ARTICLE

A Systematic Review of Head-to-Head Comparison Studies of the Roland-Morris and Oswestry Measures' Abilities to Assess Change

Anastasia N.L. Newman, MSc(PT), MSc;‡ Paul W. Stratford, PT, MSc;*† Lori Letts, OT, PhD;* Gregory Spadoni, PT, MSc, FCAMPT*§*

ABSTRACT

Purpose: To determine if the sensitivity to change of Roland-Morris Questionnaire (RMQ) and Oswestry Disability Index (ODI) scores differ when applied to patients with low back pain (LBP). A secondary purpose was to critique the methodological rigour of the identified head-to-head comparison studies. ***Methods:*** A systematic review of five online databases was performed to locate head-to-head comparison studies of the RMQ and the ODI that assessed the sensitivity to change of the two measures. Studies were eligible if they met a pre-determined set of inclusion criteria. A newly developed quality criteria form was used to evaluate the methodological rigour of head-to-head comparison studies. ***Results:*** Nine articles met the inclusion criteria. Although there was a statistically significant difference in favour of the RMQ for two studies, there was no apparent consistent advantage of one measure over the other. Frequent methodological deficiencies included no formal sample size calculation, no formal between-measure comparison, and no independent reference standard. ***Conclusion:*** There was no consistent evidence supporting one measure over the other. Many studies displayed methodological deficiencies.

Key Words: self-report; sensitivity and specificity; systematic review; reproducibility of results.

RÉSUMÉ

Objectif : Déterminer si la sensibilité au changement des résultats au questionnaire Roland-Morris (*Roland-Morris Questionnaire,* RMQ) et au questionnaire d'incapacité d'Oswestry (*Oswestry Disability Index,* ODI) diffèrent lorsqu'on les applique aux patients qui souffrent de lombalgie. Comme objectif secondaire, réaliser une analyse critique de la rigueur méthodologique des études comparatives directes sélectionnées. ***Méthode :*** Une revue systématique de cinq bases de données en ligne a été réalisée pour rechercher des études comparatives directes du RMQ et de l'ODI qui évaluaient la sensibilité au changement de ces deux mesures. Les études étaient retenues si elles satisfaisaient à un ensemble de critères d'inclusion préétabli. Un formulaire de critères de qualité nouvellement élaboré a été utilisé pour évaluer la rigueur méthodologique des études comparatives directes. ***Résultats :*** Neuf articles satisfaisaient aux critères d'inclusion. Bien que pour deux études, on ait constaté une différence statistique appréciable favorable au RQM, il n'y avait aucun avantage apparent commun pour une mesure plutôt que pour l'autre. Les lacunes méthodologiques fréquentes étaient notamment l'absence de calcul formel de la taille de l'échantillon, l'absence de critère pour la comparaison des mesures et le fait qu'il n'y avait aucune norme de référence indépendante. ***Conclusion :*** Il n'y a aucun élément probant commun permettant de privilégier une mesure plutôt qu'une autre. Plusieurs études comportaient des lacunes sur le plan méthodologique.

The past three decades have seen a growing number of patient-reported outcome measures for persons with low back pain (LBP).[1–4] A challenge for clinicians and researchers is to select the measure with the best sensitivity to change from a pool of competing measures. Sensitivity to change is "the ability of an instrument to measure change in a the state regardless of whether it is relevant or meaningful to the decision maker."[5(p.85)] We use the term competing measures to denote measures intended for the same purpose. For clinicians, using a measure with greater sensitivity to change allows them to detect with confidence smaller changes in patients over the

From the *School of Rehabilitation Science and †Department of Clinical Epidemiology and Biostatistics, McMaster University; ‡Hamilton General Hospital; §Peak Performance Physiotherapy, Hamilton, Ont.

Correspondence to: Anastasia Newman, 43–81 Valridge Dr., Ancaster, ON L9G 5B6; newmanan@gmail.com.

Contributors: All authors designed the study, collected the data, and analyzed and interpreted the data; drafted or critically revised the article; and approved the final draft.

Competing interests: None declared.

Acknowledgements: The authors sincerely thank Deborah Kennedy for her thoughtful insight and feedback on a previous draft of this manuscript.

Anastasia Newman was a student in the MSc Rehabilitation Science Program at McMaster University at the time this study took place, and it was conducted in partial fulfilment of her degree requirements.

Physiotherapy Canada **2013; 65(2);160–166; doi:10.3138/ptc.2012-12**

course of treatment.[5] For researchers, using a measure with greater sensitivity to change translates into requiring smaller sample sizes for clinical intervention studies. Of the many competing measures developed to assess outcome for persons with LBP, the Roland-Morris Questionnaire (RMQ)[2,6] and Oswestry Disability Index (ODI)[1] are cited most frequently. Moreover, an expert group recommended the use of these measures when assessing patient outcomes.[7] Given the advantages to both clinicians and researchers of using the measure with the greatest sensitivity to change, our goal was to determine whether a head-to-head comparison supports the sensitivity to change of one measure over the other.

The purpose of this systematic review was to determine if the literature supported a difference in the sensitivity to change of RMQ and ODI scores when both measures were applied to the same patients with LBP. A secondary purpose was to examine the methodological rigour associated with the conduct and reporting of head-to-head comparison studies of these outcome measures. We use the term head-to-head comparison study to denote studies where competing measures such as the RMQ and ODI are evaluated on the same patients.

METHODS

Search strategy

We conducted a systematic literature search in the following databases: (1) MEDLINE, PubMed, and Embase: 1980 to July 17, 2011; (2) AMED: 1985 to July 17, 2011; (3) CINAHL: 1994 to July 17, 2011. Search terms were *Roland*, *Oswestry*, *sensitivity to change*, *sensitivity*, *receiver operating characteristic (ROC) curve*, *reliability*, *standardized response mean (SRM)*, *effect size (ES)*, *longitudinal*, and *correlation*. Consistent with the syntax of the various databases, we applied the Boolean term AND between Roland, Oswestry, and the collection of remaining terms, which were separated by OR. We also reviewed the reference lists of selected articles and contacted authors of relevant papers by e-mail. We provided these authors with our list of relevant studies and asked if they were aware of additional studies. We also requested the raw data from their studies. Independent searches were conducted by two investigators (AN and PS).

Study selection

Articles were included if (1) the article or abstract was published in English, (2) participants were aged 18 years or older with either acute, subacute, or chronic LBP with or without surgical intervention, (3) studies used the 24-item RMQ and versions 1 or 2 of the ODI with or without cross-cultural adaptations of the measures, (4) both measures were applied to the same patients at two common time points, and (5) the results allowed a comparison of the measures' abilities to assess change. Studies were excluded if participants had LBP attributed to malignancy, spinal fracture, infection, inflammatory disease, or unstable neurological conditions. Two investigators (AN and PS) independently assessed study eligibility. In the event of a disagreement, consensus was obtained through discussion or, if necessary, an adjudicator.

Data extraction

We were unable to find assessment tools or quality criteria specific to head-to-head comparison studies of outcome measures. Accordingly, guided by the work of the COnsensus-based Standards for the selection of health status Measurement INstruments (COSMIN) group, we developed criteria specific to our purpose.[8] Our criteria considered the following topics (see Appendix): (1) purpose: one item; (2) sample characteristics: six items; (3) study design: nine items; (3) measure description: two items; (4) sample size: two items; (5) analysis: three items; (6) results: seven items; and (7) conclusion: one item. Two investigators (AN and PS) independently applied these criteria to studies fulfilling the eligibility criteria. Disagreements were resolved as described in the study selection section.

Data analysis

Because investigators often reported multiple change coefficients that at times were based on conflicting assumptions concerning the change characteristics of the sample, we were guided by the work of Stratford and Riddle in determining the most appropriate coefficients for a given study.[9] If the investigators did not declare the expected change characteristic of their sample and the reference standard had three or more response options of hierarchical structure (e.g., global rating of change [GRC]), we considered the most appropriate analysis to be a correlation between the GRC and the RMQ and ODI's change scores.

When investigators did not perform a formal head-to-head analysis of the difference in change coefficients for the RMQ and ODI, we attempted to conduct a comparison based on information published in the manuscript or additional information requested from authors. We calculated Spearman's rank order correlation coefficient between the GRC and RMQ and ODI change scores, and applied Meng's test for correlated data to evaluate the difference in coefficients.[10] This test requires knowledge of the correlation between the RMQ and ODI's change scores. When this information was not available, we estimated the correlation between these measures by pooling all available data provided by investigators responding to our request for raw data.

Because many studies reported receiver operating characteristic (ROC) curve analysis, we also compared the area under the curves (AUC). When investigators did not formally compare the AUC between measures and their data were available to us, we applied Delong's test for correlated data.[11]

All tests were two-tailed, and a difference was considered statistically significant if $p < 0.05$. Stata version

Table 1 Characteristics of Patients by Study

Study	No. of patients analyzed*	Mean age (SD)	Sex M/F	Mean duration of symptoms (SD)	Pain location Back/Thigh/Leg
Beurskens et al.[12]	76	41 (10)	42/34	70 (119) wk	NA/NA/28
Coelho et al.[13]	30	38 (14)	20/10	3.4 (2.5) y	NA
Davidson and Keating[14]	99	52 (17)	31/68	Median ≈45 d	28/40/31
Frost et al.[15]	201	42 (14)	90/111	>6 wk	NA
Grotle et al.[16]	51	38 (10)	13/38	<3 wk	31/NA/NA
	48	40 (9)	18/30	>3 mo	6/NA/NA
Kopec et al.[3]	178	NA	NA	NA	NA
Mannion et al.[17]‡	57	53 (15)	26/31	NA	NA
Maughan et al.[18]	48	52	16/32	>6 mo	NA
Stratford et al.[19]	74	41 (12)	42/32	48 (36) d	NA

*Because of missing data, in some instances the number analyzed was smaller than the reported sample size. For this reason we report the number analyzed.
‡Spinal surgery.
NA = Not available.

Table 2 Baseline Roland-Morris and Oswestry Scores

Author (no. analyzed)		RMQ	ODI
Beurskens et al.[12]	(not improved: 38)*	11.8 (5.1)	29.1 (15.2)
	(improved: 38)	12.1 (4.7)	26.2 (13.5)
Coelho et al.[13]	(all: 30)	11.1 (5.7)	32.8 (18.9)
Davidson and Keating[14]	(not improved: 47)	9.0 (5.2)	35.0 (15.0)
	(improved: 52)	9.5 (5.9)	35.0 (17.0)
Frost et al.[15]	(worse: 16)	7.3 (4.5)	28.0 (12.5)
	(same: 76)	6.4 (4.3)	22.6 (11.6)
	(better: 109)	4.9 (3.8)	18.7 (9.4)
Kopec et al.[3]		NA	NA
Grotle et al.[16]	(acute not improved: 16)	7.2 (4.6)	26.0 (16.9)
	(acute improved: 35)	9.9 (5.1)	29.8 (14.8)
	(chronic not improved: 28)	8.9 (4.5)	32.6 (11.7)
	(chronic improved: 20)	10.1 (3.9)	29.7 (9.9)
Mannion et al.[17]	(not improved: 19)	16.7 (3.1)	53.4 (12.2)
	(improved: 38)	14.1 (4.7)	40.6 (15.4)
Maughan et al.[18]	(not improved: 25)	14.0 (5.4)	35.0 (20.2)
	(improved: 23)	9.0 (6.1)	24.0 (18.2)
Stratford et al.[19]	(not improved: 31)	11.6 (6.9)	39.8 (20.4)
	(improved: 43)	11.6 (5.8)	40.3 (16.5)

*Indicates the change status of patients.
RMQ = Roland-Morris Questionnaire; ODI = Oswestry Disability Index;
NA = Not available.

10.1 (STATACorp LP, College Station, TX) was used for all data analyses.

No ethics approval was necessary for this systematic review.

RESULTS

Search results

The literature search yielded the following number of citations: PubMed 107, Embase 128, Ovid (Medline) 98, AMED 38, and CINAHL 55. After applying the eligibility criteria both reviewers identified the same nine relevant articles. PubMed, Embase, and Ovid (Medline) all included the nine articles meeting the eligibility criteria, while AMED and CINAHL included some but not all of the nine articles.

The articles were authored by Beurskens and colleagues,[12] Coelho and colleagues,[13] Davidson and Keating,[14] Frost and colleagues,[15] Grotle and colleagues,[16] Kopec and colleagues,[3] Mannion and colleagues,[17] Maughan and Lewis,[18] Stratford and colleagues.[19]

Of the nine authors contacted, four (Grotle,[16] Davidson,[14] Mannion,[17] Beurskens[12]), in addition to Stratford, a coauthor of this study, responded to our invitation to provide additional potentially relevant articles. These authors did not identify additional relevant articles.

Patient characteristics

Table 1 displays a brief summary of the patients' characteristics for the nine studies. Table 2 provides the baseline RMQ and ODI mean values by subsequent improvement status if available.

Head-to-head comparison findings

Data sets were provided for the following five studies: Beurskens,[12] Davidson,[14] Grotle,[16] Mannion,[17] and Stratford.[19]

All nine studies applied a retrospective reference standard for global ratings of change. These reference standards consisted of 3 to 15 points of discrimination. Because a hierarchy of response options was available, we interpreted the application of these reference standards to be consistent with the view that many patients were expected to truly change by different amounts. Accordingly, a correlation coefficient would be the most appropriate sensitivity to change coefficient given this anticipated change characteristic.[9] Table 3 provides a summary of Spearman's correlation coefficients between the GRC and the RMQ and ODI measures' change scores. Also, shown in this table are the test statistics (standard normal deviate "Z") and *p*-values of the formal hypotheses

Table 3 Correlation Coefficient Comparison

Author (no. analyzed)		RMQ	ODI	Difference Z, *p*-value
Beurskens et al.[12]	(76)	0.72	0.46	3.28, 0.001
Coelho et al.[13]	(30)	NA	NA	NA
Davidson and Keating[14]	(99)	0.49	0.51	0.27, 0.78
Frost et al.[15]	(201)	0.38	0.47	1.90, 0.06*
Grotle et al.[16]	(all data: 99)	0.75	0.68	1.64, 0.10
	(acute: 51)	0.74	0.67	1.04, 0.30
	(chronic: 48)	0.61	0.49	1.56, 0.12
Kopec et al.[3]	(178)	0.47	0.35	2.36, 0.018*
Mannion et al.[17]	(57)	0.67	0.69	0.49, 0.62
Maughan et al.[18]	(48)	NA	NA	NA
Stratford et al.[19]	(74)	0.56	0.53	0.48, 0.63

*Based on an estimated correlation of 0.72 between RMQ and ODI change scores.
RMQ = Roland-Morris Questionnaire; ODI = Oswestry Disability Index; NA = Not available.

Table 4 Receiver Operating Curve Area Comparison

Author (no. analyzed)		RMQ	ODI	Difference χ^2_1, *p*-value
Beurskens et al.[12]	(76)	0.93	0.76	8.58, 0.003
Coelho et al.[13]	(30)	0.82	0.73	NA
Davidson and Keating[14]	(101)	0.76	0.77	0.04, 0.85
Frost et al.[15]	(201)	0.69	0.75	NA
Grotle et al.[16]	(GRC* all data: 99)	0.89	0.84	1.95, 0.16
	(GRC acute: 51)	0.93	0.87	0.66, 0.42
	(GRC chronic: 48)	0.83	0.75	0.23, 0.23
	(acute chronic: 99)	0.73	0.71	0.29, 0.59
Kopec et al.[3]		NA	NA	NA
Mannion et al.[17]	(57)	0.84	0.85	0.11, 0.75
Maughan et al.[18]	(48)	0.64	0.67	NA
Stratford et al.[19]	(74)	0.85	0.82	0.58, 0.45

*Global rating of change.
RMQ = Roland-Morris Questionnaire; ODI = Oswestry Disability Index; NA = Not available.

tests that the correlation coefficients for the two measures within a study are equal. The test statistics for Frost's and Kopec's studies were estimated using the pooled between-measure correlation of 0.72 obtained from the five studies providing raw data. The studies of Beurskens and colleagues ($Z = 3.28$, $p = 0.001$) and Kopec and colleagues ($Z = 2.36$, $p = 0.018$) showed statistically significant differences in favour of the RMQ.

Although we believe the correlation analysis is the most appropriate (i.e., the applied rating scales allowed for multiple levels of change),[9] all investigators with the exception of Kopec and colleagues[3] reported the results of ROC curve analyses. For this reason we also present a summary of these results and our formal head-to-head analyses in Table 4. Only the study by Beurskens and colleagues yielded a statistically significant difference, which was in favour of the RMQ.

Methodological findings

Table 5 summarizes our methodological review of the nine studies. We found that authors consistently (i.e., at least 8 of 9 Yes responses) provided a clear statement of the purpose; description of the sample; setting and study design including the interval between assessments, description of the reference standard, sample size, baseline and change RMQ and ODI summary values; and individual measure inferential statistics. In contrast, authors consistently (i.e., at least 7 of 9 No or Can't tell responses) provided inadequate descriptions of the measurement conditions, did not justify the interval between assessments, did not state the sample's change characteristic, did not apply a reference standard that was independent of the patients' measures responses, did not perform a sample size or power calculation, neglected to comment on the patients lost to follow-up, and failed to attempt a formal statistical comparison of the measures' abilities to detect change (Table 5).

DISCUSSION

Our systematic review found no substantive evidence supporting a difference in the sensitivity to change of the RMQ and ODI. However, our review did reveal consistent deficiencies in the conduct and reporting of head-to-head comparison studies of these measures.

Several factors can influence the quality of a systematic review. These include the extent to which relevant studies have been identified and the accuracy of the interpretation of their results. We systematically searched five databases and petitioned authors of relevant studies to supplement our reference list. Four authors acknowledged our request; however, none provided additional studies fulfilling our eligibility criteria. Accordingly, we believe that we included all relevant studies appearing in the literature at the time of our investigation.

Subsequent to the completion of our study, an additional article by Monticone and colleagues comparing the sensitivity to change of the Italian versions of the ODI and the RMQ was published.[20] Their sample consisted of 179 patients with subacute and chronic LBP. These patients were recruited from four rehabilitation units and completed both outcome measures before and after an eight-week rehabilitation program. The reference standard was a fivepoint global perception of change. The authors reported correlation coefficients between the global perception of change and the RMQ and ODI to be 0.287 and 0.431 respectively. The authors did not perform a formal comparison of these coefficients; neither did they provide the correlation between RMQ and ODI change scores. Applying the correlation between RMQ and ODI change scores of 0.72 obtained from our

Table 5 Methodological Summary*

Criteria	Beurskens et al.[12]	Coelho et al.[13]	Davidson and Keating[14]	Frost et al.[15]	Grotle et al. (a)[16]	Grotle et al. (b)[16]	Kopec et al.[3]	Mannion et al.[17]	Maughan et al.[18]	Stratford et al.[19]
Purpose										
Was the purpose/research question clearly stated?	Y†	Y	Y	Y	Y	Y	Y	Y	Y	Y
Sample characteristics										
Were eligibility criteria clearly stated?	N	Y	Y	Y	Y	Y	Y	N	Y	Y
Did the authors provide descriptive statistics of age?	Y	Y	Y	Y	Y	Y	N	Y	Y	Y
Was a gender distribution provided?	Y	Y	Y	Y	Y	Y	N	Y	Y	Y
Were descriptive statistics provided concerning the duration of back pain before the study?	Y	Y	Y	N	Y	Y	N	N	Y	Y
Were the patients' working status provided?	N	N	Y	N	Y	Y	N	N	Y	Y
Was the distribution of pain pattern provided?	Y	N	Y	N	Y	Y	N	N	N	N
Study design										
Was the study design explicitly stated?	Y	N	Y	Y	Y	Y	N	N	Y	Y
Was the setting of the study stated?	N	Y	Y	Y	Y	Y	Y	Y	Y	Y
Were the measurement conditions similar for both measures?	CT	CT	CT	CT	CT	CT	CT	CT	CT	Y
Was the interval between assessments specified?	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Was the interval between assessments justified?	N	N	Y	N	N	N	N	N	N	N
Was the expectation of the sample's change characteristics stated?	N	N	N	N	N	N	N	N	N	N
Interval between assessments	5 wk	6 wk	6 wk	12 mo	4 wk 3 mo	4 wk 3 mo	4–6 mo	6 mo	5 wk	4–6 wk
Reference standard for change	7-pt. GRC	7-pt. GRC	7-pt. GRC	3-pt. GRC	Expected clinical course	6-pt. GRC	15-pt. GRC	6-pt. GRC	7-pt. GRC	15-pt. GRC
Was the reference standard independent of measures' responses?	N	N	N	N	Y	N	N	N	N	N
Measure description										
Questionnaire Language††	D	BP	E	E	NW	NW	E/F	G	E	E
Version used	RMQ (Y/N) ODI 1.0	RMQ ODI 1.0	RMQ ODI 2.0	RMQ (Y/N) ODI 2.1	RMQ ODI 2.0	RMQ ODI 2.0	RMQ ODI 1.0	RMQ (Y/N) ODI 2.1	RMQ (Y/N) ODI 2.0	RMQ ODI 1.0
Sample Size										
Was a formal sample size calculation done?	N	N	N	N	N	N	N	N	N	N
Sample size: enrolled/analyzed§	81/76	30/30	207/99	286/201	104/104	104/104	242/178	63/57	63/48	88/74
Analysis										
Was the choice of analysis consistent with the samples expected change characteristics?	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT
Was a formal comparison of the measures attempted?	N	N	Y	N	N	N	N	N	N	Y
Did the formal analysis account for dependent data?	Not attempted	Not attempted	Y	Not attempted	Not attempted	Not attempted	Not attempted	Not attempted	Not attempted	CT
Results										
Was the proportion of unanswered/multiple response questions reported?	N	N	N	N	Y	Y	Y	N	N	Y
Did the authors comment on patient follow-up losses?	Y	N	Y	N	N	N	N	N	N	N
Were descriptive statistics of each measure provided for pre-scores?	Y	Y	Y	Y	Y	Y	N	Y	Y	Y
Were descriptive statistics of each measure provided for post-scores?	Y	Y	Y	Y	N	N	N	Y	Y	Y
Were descriptive statistics of each measure provided for change scores?	Y	Y	Y	Y	Y	Y	N	Y	Y	Y
Were individual measure change statistics with p-value/CIs provided?	N	Y	Y	Y	Y	Y	N	Y	N	Y
Were between measure comparison statistics with p-value/CIs provided?	N	N	Y	N	Y	Y	N	N	Y	Y
Conclusion										
Were the authors' conclusions consistent with the results?	CT	Y	Y	CT	Y	Y	Y	Y	Y	Y

*Ratings in this table are specific to the number of patients analyzed.

†Sample size enrolled and analyzed for this RMQ ODI head-to-head comparison.

E = English; NW = Norwegian; D = Dutch; F = French; BP = Brazilian-Portuguese; G = German; Grotle (a) = Expected Clinical Course as reference standard; Grotle (b) = Global Change Index as reference standard; GRC = Global rating of change; RMQ = RMQ-Original; RMQ (Y/N) = RMQ Yes/No response option; ODI version; Y = Yes, N = No, CT = Can't tell; NA = None available.

pooled data yielded $Z = 1.69$, $p = 0.09$. Monticone and colleagues also performed an ROC curve analysis that produces areas of 0.64 (95% CI, 0.55–0.72) for the RMQ and 0.71 (95% CI, 0.64–0.79) for the ODI. The authors did not perform a formal between-measure comparison.

The identified studies showed substantial differences in patient characteristics, duration of symptoms, language of instruments, and version of ODI. For these reasons it was not possible to quantitatively pool the results across all nine studies. We found that when a correlation analysis was applied, the studies of Beurskens and colleagues[12] and Kopec and colleagues[3] yielded statistically significant differences in favour of the RMQ. When ROC curve analysis was performed, only the study of Beurskens and colleagues demonstrated a statistically significant difference.[12] This difference also favoured the RMQ. In spite of these findings in favour of the RMQ, just under half of the studies produced point estimates of sensitivity to change in favour of the ODI. Our interpretation of these results is that there is no clear evidence supporting a difference in the sensitivity to change of the RMQ and ODI.

We chose to include only head-to-head comparison studies in our investigation; however, independent estimates of the sensitivity to change of the RMQ and ODI can be found in many more investigations.[21–26] Our reasoning for including only head-to-head comparison studies was based on the observation of Messick who noted that properties such as reliability and validity are not properties of a measure, but rather of a measure in a particular context.[27] Accordingly, we felt that a meaningful comparison between measures could be obtained only when they were evaluated on the same patients, in the same setting, and at the same time points.

Application of our methodological criteria revealed that investigators have consistently stated the study's purpose, provided a description of the sample's demographic and baseline characteristics, specified the interval between assessments, and provided descriptive statistics of the change. In contrast, investigators have rarely if ever justified the interval between assessments, stated the expected change characteristic of the sample, justified the sample size, applied a reference standard that is independent of the measures under investigation, commented on the losses to follow-up, and performed a formal statistical comparison of the between-measure difference in sensitivity to change coefficients.

Our study has several limitations. First, because we searched for and included only studies or abstracts published in English, we do not know the extent to which studies that would have otherwise met our eligibility exist in other languages. A second limitation is that for some studies we had insufficient data to perform a formal statistical comparison of the change coefficients.

CONCLUSION

Our findings do not provide strong evidence supporting a difference in the sensitivity to change of the RMQ and ODI. However, our results do suggest that our quality criteria form demonstrated that head-to-head comparison studies of the RMQ and ODI consistently had methodological deficiencies in the following areas: justifying the interval between assessments, stating the expected change characteristic of the sample, justifying the sample size, applying a reference standard that is independent of the measures under investigation, commenting on the losses to follow-up, and performing a formal statistical comparison of the between measure difference in sensitivity to change coefficients.

KEY MESSAGES

What is already known on this topic

The RMQ and ODI are the two most frequently recommended and cited patient-reported outcome measures for persons with LBP. Individual studies examining the sensitivity to change of these measures have produced conflicting results.

What this study adds

Our systematic review found no clear evidence supporting a difference in the sensitivity to change of the RMQ and ODI. To our knowledge this investigation is the first to comment on and suggest quality criteria for the evaluation of head-to-head comparison studies of competing measures' abilities to detect change.

REFERENCES

1. Fairbank JC, Couper J, Davies JB, et al. The Oswestry low back pain disability questionnaire. Physiotherapy. 1980;66(8):271–3. Medline:6450426
2. Roland M, Morris R. A study of the natural history of back pain. Part I: development of a reliable and sensitive measure of disability in low-back pain. Spine. 1983;8(2):141–4. http://dx.doi.org/10.1097/00007632-198303000-00004. Medline:6222486
3. Kopec JA, Esdaile JM, Abrahamowicz M, et al. The Quebec Back Pain Disability Scale. Spine. 1995;20(3):341–52. http://dx.doi.org/10.1097/00007632-199502000-00016. Medline:7732471
4. Stratford PW, Binkley JM, Riddle DL. Development and initial validation of the back pain functional scale. Spine. 2000;25(16):2095–102. http://dx.doi.org/10.1097/00007632-200008150-00015. Medline:10954642
5. Liang MH. Longitudinal construct validity: establishment of clinical meaning in patient evaluative instruments. Med Care. 2000;38(9 Suppl):II84–90. Medline:10982093
6. Roland M, Morris R. A study of the natural history of low-back pain. Part II: development of guidelines for trials of treatment in primary care. Spine. 1983;8(2):145–50. http://dx.doi.org/10.1097/00007632-198303000-00005. Medline:6222487
7. Deyo RA, Battie M, Beurskens AJ, et al. Outcome measures for low back pain research. A proposal for standardized use. Spine. 1998;23(18):2003–13. http://dx.doi.org/10.1097/00007632-199809150-00018. Medline:9779535
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(1):22. http://dx.doi.org/10.1186/1471-2288-10-22. Medline:20298572
9. Stratford PW, Riddle DL. Assessing sensitivity to change: choosing the appropriate change coefficient. Health Qual Life Outcomes. 2005;3(1):23. http://dx.doi.org/10.1186/1477-7525-3-23. Medline:15811176

10. Meng X, Rosenthal R, Rubin DB. Comparing correlated correlation coefficients. Psychol Bull. 1992;111(1):172–5. http://dx.doi.org/10.1037/0033-2909.111.1.172.
11. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837–45. http://dx.doi.org/10.2307/2531595. Medline:3203132
12. Beurskens AJHM, de Vet HCW, Köke AJA. Responsiveness of functional status in low back pain: a comparison of different instruments. Pain. 1996;65(1):71–6. http://dx.doi.org/10.1016/0304-3959(95)00149-2. Medline:8826492
13. Coelho RA, Siqueira FB, Ferreira PH, et al. Responsiveness of the Brazilian-Portuguese version of the Oswestry Disability Index in subjects with low back pain. Eur Spine J. 2008;17(8):1101–6. http://dx.doi.org/10.1007/s00586-008-0690-1. Medline:18512083
14. Davidson M, Keating JL. A comparison of five low back disability questionnaires: reliability and responsiveness. Phys Ther. 2002;82(1):8–24. Medline:11784274
15. Frost H, Lamb SE, Stewart-Brown S. Responsiveness of a patient specific outcome measure compared with the Oswestry Disability Index v2.1 and Roland and Morris Disability Questionnaire for patients with subacute and chronic low back pain. Spine. 2008;33(22):2450–7, discussion 2458. http://dx.doi.org/10.1097/BRS.0b013e31818916fd. Medline:18824951
16. Grotle M, Brox JI, Vøllestad NK. Concurrent comparison of responsiveness in pain and functional status measurements used for patients with low back pain. Spine. 2004;29(21):E492–501. http://dx.doi.org/10.1097/01.brs.0000143664.02702.0b. Medline:15507789
17. Mannion AF, Junge A, Grob D, et al. Development of a German version of the Oswestry Disability Index. Part 2: sensitivity to change after spinal surgery. Eur Spine J. 2006;15(1):66–73. http://dx.doi.org/10.1007/s00586-004-0816-z. Medline:15856340
18. Maughan EF, Lewis JS. Outcome measures in chronic low back pain. Eur Spine J. 2010;19(9):1484–94. http://dx.doi.org/10.1007/s00586-010-1353-6. Medline:20397032
19. Stratford PW, Binkley J, Solomon P, et al. Assessing change over time in patients with low back pain. Phys Ther. 1994;74(6):528–33. Medline:8197239
20. Monticone M, Baiardi P, Vanti C, et al. Responsiveness of the Oswestry Disability Index and the Roland Morris Disability Questionnaire in Italian subjects with sub-acute and chronic low back pain. Eur Spine J. 2012;21(1):122–9. http://dx.doi.org/10.1007/s00586-011-1959-3. Medline:21823035
21. Demoulin C, Ostelo R, Knottnerus JA, et al. What factors influence the measurement properties of the Roland-Morris disability questionnaire? Eur J Pain. 2010;14(2):200–6. http://dx.doi.org/10.1016/j.ejpain.2009.04.007. Medline:19443246
22. 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(1):45–52. http://dx.doi.org/10.1016/j.jclinepi.2005.03.018. Medline:16360560
23. Hall AM, Maher CG, Latimer J, et al. The patient-specific functional scale is more responsive than the Roland Morris disability questionnaire when activity limitation is low. Eur Spine J. 2011;20(1):79–86. http://dx.doi.org/10.1007/s00586-010-1521-8. Medline:20628767
24. Küçükdeveci AA, Tennant A, Elhan AH, et al. Validation of the Turkish version of the Roland-Morris Disability Questionnaire for use in low back pain. Spine. 2001;26(24):2738–43. http://dx.doi.org/10.1097/00007632-200112150-00024. Medline:11740366
25. Taylor SJ, Taylor AE, Foy MA, et al. Responsiveness of common outcome measures for patients with low back pain. Spine. 1999;24(17):1805–12. http://dx.doi.org/10.1097/00007632-199909010-00010. Medline:10488511
26. Fritz JM, Irrgang JJ. A comparison of a modified Oswestry Low Back Pain Disability Questionnaire and the Quebec Back Pain Disability Scale. Phys Ther. 2001;81(2):776–88. Medline:11175676
27. Messick S. Validity. In: Linn RL, editor. Educational measurement. 3rd ed. Phoenix: ORYZ Press; 1993. p. 14.