retrospective single-item questions

PROM	Exact agreement (%)	T1 > T2 (%)	T1 < T2 (%)	Weighted kappa
GA _{BACK} ($n = 96$)	74	12.5	13.5	0.86
GA _{LEG} ($n = 96$)	65.5	18	16.5	0.75
SF-36 _{GH} ($n = 94$)	69	19	12	0.81

Exact Agreement: proportion giving same response at T1 as at T2; T1 > T2: proportion that responded as having a better outcome at T1 than at T2; T1 < T2: proportion that responded as having a worse outcome at T1 than at T2%

occasions and not to systematic differences. It has important implications.

For instance, consider a trial exploring a possible difference in outcome between two groups undergoing posterolateral fusion with or without interbody fusion, and the outcome variable is NRS_{BACK}. Then—according to the present study—both groups need to reach a change of 3.6 before there is a 95% certainty that the change from baseline is not a mere chance. If—and only if—both groups reach this level of improvement, the research question can be answered.

In other studies on low back pain populations, using the same definition of SDC as in this paper, the SDCs were also rather high: 2.4–4.7 for NRS_{BACK}, 11–16.7 for ODI, and 0.28–0.58 for the EQ-5D [4, 5, 23].

The MIC corresponds to the minimal level of change that makes the efforts of the surgery worthwhile. A statistically detectable change does not reveal any information about its value in real life. That estimation has to be based on opinions of the persons undergoing the treatment. Accepting the opinion-based MIC does however not allow for the exclusion of the SDC!

If we recycle the example above but change the research question to whether there is a *clinically* important difference between the groups or not, a MIC in NRS_{BACK} of 2.9 must be reached by both groups before the question can be answered. Note that the answer should not be given in terms of a mean difference between the groups, but rather as the percentage in each group reaching the MIC cut-point. However, as the SDC was 3.6, a change of 2.9 may just be a measurement error—no matter the importance of personal opinions.

As long as it can be shown that the MIC estimate exceeds the SDC it can be used separately. But as soon as it is the opposite way, both the SDC and the MIC should be presented in such a manner that the reader can get a clear picture of the true degree of change. This simultaneous usage of both a distribution-based cut-off value and anchor-based estimate has earlier been advocated by Terwee and colleagues [13].

If the SDC by far outreaches the MIC, as was the case for EQ-5D, the use of that PROM should not be accepted, simply because the size of the error is too large to make sound inferences. Why this was the case for EQ-5D in the current study is not clear. Variations in measurement-of-change estimates for this particular PROM stretch from 0.15 to 0.45 [24]. In this study, the SDC was 0.48 and the MIC was 0.10–0.18 depending on which diagnosis group the calculations were based. A possible explanation is that the preference-based summary index systematically divides the population in two, making it difficult to define an SDC, which is based on dispersion.

Based on the large Swespine database, the MIC values in this study may be considered credible. However, it must be remembered that the MIC is anchored to a retrospective single-item transition question, requiring that each patient remembers his or her health state prior to their operation. Also demanded is an honest response about the degree of improvement or deterioration where the patient excludes factors such as disappointment, gratitude, insurance, sick leave or work-related issues. The human nature probably makes sure that recall bias and response shift will always have an impact on the response to these types of questions.

The PPV of 0.88 for NRS_{LEG} indicates the probability that patients with a change exceeding the MIC, also classified themselves as being importantly improved on the anchor. The NPV of 0.64 is the probability that patients with a change less than the MIC self-assessed a non-important improvement on the anchor.

The reliability of the retrospective single-item questions, interpreted by their weighted kappa values, was almost perfect (above 0.8) or substantial (0.75) according to Landis and Koch [25]. A high weighted kappa also indicates that misclassifications mainly occurred between adjacent response options.

Conclusion

A consequence of large measurement errors in PROMs, is the need of considerable change in outcome in order to distinguish a random error from true change.

Acknowledgements Open access funding provided by University of Gothenburg. The authors would like to thank the Neubergh foundation, the Region and the Regional Agreement on Medical Training and Clinical Research (ALF) and GHP Spine Center Göteborg for their financial support.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.


References

1. Chapman JR, Norvell DC, Hermsmeyer JT, Bransford RJ, DeVine J, McGirt MJ, Lee MJ (2011) Evaluating common outcomes for measuring treatment success for chronic low back pain. *Spine* 36(21 Suppl):S54–S68. <https://doi.org/10.1097/BRS.0b013e31822ef74d>
2. Polit DF, Yang FM (2016) Measurement and the measurement of change: a primer for the health professions. Wolters Kluwer, Philadelphia
3. Chiarotto A, Terwee CB, Kamper SJ, Boers M, Ostelo RW (2018) Evidence on the measurement properties of health-related quality of life instruments is largely missing in patients with low back pain: a systematic review. *J Clin Epidemiol* 102:23–37. <https://doi.org/10.1016/j.jclinepi.2018.05.006>
4. Chiarotto A, Maxwell LJ, Ostelo RW, Boers M, Tugwell P, Terwee CB (2018) Measurement properties of visual analogue scale, numeric rating scale, and pain severity subscale of the brief pain inventory in patients with low back pain: a systematic review. *J Pain Off J Am Pain Soc*. <https://doi.org/10.1016/j.jpain.2018.07.009>
5. Chiarotto A, Maxwell LJ, Terwee CB, Wells GA, Tugwell P, Ostelo RW (2016) Roland-morris disability questionnaire and Oswestry disability index: which has better measurement properties for measuring physical functioning in nonspecific low back pain? *Syst Rev Meta Anal Phys Ther* 96(10):1620–1637. <https://doi.org/10.2522/ptj.20150420>
6. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HC (2010) The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual Life Res Int J Qual Life Asp Treat Care Rehabil* 19(4):539–549. <https://doi.org/10.1007/s11136-010-9606-8>
7. Wyrwich KW, Tierney WM, Wolinsky FD (1999) Further evidence supporting an SEM-based criterion for identifying

- meaningful intra-individual changes in health-related quality of life. *J Clin Epidemiol* 52(9):861–873
8. Wells G, Beaton D, Shea B, Boers M, Simon L, Strand V, Brooks P, Tugwell P (2001) Minimal clinically important differences: review of methods. *J Rheumatol* 28(2):406–412
 9. Terwee CB, Roorda LD, Dekker J, Bierma-Zeinstra SM, Peat G, Jordan KP, Croft P, de Vet HC (2010) Mind the MIC: large variation among populations and methods. *J Clin Epidemiol* 63(5):524–534. <https://doi.org/10.1016/j.jclinepi.2009.08.010>
 10. King MT (2011) A point of minimal important difference (MID): a critique of terminology and methods. *Expert Rev Pharmacoecon Outcomes Res* 11(2):171–184. <https://doi.org/10.1586/erp.11.9>
 11. Copay AG, Subach BR, Glassman SD, Polly DW Jr, Schuler TC (2007) Understanding the minimum clinically important difference: a review of concepts and methods. *Spine J Off J N Am Spine Soc* 7(5):541–546. <https://doi.org/10.1016/j.spinee.2007.01.008>
 12. Guyatt G, Walter S, Norman G (1987) Measuring change over time: assessing the usefulness of evaluative instruments. *J Chronic Dis* 40(2):171–178
 13. Terwee CB, Roorda LD, Knol DL, De Boer MR, De Vet HC (2009) Linking measurement error to minimal important change of patient-reported outcomes. *J Clin Epidemiol* 62(10):1062–1067. <https://doi.org/10.1016/j.jclinepi.2008.10.011>
 14. Parai C, Hagg O, Lind B, Brisby H (2018) The value of patient global assessment in lumbar spine surgery: an evaluation based on more than 90,000 patients. *Eur Spine J Off Publ Eur Spine Soc Eur Spinal Deform Soc Eur Sect Cerv Spine Res Soc* 27(3):554–563. <https://doi.org/10.1007/s00586-017-5331-0>
 15. Ware JE Jr, Sherbourne CD (1992) The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care* 30(6):473–483
 16. Stromqvist B, Fritzell P, Hagg O, Jonsson B, Sanden B (2013) Swespine: the Swedish spine register: the 2012 report. *Eur Spine J Off Publ Eur Spine Soc Eur Spinal Deform Soc Eur Sect Cerv Spine Res Soc* 22(4):953–974. <https://doi.org/10.1007/s00586-013-2758-9>
 17. Fairbank JC, Pynsent PB (2000) The Oswestry disability index. *Spine* 25(22):2940–2952 **discussion 2952**
 18. Group TE (1990) EuroQol—a new facility for the measurement of health-related quality of life. *Health policy* 16(3):199–208
 19. Parai C, Hagg O, Lind B, Brisby H (2019) Follow-up of degenerative lumbar spine surgery-PROMs stabilize after 1 year: an equivalence study based on Swespine data. *Eur Spine J Off Publ Eur Spine Soc Eur Spinal Deform Soc Eur Sect Cerv Spine Res Soc*. <https://doi.org/10.1007/s00586-019-05989-0>
 20. Deyo RA, Centor RM (1986) Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. *J Chronic Dis* 39(11):897–906
 21. Altman DG, Bland JM (1994) Diagnostic tests 2: predictive values. *BMJ (Clinical research ed)* 309(6947):102. <https://doi.org/10.1136/bmj.309.6947.102>
 22. Cicchetti DV (1994) Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess* 6(4):284
 23. van der Roer N, Ostelo RW, Bekkering GE, van Tulder MW, de Vet HC (2006) Minimal clinically important change for pain intensity, functional status, and general health status in patients with nonspecific low back pain. *Spine* 31(5):578–582. <https://doi.org/10.1097/01.brs.0000201293.57439.47>
 24. Coretti S, Ruggeri M, McNamee P (2014) The minimum clinically important difference for EQ-5D index: a critical review. *Expert Rev Pharmacoecon Outcomes Res* 14(2):221–233. <https://doi.org/10.1586/14737167.2014.894462>
 25. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* 33(1):159–174

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

C. Parai^{1,2}  · O. Hägg² · B. Lind^{1,2} · H. Brisby^{1,3}

¹ Department of Orthopaedics, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

² Spine Center Göteborg, Gothenburg, Sweden

³ Department of Orthopaedics, Sahlgrenska University Hospital, Gothenburg, Sweden