Different minimally important clinical difference (MCID) scores lead to different clinical prediction rules for the Oswestry disability index for the same sample of patients

Julie Schwind, Kenneth Learman, Bryan O'Halloran, Christopher Showalter, Chad Cook

Walsh University, North Canton, OH, USA

Background: Minimal clinically important difference (MCID) scores for outcome measures are frequently used evidence-based guides to gage meaningful changes. There are numerous outcome instruments used for analyzing pain, disability, and dysfunction of the low back; perhaps the most common of these is the Oswestry disability index (ODI). A single agreed-upon MCID score for the ODI has yet to be established. What is also unknown is whether selected baseline variables will be universal predictors regardless of the MCID used for a particular outcome measure.
Objective: To explore the relationship between predictive models and the MCID cutpoint on the ODI.
Setting: Data were collected from 16 outpatient physical therapy clinics in 10 states.
Design: Secondary database analysis using backward stepwise deletion logistic regression of data from a randomized controlled trial (RCT) to create prognostic clinical prediction rules (CPR).
Participants and Interventions: One hundred and forty-nine patients with low back pain (LBP) were enrolled in the RCT. All were treated with manual therapy, with a majority also receiving spine-strengthening exercises.
Results: The resultant predictive models were dependent upon the MCID used and baseline sample characteristics. All CPR were statistically significant ($P < 0.01$). All six MCID cutpoints used resulted in completely different significant predictor variables with no predictor significant across all models.
Limitations: The primary limitations include sub-optimal sample size and study design.
Conclusions: There is extreme variability among predictive models created using different MCIDs on the ODI within the same patient population. Our findings highlight the instability of predictive modeling, as these models are significantly affected by population baseline characteristics along with the MCID used. Clinicians must be aware of the fragility of CPR prior to applying each in clinical practice.

Keywords: Low back pain, Clinical prediction rule, Minimal clinically important difference, Prognosis

Introduction

In today's healthcare environment an increased focus on providing cost-efficient care without compromising patient outcome has emphasized the importance of sound clinical decision making. A key element in clinical decision making is the ability to gage when meaningful changes occur in the patient's condition. Minimal clinically important difference (MCID) scores for outcome measures are frequently used evidence-based guides to gage meaningful changes.[1]

Correspondence to: Chad Cook, Walsh University, North Canton, OH, USA. Email: jmmtcook@gmail.com

A MCID score is defined as the minimal change in score on an outcome instrument that coincides with the patient's perception of beneficial change or recovery.[2] The MCID score is a single 'cut point', that is, a point estimate that either represents a change in the score (initial score minus the final score or a percentage-based change score from baseline) or a particular value for the final score.

As stated, MCID scores are calculated for a number of different outcome instruments. There are numerous outcome instruments used for analyzing pain, disability, and dysfunction of the low back; perhaps the most common of these is the Oswestry disability index (ODI). The ODI is a self-administered questionnaire

DOI 10.1179/2042618613Y.0000000028

containing 10 sections including pain intensity, personal care, lifting, walking, sitting, standing, sleeping, sex life, social life, and traveling. Each section contains six statements that are scored from 0 to 5, with 0 representing no difficulty in the activity and 5 representing maximal difficulty. The scores from each section are totaled and divided by the total possible score to obtain a final percentage of disability, with a higher per cent indicating greater disability.[3] The ODI is a valid, reliable, and responsive clinical tool for analyzing disability status in individuals with low back pain (LBP).[2,4,5]

A single agreed-upon MCID score has yet to be established for the ODI.[1] There have been a number of suggestions for MCIDs, which have mostly been represented by required change scores that have been anchored to a patient's perception of clinical importance. Minimal clinically important difference change score cutpoints for the ODI that have been advocated include: 50% change,[6] 30% change,[7,8] 17-point change,[9] 10-point change,[10,11] and 5- (and sometimes 6-) point change.[12–14] The creators[3] of the ODI suggested that a final ODI score of ≤20% represented no disability. To our knowledge, the assumption that an ODI score of ≤20% represents no disability has not been supported beyond the original publication.[3]

For obvious reasons, this lack of clarity for the proper MCID may cause confusion among clinicians. In an attempt to explain the variations, authors have suggested a multitude of reasons for MCID variances, including: (i) the lack of a standardized methodological calculation approach;[1] (ii) the patient recall bias when using an anchor-based calculation approach;[1,15] (iii) the dependence on the sample size and lack of patient perspective when using distribution-based calculations;[1,16,17] and (iv) the influence of patient demographics and baseline characteristics.[1,18,19] Worth noting is that depending on interpretation by a clinician one may find very different results when scrutinizing the effectiveness of a particular intervention. For example, adoption of a MCID may result in fewer 'successes' with a 50% change from baseline than adoption of a 30% change score.

Baseline patient characteristics used during predictive modeling have been found to determine good or poor prognosis in patients with LBP regardless of the treatment provided.[20] Prognostic clinical prediction rules (CPR) are created by identifying baseline predictive characteristics of patients who are inclined to improve irrespective of the treatment provided. Recently, it was found that meeting the CPR for spinal manipulation,[21,22] which includes the presence of four out of five predictive variables at baseline (no pain below knee, symptoms less than 16 days, fear avoidance beliefs questionnaire work subscale [FABQ-w]<19, 1+ hips with an internal rotation range of motion of >35°, 1+hypomobile lumbar segment) is a universal predictor for good prognosis in patients with LBP regardless of the outcome tool used.[23] What is unknown is whether the selected baseline variables will be universal predictors regardless of the MCID used for a particular outcome measure. Therefore, the purpose of this study was to identify predictive variables for recovery in patients with LBP based on variable MCID cutpoints for the ODI. We hypothesized that the predictive model will be the same regardless of the MCID cutpoint used, since each cutpoint (50% change, 30% change, 17-point change, 10-point change, five-point change, and a final ODI score of ≤20%) has been found previously within the literature (with the exception of the original authors' recommendation of ≤20%) to indicate meaningful recovery.

Method

Design

The study was a secondary database analysis using predictive modeling of a randomized controlled trial (RCT) comparing thrust and non-thrust manipulation in the treatment of LBP.[24] The RCT was registered with ClinicalTrials.gov: Identifier NCT01438203. Specific details of the trial are published elsewhere.[24] The study was approved by the Walsh University Human Ethics Review Board.

Participants

The RCT enrolled 149 patients with mechanically reproducible LBP who were aged 18 years or older. Participant exclusion criteria included the presence of a tumor, metabolic diseases, rheumatoid arthritis, osteoporosis, prolonged history of steroid use, past surgical history of the lumbar spine, current pregnancy, or signs consistent with nerve root compression (any of the following: reproduction of low back or leg pain with straight leg raise of <45°, muscle weakness involving a major muscle group of the lower extremity, diminished lower extremity muscle stretch reflex, or diminished or absent sensation to pinprick in any lower extremity dermatome).

Intervention

The intervention was a comprehensive rehabilitation intervention that included either thrust or non-thrust manipulation for the first two visits only, followed by physical therapist-directed treatment approach until discharge. In the RCT, there were no differences in any of the outcomes, including the ODI, between the thrust and non-thrust manipulation interventions.[24] All patients in the RCT were treated by one of the 17 highly trained physical therapists who had extensive manual therapy training including certification in orthopedic manual therapy or were Fellows within

the American Academy of Orthopedic Manual Physical Therapists.

Minimal clinically important difference scores used for the ODI during creation of the predictive models

The modified version of the ODI was used for this study.[14] Within the modified version,[14] the sex-related question was replaced with a question associated with social life. Six MCID cutpoints were used in constructing the predictive models. These included:

1. a 50% change in disability from the initial to final ODI calculated as [(ODI raw $score_{initial}$−-ODI raw $score_{final}$)/ODI raw $score_{initial} \times 100\%] \geq 50\%$;[6]
2. a 30% change calculated as 30% change [(ODI raw $score_{initial}$−ODI raw $score_{final}$)/ODI raw $score_{initial} \times 100\%] \geq 30\%$;[7,8]
3. a 17-point decrease calculated as 17-point change [ODI total $score_{initial}$−ODI total $score_{final}] \geq 17$;[9]
4. a 10-point decrease calculated as 10-point change [ODI total $score_{initial}$−ODI total $score_{final}] \geq 10$;[10,11]
5. a 5–6-point decrease calculated as 5–6-point change [ODI total $score_{initial}$−ODI total $score_{final}] \geq 5$;[12,13,14]
6. a final ODI score of 20% or less.

Prognostic variables used in the predictive models

Ten prognostic variables were selected based on prior representation within the published literature, or based on expectations derived from clinical experience. Irrespective of whether the subject met the CPR for spinal manipulation,[30] duration of symptoms (weeks), and age,[27] the variables, body mass index (BMI),[25,26] NPRS at baseline,[27] ODI at baseline,[28] fear avoidance beliefs questionnaire work subscale (FABQ-w) at baseline,[29] are well-represented within the literature, have been acknowledged as prognostic variables, and are the same variables used in a paper by Cook and colleagues.[23] The FABQ-w[31] is a subset of the full FABQ and includes a seven-item questionnaire examining the subjects' beliefs on the relationship between their work and pain. The CPR for manipulation[21,22] has been proposed to be both prescriptive and prognostic[32] and was coded as present or not-present during the initial baseline visit with an operational definition of meeting the rule set for the presence of at least four out of five variables.

The variables irritability[32,33] and medical diagnosis (using the International Classification of Diseases, Ninth Edition (ICD-9 code)) were also used by Cook and colleagues[23] and were found to be prognostic. Irritability was a concept promoted by Maitland[34] and includes three related constructs: (i) the vigor of the activity required to provoke a patient's symptoms; (ii) the severity of those symptoms; and (iii) the time it takes for the symptoms to subside once aggravated (i.e. pain persistence). The variable was dichotomously coded, as recommended by Maitland,[34] as present or not-present; with present qualified as any one or more excessive findings recognized on the three identifiers. We created a dichotomous variable titled 'diagnosis' by combining all strains and sprains (ICD-9 codes 847.2 and 846.0) into one group and combining the remaining codes (722.1; 724.2; 724.6, and others) into another category.

Data analysis

All analyses were performed using *Statistical Package for the Social Sciences* (SPSS) version 18.0 (233 South Wacker Drive, Chicago, Illinois, 60606, USA). Baseline characteristics including means, standard deviations, and frequencies were reported.

Logistic regression analysis

A backward elimination binary logistic regression analysis was completed for each of the six dependent variables (definitions of MCID used for the ODI) using stepwise deletion ($p = 0.05$ enter and 0.10 exit). Backward elimination involves starting with all predictive variables, testing the deletion of each variable using a chosen model comparison criterion, deleting the variable (if any) that improves the model the most by being deleted, and repeating this process until no further improvement is possible. All six dependent variables were dichotomized into successful and unsuccessful. Success for 50% improvement on the ODI was calculated as [(ODI raw $score_{initial}$ − ODI raw $score_{final}$)/ODI raw $score_{initial} \times 100\%] \geq 50\%$. Success for 30% improvement on the ODI was calculated as [(ODI raw $score_{initial}$ − ODI raw $score_{final}$)/ODI raw $score_{initial} \times 100\%] \geq 30\%$. Success for the 17, 10, and five-point improvement models on the ODI was calculated as [ODI total $score_{initial}$ − ODI total $score_{final}] \geq 17$, 10, or five, respectively. Success for a final ODI score of 20% or less was an ODI total $score_{final}$ of $\leq 20\%$. If a baseline score was already below the cutpoint for success, the N was adjusted to exclude those patients. A second logistic regression was completed to account for the differences in sample size between models using the same process, but only including patients with an initial ODI above 20%. For all regression calculations a P value of ≤ 0.05 was considered significant.

Goodness of fit and collinearity

A Nagelkerke R^2 was used to assess goodness of fit, as it theoretically represents the proportion of variance in the criterion that is explained by the predictors. The Nagelkerke R^2 is a version of the Cox and Snell R^2, which overcomes the problem that this statistic has of not being able to reach its maximum value. Although controversial, the Nagelkerke R^2 is often used to determine how well the predictors explain the model variance (or, proposed explanation of the proportion of variance).[35] The R^2 was run for each of the six models.

Collinearity is a measure of the correlation between (and thereby, redundancy of) the predictive variables in the model. Collinearity is usually measured in terms of variance inflation factors (VIFs) for their predictors.[36] Variance inflation factors were calculated for each covariate in all six models. Ideally, VIF values should be low with a mean VIF close to 1, indicative of minimal collinearity, cut-offs of 5, 10, and sometimes 30 have been suggested as indicating problematic levels of multicollinearity.[36]

Results

Table 1 summarizes the descriptive statistics of the sample population including frequencies, means, standard deviations, ranges, and percentages. The age range of the sample was diverse including individuals 18–88 years of age with a balanced mix of females and males (53% and 47%, respectively). The majority was Caucasian (91.3%), most did not display irritability at baseline (73.2%), and half (50.3%) demonstrated acute LBP. Body mass index (BMI) ranged from 18.7 to 46.7 with a mean of 26.4. The duration of symptoms ranged from 1 to 1000 weeks with an average of 33.9 weeks, the total visits ranged from 3 to 28, and the total days in care ranged from 3 to 150. The descriptive statistics were also reported for the subgroup of patients with a baseline ODI score above 20% to compare to the original sample. The subgroup of 107 patients had the same age range with the average age slightly increased to 49.7 years. The subgroup had a greater ratio of females (57%) to males (43%) and a higher percentage of irritable patients (31.8%). The characteristics of race, BMI, baseline LBP classification (acute = <6 weeks of symptoms, sub-acute = 6 weeks to 6 months of symptoms, and chronic = >6 months of symptoms), diagnosis, baseline numeric pain scale rating, and treatment group allocation were all very similar to the original sample of 149 patients. The mean symptom duration of the subgroup decreased to 26.4 weeks with a range of 1–312 and baseline ODI increased to an average of 37.7%.

Table 1 Descriptive statistics of the sample. The first two columns represent the full sample (*N* = 149) whereas the second two columns (*N* = 107) represent subjects with a baseline Oswestry disability index (ODI) of <20

	(*N* = 149)		(*N* = 107)		
	Full sample		Partial sample of ODI <20		
Variable	Mean (SD)/frequency	Ranges/percentages	Mean (SD)/frequency	Ranges/percentages	*P* value
Age (years)	48.2 (14.9)	18–88 years	49.7 (15.1)	18–88 years	0.43
Gender	70=male	47% = male	46 = male	43% = male	0.61
	79=female	53% = female	61 = female	57% = female	
Irritability	39=irritable	26.2% = irritable	34 = irritable	31.8% = irritable	0.42
	109=not irritable	73.2% = not irritable	73 = not irritable	68.2% = not irritable	
	1=missing	0.7% = missing	0 = missing	0% = missing	
Race/culture	136=white	91.3% = white	96 = white	89.7% = white	0.99
	3 = black	2% = black	2 = black	1.9% = black	
	3 = hispanic	2% = hispanic	3 = hispanic	2.8% = hispanic	
	3 = asian	2% = asian	3 = asian	2.8% = asian	
	2 = other	1.3% = other	2 = other	1.9% = other	
	2 = missing	1.3% = missing	1 = missing	0.9% = missing	
BMI	26.4 (4.8)	18.7–46.7	26.6 (5.0)	18.7–46.7	075
Baseline LBP classification	75 = acute	50.3% = acute	54 = acute	50.5% = acute	0.98
	43 = sub-acute	28.9% = sub-acute	30 = sub-acute	28.0% = sub-acute	
	31 = chronic	20.8% = chronic	23 = chronic	21.5% = chronic	
	0 = missing	0% = missing	0 = missing	0% = missing	
Symptom duration (weeks)	33.9 (98.9)	1–1000 weeks	26.4 (50.2)	1–312 weeks	0.47
Diagnosis	70 = sprains and strains	46.9% = sprains and strains	52 = sprains and strains	48.6% = sprains and strains	0.89
	72 = lumbago and degenerative conditions	48.4% = lumbago and degenerative conditions	50 = lumbago and degenerative conditions	46.7% = lumbago and degenerative conditions	
	7 = missing	4.7% = missing	5 = missing	4.7% = missing	
Total visits	6.9 (4.6)	3–28 visits	7.6 (4.5)	2–28 visits	0.23
Total days in care	35.7 (29.9)	3–150 days	36.6 (30.5)	1–160 days	0.81
NPRS at baseline	5.2/10 (2.1)	1/10–10/10	5.5/10 (1.9)	1/10–10/10	0.24
ODI at baseline	30.6 (15.7)	2–78	37.7 (12.5)	22–78	<0.01
FABQ-w at baseline	11.9 (10.7)	0–43	13.06 (11.0)	0–43	0.39
Met CPR	71 = met rule	49% = met rule	46 = met rule	43% = met rule	0.54
	78 = did not meet rule	51% = did not meet rule	61 = did not meet rule	57% = did not meet rule	
Treatment group	73 = non-thrust manipulation	49% = non-thrust manipulation	52 = non-thrust manipulation	48.6% = non-thrust manipulation	0.95
	76 = thrust manipulation	51% = thrust manipulation	55 = thrust manipulation	51.4% = thrust manipulation	

Table 2 displays the results for the logistic regression models, with all models deemed significant ($P < 0.01$). After the backward stepwise deletion, three variables (meeting the CPR, younger age, and diagnosis of lumbago and degenerative disease) were significantly associated with the 50% change in the ODI, four variables (lower baseline FABQ, shorter symptom duration, younger age, and diagnosis of lumbago and degenerative disease) were significantly associated with the 30% change in the ODI, two variables (higher baseline ODI and meeting the CPR) were significantly associated with the 17-point reduction in the ODI, two variables (higher baseline ODI and younger age) were significantly associated with both the 10-point and five-point change in the ODI, and three variables (lower baseline ODI, younger age, and meeting the CPR) were significantly associated with a final ODI less than or equal to 20%. There was no single variable that was significant across all six models. *Younger age* was significant in five out of six models, followed by a *higher baseline ODI score* (three of six), *meeting the CPR* (three of six), *diagnosis of lumbago and degenerative disease* (two of six), and *lower baseline FABQ*, *shorter symptom duration, and lower baseline ODI* (each one of six). The Nagelkerke R^2 for each model is also reported in Table 2, indicating the proposed explanation of the proportion of variance. The model with the highest Nagelkerke R^2 involved the dependent variable of a final ODI score of 20% or less, with a lower baseline ODI, younger age, shorter symptom duration, diagnosis, and meeting the CPR accounting for 49.1% of predictive factors. The explanation of proportion of variance of all other models was not even half of that of a final ODI score of 20% or less, ranging from 24.5% down to 13.9%.

After adjusting the sample to include only patients with a baseline ODI score greater than 20%, a second logistic regression produced six different, yet significant, models with the results presented in Table 3. Once again, there was no one predictive variable that was individually significant across all six models. *Diagnosis of lumbago and degenerative disease* was significant in four out of six models. *Younger age*, *shorter symptom duration, and higher baseline ODI* were each significant in three out of six models, however, not commonly significant in any model. *Meeting the CPR* followed in two of six and a new predictive variable of *lower BMI* was individually significant in the five-point change model only. A *lower baseline ODI* was found to be significant only in the final ODI of 20% or less model. A final ODI of 20% or less produced the model with the highest Nagelkerke R^2 of 49.1%, followed by the 50% change model at 27.1%, the 17-point change model at 26.0%, the 10-point change model at 21.7%, the five-point

Table 2 Logistic regression modeling including final predictor variables, individual *P* values for each model variable, as well as odds ratios and 95% confidence intervals

Model	Variables	Individual *P* value	Odds ratio (95% confidence interval)	Nagelkerke R^2	Model *P* value	% Correct
ODI 50% change	Met CPR	0.005	2.916 (1.380–6.163)	0.231	0.000	57.0%
(*N*=149)	Younger age	0.003	1.041 (1.014–1.069)			
	Diagnosis	0.014	0.385(0.180–0.824)			
ODI 30% change	Lower baseline FABQ	0.044	1.039 (1.001–1.079)	0.214	0.000	70.5%
(*N*=149)	Shorter symptom duration	0.012	1.008 (1.002–1.014)			
	Younger age	0.001	1.054 (1.023–1.087)			
	Diagnosis	0.012	0.337 (0.144–0.788)			
ODI 17-point	Higher baseline ODI	0.010	0.954 (0.921–0.989)	0.245	0.000	48.3%
change	Shorter symptom	0.064	1.009 (0.999–1.020)			
(*N*=116)[a]	duration					
	Diagnosis	0.063	0.441 (0.186–1.046)			
	Met CPR	0.007	3.387 (1.402–8.182)			
ODI 10-point	Higher baseline ODI	0.002	0.951 (0.921–0.981)	0.172	0.001	66.7%
change	Younger age	0.020	1.035 (1.005–1.065)			
(*N*=138)[b]	Diagnosis	0.091	0.490 (0.215–1.120)			
ODI five-point	Higher baseline ODI	0.008	0.956 (0.924–0.989)	0.139	0.006	79.2%
change	Younger age	0.043	1.034 (1.001–1.068)			
(*N*=144)[c]	Diagnosis	0.082	0.442 (0.176–1.110)			
ODI final ≤20%	Lower baseline ODI	0.000	1.113 (1.055–1.175)	0.491	0.000	56.1%
(*N*=107)[d]	Younger age	0.002	1.064 (1.023–1.105)			
	Shorter symptom duration	0.083	1.009 (0.999–1.018)			
	Diagnosis	0.061	0.366 (0.128–1.047)			
	Met CPR	0.029	3.288 (1.130–9.566)			

ODI = Oswestry disability index; CPR = Clinical prediction rule; FABQ-w = Fear avoidance beliefs questionnaire work subscale.
[a] *N* adjusted for 33 patients with baseline ODI of less than 17%; [b] *N* adjusted for 11 patients with baseline ODI of less than 10%; [c] *N* adjusted for five patients with baseline ODI less than 5%; [d] *N* adjusted for 42 patients with baseline ODI of less than 20%.

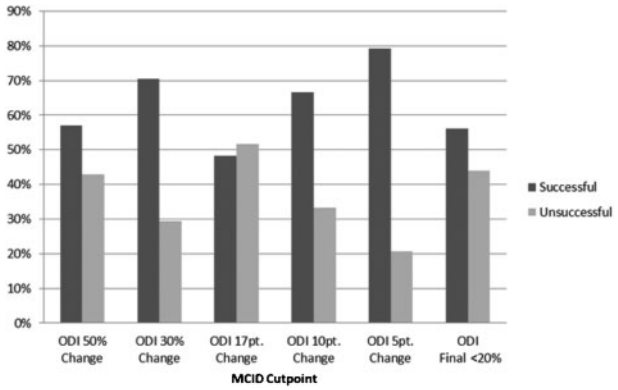

Figure 1 Patient success rate of full sample ($N = 149$).

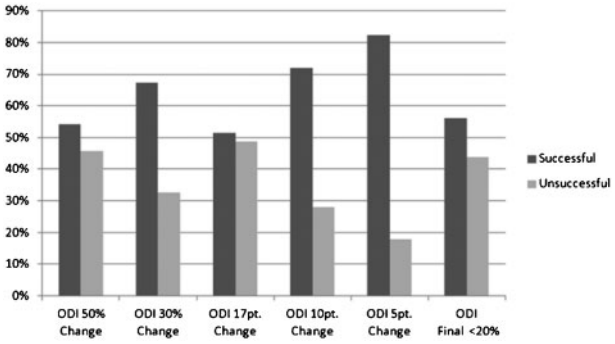

Figure 2 Patient success rate of sub-group ($N = 106$).

change model at 21.0%, and the 30% change model provided the lowest Nagelkerke R^2 at 19.1%. The VIFs for all models ranged from 1.001 to 1.121 indicating minimal to no covariance between the predictive variables within each model.

Figure 1 displays the success rate of the full sample based on the MCID cutpoint used. The five-point change produced the highest success rate with 79.2% of patients defined as 'recovered'. A MCID cutpoint of 17 points manufactured the lowest success rate at 48.3% of patients 'recovered'. Figure 2 shows the success rates of the subgroup of 107 patients with a baseline ODI greater than 20%. The five-point change continues to show the highest success rate at 82.2%, followed by a 10-point change (72% recovered), a 30% change (67.3% recovered), a final ODI of 20% or less (56.1% recovered), a 50% change (54.2% recovered), and a 17-point change (51.4% recovered).

Discussion

Our study sought to determine the relationship between MCID and predictive variables within a single patient population and the outcome measures derived from that population. The results of this study show that different MCIDs used on the same outcome measure across the same patient population lead to notably different predictive models. In

Table 3 Logistic regression modeling (only cases with initial Oswestry disability index (ODI) greater than 20%), including final predictor variables, individual *P* values for each model variable, as well as odds ratios and 95% confidence intervals

Model	Variables	Individual *P* value	Odds ratio (95% confidence interval)	Nagelkerke R^2	Model *P* value	% correct
ODI 50% change (*N*=107)	Younger age	0.020	1.039 (1.006–1.072)	0.271	0.000	54.2%
	Shorter symptom duration	0.062	1.009 (1.000–1.018)			
	Diagnosis	0.004	0.254 (0.100–0.645)			
	Met CPR	0.070	2.347 (0.933–5.906)			
ODI 30% change (*N*=107)	Younger age	0.029	1.036 (1.004–1.069)	0.191	0.002	67.3%
	Shorter symptom duration	0.011	1.012 (1.003–1.022)			
	Diagnosis	0.043	0.381 (0.149–0.972)			
ODI 17-point change (*N*=107)	Higher baseline ODI	0.037	0.961 (0.926–0.998)	0.260	0.000	51.4%
	Younger age	0.093	1.027 (0.996–1.060)			
	Shorter symptom duration	0.049	1.010 (1.000–1.020)			
	Diagnosis	0.029	0.354 (0.140–0.897)			
	Met CPR	0.034	2.722 (1.077–6.882)			
ODI 10-point change (*N*=107)	Higher baseline ODI	0.010	0.939 (0.894–0.985)	0.217	0.002	72.0%
	Younger age	0.058	1.034 (0.999–1.070)			
	Shorter symptom duration	0.031	1.010 (1.001–1.018)			
	Diagnosis	0.039	0.346 (0.126–0.949)			
ODI five-point change (*N*=107)	Higher baseline ODI	0.014	0.924 (0.807–0.984)	0.210	0.008	82.2%
	Lower BMI	0.016	1.160 (1.028–1.308)			
	Shorter symptom duration	0.083	1.009 (0.999–1.019)			
	Diagnosis	0.077	0.348 (0.108–1.121)			
ODI final $\leq$20% (*N*=107)	Lower baseline ODI	0.000	1.113 (1.055–1.175)	0.491	0.000	56.1%
	Younger age	0.002	1.064 (1.023–1.105)			
	Shorter symptom duration	0.083	1.009 (0.999–1.018)			
	Diagnosis	0.061	0.366 (0.128–1.047)			
	Met CPR	0.029	3.288 (1.130–9.566)			

ODI = Oswestry disability index; CPR = Clinical prediction rules; FABQ-w = Fear avoidance beliefs questionnaire work subscale.

addition, after modifying our analyses for baseline ODI scores, we found different predictive models, using different MCID interpretations. We feel that there are a number of reasons why the results are different among predictive models including variations in baseline characteristics, differences in success ratios depending on MCID interpretation, challenges to using a single point estimate to report change, and inherent fragility of predictive modeling.

Previous research has suggested that patient baseline characteristics, especially severity measures, can significantly influence the MCID.[1] Wright *et al.*[1] report, 'the use of the MCID score alone when determining treatment effects may be limited due to the inherent limitations of the methodology and baseline dependency of the sample'. In a study by Terwee *et al.*,[18] a large variation was found in MCID scores on the ODI using the same methodology across several studies, suggesting that the differences could be explained by baseline population characteristics. In our second logistic regression analyses, we removed the subjects who failed to present with an initial ODI score of 20%, and our models were notably different. It is likely that enrolling only subjects with higher disability would result in different findings as well.

In clinical practice, MCIDs are utilized on outcome measures to determine whether or not a patient is successfully responding to treatment. Thus, difference in one's interpretation of 'success' will lead to proportional differences between success and failure, when in reality the successive nature of recovery is likely present along a continuum. A MCID is a single point estimate and weaknesses of single point estimates involve the focus on measures of central tendencies, measures which may or may not truly represent change in a patient condition. The different MCID values for the ODI used in our study were captured from measures of central tendencies in different studies under different conditions. This is of considerable importance as the medical profession shifts from fee for service to fee for performance; recognizing the instability of the MCID and the need for a stable quality measure becomes more important to assure correct reimbursement.

The findings of our study also inherently suggest the fragility of predictive models. Without a universally stable definition of meaningful change, it is likely that different MCID scores will result in very different predictive variables in different patient populations. Clinical predictive rules are forms of predictive models and are assumed to be robust clinical tools that are usable in any given clinical situation. Our findings hint that this may not be the case. In our study, there were no variables that were universal predictors across all models, regardless of the baseline ODI measure. It is highly likely that models that do not involve universal predictors are not transferable across different patient spectrums (samples), as well as models that use different MCIDs.

Is there evidence to support which MCID is the best for determining clinically important differences? We would argue that there is not enough information available to support *any* of the choices provided in this manuscript. In an attempt to identify the most robust predictive models, we used the Nagelkerke R^2 to assess goodness of fit. The Nagelkerke R^2 theoretically represents the proportion of variance in the criterion that is explained by the predictors. Although it is not as discerning a measure as the R^2 used for a linear regression analysis, it does help define which predictive model is best able to explain the proportion of variance within the dedicated models. Our study findings suggest that the most robust predictive model for determining proportion of variance in the criterion is the original authors'[3] definition (an ODI score of 20% or less), a value that has yet to be analyzed in any study, outside of ours. Further, this does not suggest that a $\leq$20% of the final ODI is best at discriminating clinically important recovery from disability, only that the variables within the model fit best when that definition was used as the dependent variable. All other models which used change scores or end point score were much less robust. We feel that the minimally clinical important change score is a unique concept that is likely *different* for *each* patient and that one value does not appropriately capture change for everyone. We endorse none of the measures presented herein.

Limitations

There are limitations to our study including patient population and study design. Our study utilized the data from a RCT, which is not the optimal study design for prognostic studies. A larger sample size would have allowed for better generalizations, indicating a cohort study as a more appropriate form of design. Furthermore, a larger sample size would improve the precision of regression-based estimates.

Conclusion

The main finding that our secondary database analysis demonstrates is the extreme variability between predictive models created by using different MCIDs on the ODI within the same patient population. Our findings highlight the instability of predictive modeling, as these models are significantly affected by population baseline characteristics along with the MCID used. The fragility of predictive modeling creates difficulty applying CPR to clinical practice and suggests that CPR may not be reliable across all patient populations.

References

1 Wright A, Hannon J, Hegedus EJ, Kavchak AE. Clinimetrics corner: a closer look at the minimal clinically important difference (MCID). J Man Manip Ther 2012;20:164–70.
2 Rocchi MBL, Sisti D, Benedetti P, Valentini M, Bellagamba S, Federici A. Critical comparison of nine different self-administered questionnaires for the evaluation of disability caused by low back pain. Eura Medicophys 2005;41:275–81.
3 Fairbank J, Couper J, Davies J, O'Brien JP. The Oswestry low back pain questionnaire. Physiotherapy 1980;66:271–3.
4 Vianin M. Psychometric properties and clinical usefulness of the Oswestry Disability Index. J Chiropr Med 2008;7:161–3.
5 Davidson M, Keating JL. A comparison of five low back disability questionnaires: reliability and responsiveness. Phys Ther 2002;82:8–24.
6 Fritz JM, Herbert J, Koppenhaver S, Parent E. Beyond minimally important change: defining a successful outcome of physical therapy for patients with low back pain. Spine 2009;34:2803–9.
7 Gatchel RJ, Mayer TG. Testing minimal clinically important difference: consensus or conundrum? Spine J 2010;10:321–7.
8 Ostelo RWJG, Deyo RA, Stratford P, Waddell G, Croft P, Von Korff M, *et al*. Interpreting change scores for pain and functional status in low back pain: towards international consensus regarding minimal important change. Spine 2008; 33:90–4.
9 Maughan EF, Lewis JS. Outcome measures in chronic low back pain. Eur Spine J 2010;19:1484–94.
10 Ostelo RWJG, deVet HCW. Clinically important outcomes in low back pain. Best Pract Res Clin Rheumatol 2005;19:593–607.
11 Hagg O, Fritzell P, Nordwall A. The clinical importance of changes in outcome scores after treatment for chronic low back pain. Eur Spine J 2003;12:12–20.
12 Cleland JA, Whitman JM, Houser JL, Wainner RS, Childs JD. Psychometric properties of selected tests in patients with lumbar spinal stenosis. Spine J 2012;12:921–31.
13 Lauridsen HH, Manniche C, Korsholm L, Grunnet-Nilsson N, Hartvigsen J. What is an acceptable outcome of treatment before it begins? Methodological considerations and implications for patients with chronic low back pain. Eur Spine J 2009;18:1858–66.
14 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;81:776–88.
15 Norman GR, Stratford PW, Regehr G. Methodological problems in the retrospective computation to change: the lesson of Cronbach. J Clin Epidemiol 1997;50:869–79.
16 Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol 2003;56:395–407.
17 Copay AG, Subach BR, Glassman SD, Polly DW Jr, Schuler TC. Understanding the minimum clinically important difference: a review of concepts and methods. Spine J 2007;7:541–6.
18 Terwee CB, Roorda LD, Decker J, Bierma-Zeinstra SM, Peat G, Jordan KP, *et al*. Mind the MIC: a large variation among populations and methods. J Clin Epidemiol 2010;63:524–34.
19 Wang YC, Hart DL, Stratford PW, Mioduski JE. Baseline dependency of minimal clinically important improvement. Phys Ther 2011;91:675–88.
20 Grotle M, Foster NE, Dunn KM, Croft P. Are prognostic indicators for poor outcome different for acute and chronic low back pain consulters in primary care? Pain 2010;151:790–7.
21 Flynn T, Fritz J, Whitman J, Wainner R, Magel J, Rendeiro D, *et al*. A clinical prediction rule for classifying patients with low back pain who demonstrate short term improvement with spinal manipulation. Spine 2002;27:2835–43.
22 Childs JD, Fritz JM, Flynn TW, Irrgang JJ, Johnson KK, Majkowski GR, *et al*. Validation of a clinical prediction rule to identify patients with low back pain likely to benefit from spinal manipulation. Ann Intern Med 2004;141:920–8.
23 Cook CE, Learman KE, O'Halloran BJ, Showalter CR, Kabbaz VJ, Goode AP, *et al*. Which prognostic factors for low back pain are generic predictors of outcome across a range of recovery domains? Phys Ther 2013;93(1):32–40.
24 Cook C, Learman K, Showalter C, Kabbaz V, O'Halloran B. Early use of thrust manipulation versus non-thrust manipulation: A randomized clinical trial. Man Ther 2012; Oct 2. doi:pii: S1356-689X(12)00189-0. 10.1016/j.math.2012.08.005. [Epub ahead of print].
25 Vincent HK, Omli MR, Day T, Hodges M, Vincent KR, George SZ. Fear of movement, quality of life, and self-reported disability in obese patients with chronic lumbar pain. Pain Med 2011;12:154–64.
26 Heinrich M, Hafenbrack K, Michel C, Monstadt D, Marnitz U, Klinger R. Measures of success in treatment of success in treatment of chronic low back pain: pain intensity, disability and functional capacity: determinants of success in a multimodal day clinic setting. Schmerz 2011;25:282–9.
27 Melloh M, Elfering A, Egli Presland C, Roeder C, Barz T, Rolli Salathé C, *et al*. Identification of prognostic factors for chronicity in patients with low back pain: a review of screening instruments. Int Orthop 2009;33:301–13.
28 Steenstra IA, Verbeek JH, Heymans MW, Bongers PM. Prognostic factors for duration of sick leave in patients sick listed with acute low back pain: a systematic review of the literature. Occup Environ Med 2005;62:851–60.
29 Cleland JA, Fritz JM, Brennan GP. Predictive validity of initial fear avoidance beliefs in patients with low back pain receiving physical therapy: is the FABQ a useful screening tool for identifying patients at risk for a poor recovery? Eur Spine J 2008;17:70–9.
30 Kent P, Keating JL, Leboeuf-Y de C. Research methods for subgrouping low back pain. BMC Med Res Methodol 2010;10:62.
31 Waddell G, Newton M, Henderson I, Somerville D, Main CJ. A fear-avoidance beliefs questionnaire (FABQ) and the role of fear-avoidance beliefs in chronic low back pain and disability. Pain 1993;52:157–68.
32 Barakatt ET, Romano PS, Riddle DL, Beckett LA, Kravitz R. An exploration of Maitland's concept of pain irritability in patients with low back pain. J Man Manip Ther 2009;17:196–205.
33 Barakatt ET, Romano PS, Riddle DL, Beckett LA. The reliability of Maitland's irritability judgments in patients with low back pain. J Man Manip Ther 2009;17:135–40.
34 Maitland GD. Vertebral manipulation. 5th ed. Oxford: Butterworth-Heinemann; 1997, pp.93–114.
35 Field A. Discovering statistics using SPSS. Los Angeles, CA: Sage Publications; 2009.
36 O'Brien RM. A caution regarding rules of thumb for variance inflation factors. Qual Quant 2007;41:673–90.