ALEXIS A. WRIGHT, PT, PhD, DPT[1] • **CHAD E. COOK,** PT, PhD, MBA, OCS, FAAOMPT[2] • **G. DAVID BAXTER,** TD, BSc, DPhil[3] • **JOHN D. DOCKERTY,** PhD[4] • **J. HAXBY ABBOTT,** PhD, MScPT, FNZCP[5]

A Comparison of 3 Methodological Approaches to Defining Major Clinically Important Improvement of 4 Performance Measures in Patients With Hip Osteoarthritis

STUDY DESIGN: Prospective cohort study.

OBJECTIVES: To establish the major clinically important improvement (MCII) of the timed up-and-go test (TUG), 40-meter self-paced walk test (40-m SPWT), 30-second chair stand (30 CST), and a 20-cm step test in patients with hip osteoarthritis (OA) undergoing physiotherapy treatment. As a secondary aim, a comparison of methods was employed to evaluate the effect of method on the reported MCII.

BACKGROUND: Minimal clinically important difference scores are commonly used by rehabilitation professionals to determine patient response following treatment. A gold standard for calculating MCII has yet to be determined, which has resulted in problems of interpretation due to varied results.

METHODS: As part of a randomized controlled trial, 65 patients were randomized into a physiotherapy treatment group for hip OA, in which they completed 4 physical performance measures at baseline and 9 weeks. Upon completion of physiotherapy, patients assessed their response to treatment on a 15-point global rating of change scale (GRCS). MCII was estimated using 3 variations of an anchor-based method, based on the patient's opinion.

RESULTS: A comparison of 3 methods resulted in the following change scores being best associated with our definition of MCII: a reduction equal to or greater than 0.8, 1.4, and 1.2 seconds for the TUG; an increase equal to or greater than 0.2, 0.3, and 0.2 m/s for the 40-m SPWT; an increase equal to or greater than 2.0, 2.6, and 2.1 repetitions for the 30 CST; an increase equal to or greater than 5.0, 12.8, and 16.4 steps for the 20-cm step test.

CONCLUSION: The variation in methods provided very different results. This illustrates the importance of comparing methodologies and reporting a range of values associated with the MCII, as such values vary, depending upon the methodology chosen. *J Orthop Sports Phys Ther 2011;41(5):319-327, Epub 18 February 2011. doi:10.2519/jospt.2011.3515*

KEY WORDS: *outcome assessment, rehabilitation, task performance and analysis, timed up and go*

For rehabilitation professionals engaged in the treatment of osteoarthritis (OA), it is standard practice to perform objective assessments of physical function (using both physical performance and self-report measures) to obtain a picture of patient status. These assessments also serve as a baseline value for estimating changes (treatment effects) over time. While reliability and validity of some commonly used physical performance measures have been investigated in an OA population, most require further research regarding clinical utility and responsiveness.[14]

Physical performance measures have often been criticized, as detailed testing of their measurement properties has not been extensively reported.[14,22,33] Measures of responsiveness have commonly been reported as statistically significant change scores, which are useful in establishing the threshold of change needed beyond measurement error.[10] Investigation of minimal clinically important differences (MCIDs) of physical performance measures is warranted, as these have become commonly used outcome measures in the treatment of OA.[8,12,17,35-37] At present, the responsiveness (in terms of MCID) of the timed up-and-go (TUG) test, 40-meter self-paced walk test (40-m SPWT), 30-second chair stand (30 CST), and 20-cm step test has not been investigated in

[1]Centre for Physiotherapy Research, School of Physiotherapy, University of Otago, Dunedin, New Zealand. [2]Chair and Professor, Division of Physical Therapy, Walsh University, North Canton, OH. [3]Dean, School of Physiotherapy, University of Otago, Dunedin, New Zealand. [4]Associate Dean, Preventive and Social Medicine, Dunedin School of Medicine, University of Otago, Dunedin, New Zealand. [5]Senior Research Fellow, Orthopaedic Surgery Section, Department of Medical and Surgical Sciences, Dunedin School of Medicine, University of Otago, Dunedin, New Zealand. The MOA trial is funded by a grant from the Health Research Council of New Zealand (07/199 and 07/200) and a grant-in-aid from the New Zealand Lottery Grants Board (MR212664). The study was granted ethical approval by the Lower South Regional Ethics Committee of the New Zealand Ministry of Health (ethics reference: LRS/07/11/044). Address correspondence to Dr Alexis A. Wright, University of Illinois Medical Center, Fellowship in Orthopedic Manual Physical Therapy, 1801 W Taylor St, Chicago, IL 60612. E-mail: lexie.wright@gmail.com

TABLE 1 Description of Selected Physical Performance Measures

Measure	Description
Timed up-and-go test, s	Participants are asked to rise from a standard arm chair, walk as quickly but as safely as possible to a mark 3 m away, turn around, and return to the seated chair position. Subjects are timed for this test.[14,26]
40-m self-paced walk test, m/s	Participants are asked to walk as quickly but as safely as possible to a mark 10 m away, return, and repeat for a total distance of 40 m. Subjects are timed for this test and data are expressed as speed.[14]
30-s chair stand, n repetitions	Participants are asked to rise from a seated position to a standing position, with their arms folded across their chest, as many times as possible in 30 s. The number completed is recorded for this test.[11,18,28]
20-cm step test, n repetitions	Participants are asked to step up onto and down from a step 20-cm in height as many times as possible. The involved lower extremity acts as the working lower extremity, so that the patient steps up with the involved lower extremity and down with the uninvolved lower extremity. The number completed is recorded for this test. The test is discontinued if the maximum number of 50 steps is reached.[19-21]

our specific patient population.

Responsiveness, as investigated with the MCID, indicates whether patients experience a beneficial change following treatment that would mandate a change in patient management, in the absence of troublesome side effects (eg, increased irritability to treatment), excessive costs, and inconveniences (frequency of visits).[5,10] Recent literature has suggested that the reporting of MCID provide a more definite response or major clinically important improvement (MCII), thereby reducing the chance for error potentially associated with minimal improvement scores.[3]

Various methodological approaches to calculating MCID and MCII have been reported. These, however, have resulted in problems of interpretation, due to varying results of the different methods of calculation. In general, methodological approaches can be classified into 2 broad groups: anchor based and distribution based. Anchor-based approaches, as used in longitudinal studies, compare changes in the outcome measure score to an external criterion standard. Most commonly, the global rating of change score (GRCS) is used to define the MCID.[6,7] The GRCS is a self-report measure that asks patients to rate their current health status as worse, better, or the same on a numeric continuum and has been used for a number of different health-related conditions. It has been suggested that clinically important change include some form of patient report, which has necessitated the use of an anchor-based approach.[10]

The purpose of this study was to establish the MCII of the TUG, 40-m SPWT, 30 CST, and the 20-cm step test using 3 anchor-based methods in patients who had non-end stage hip OA and were undergoing a course of physiotherapy treatment.

METHODS

Participants

The sample consisted of patients with a clinical diagnosis of hip OA, who were part of a larger, randomized controlled trial designed to investigate the long-term effectiveness of 3 different physiotherapy programs, as compared to usual care, in patients with OA of the hip and knee.[1] The current study focused only on those 70 patients (23 in the exercise therapy group, 25 in the manual therapy group, and 22 in the exercise and manual therapy group) with OA of the hip, who were randomized into a physiotherapy treatment group. Patients assigned to the usual-care group were excluded from analysis, given that MCII scores are in reference to a measure's responsiveness to treatment.

The sample represented consecutive patients, from March 2008 to March 2009, who fulfilled the eligibility criteria. All patients agreed to be enrolled in the study and provided their signed informed consent. The study was granted ethical approval by The Lower South Regional Ethics Committee of the New Zealand Ministry of Health. Details of inclusion and exclusion criteria are described in detail elsewhere[1]; however, in brief, participants were included if they met clinical criteria for diagnosis of OA of the hip (criteria as established by the American College of Rheumatology [ACR]) and were able to walk 10 m without an assistive device at their baseline assessment.[1,2] Exclusion criteria included previous hip joint replacement surgery of the affected joint, an inability to comprehend and complete study assessments, or an inability to comply with instructions.[1]

Clinical Measures

TABLE 1 outlines the clinical measures used in this study. The TUG, 40-m SPWT, 30 CST, and the 20-cm step test were investigated for responsiveness and interrater reliability.

The GRCS at the 9-week follow-up was used as the anchor in the study. The GRCS is a measure of patient perception that asks respondents to rate the change in their symptoms. To minimize ambiguity and ensure valid information, the GRCS was worded as a condition-specific and construct-specific functional scale that specifically referred to hip OA and physical function, based on the recommendations of Kamper et al.[13] The question reads, "Please imagine how you would have described your level of function 9 weeks ago when you entered this trial. How do you rate your overall strength, endurance, and agility in performing activities today as compared to 9 weeks ago as far as your osteoarthritis of the left/right hip is concerned? Think of activities such as getting in/out of a car, standing up from a chair, walking for longer duration, walking more quickly, feeling steady on your feet without losing balance, performing daily activities such as putting on shoes and socks."

The GRCS used in this study had 15 possible numerical values corresponding to verbal descriptions ranging from +7 ("A very great deal better") to −7 ("A very great deal worse"), as described by Jaeschke et al.[10] The GRCS has been well validated and extensively used in research as an outcome measure and to compare outcome measures.[4,25]

Reliability

Interrater reliability of selected physical performance measures was assessed in a subsample of patients with a diagnosis of hip or knee OA who were part of the larger clinical trial. Two raters, who were trained physiotherapists, not involved in the intervention, and blinded to group allocation, performed the testing. These raters were considered to be representative of the population of practicing physiotherapists. Based on 2 measurements (1 by each rater) of each patient, it was estimated that 33 patients were required for the reliability assessment. Given that the desired level of reliability was 0.7, a significance level of 0.05 with a 1-sided test for this sample size would have 80% power to show a level of reliability of at least 0.4 (according to the classification of Landis and Koch,[15] 0.4 represents fair reliability and 0.7 substantial reliability).[15] Prior to performing any tests and measures, a complete guide to examination was developed, including operational definitions of each test. The 2 raters also went through training sessions to standardize their performance prior to the study. On the day of their entry into the clinical trial, rater 1 performed the baseline examination for each patient. The patients were scheduled for a second appointment within 7 days, prior to scheduling treatment. At this second appointment, rater 2, who was blinded to the findings of rater 1, assessed the physical performance measures. It was important that the patients were retested before they received any intervention with the potential to change the clinical features being measured. Each patient was examined in the same environment with the same equipment on both visits to reduce measurement error associated with external factors.

Intervention

Standardized interventions were provided at the School of Physiotherapy, University of Otago, under the supervision of licensed practicing physiotherapists (n = 5). These 5 therapists had been previously trained to administer the intervention protocols in a standardized manner. Patients underwent a 9-session physiotherapy program and were randomly allocated to receive either (*a*) manual therapy, (*b*) exercise therapy, or (*c*) both manual therapy and exercise therapy. The details of the intervention have been described elsewhere.[1]

Responsiveness

The investigation of responsiveness depends on the research design being employed during a period when change is expected.[14,29] Based on previous results,[9] it was recognized that physical performance measures could determine whether functional change had occurred following our physiotherapy program. A baseline examination was performed for all patients, and physical performance measures were repeated after the treatment period (at approximately 9 weeks). An estimate of meaningful change was obtained by having the patient complete a 15-point functional GRCS the day of the posttreatment follow-up visit.

Statistical Analysis

All statistical analyses were performed using Stata, Version 10.0 (StataCorp, College Station, TX). Descriptive statistics, as well as means and standard deviations, for baseline, 9-week, and change scores were calculated for the 4 physical performance measures.

An intraclass correlation coefficient ($ICC_{2,1}$) was used with continuous and ordinal variables and corresponding standard errors of measurement (SEMs) were calculated according to the following equation: $SEM = SD \times \sqrt{1 - ICC}$.[27] One SEM has been described as the preferred method for establishing the minimal detectable change (MDC) to identify important change beyond measurement error.[16,39] The SEM quantifies the measurement error in the units of the original measurement, which provides a clinically meaningful value to be interpreted by the clinician.[32] Ninety-five percent confidence intervals (CIs) were calculated for all reliability coefficients.

The MCII for the TUG, 40-m SPWT, 30 CST, and 20-cm step test was calculated using 3 variations associated with the anchor-based approach: (1) sensitivity- and specificity-based approach, (2) within-patients score change, and (3) between-patients score change.[6] While a fourth anchor-based method exists (social comparison approach), this method is not widely used and was therefore excluded from the analysis.[6] The GRCS was used as an external criterion, based on the patient's subjective perception of transition effects before and after receiving physiotherapy treatment. Criterion scores were dichotomized to identify those patients who experienced a major clinically meaningful reduction of symptoms.

Based on author consensus and a study of the literature, for our analyses we targeted a GRCS change of greater than +5 to represent important change and a score of +5 or lower to represent unimportant change. Several authors have used and identified scores of greater than +5 on the GRCS as reflective of care-terminating behavior; in other words, this value has been associated with change in care-seeking behavior in patients with low back pain.[24,31,40] We felt that targeting a score of greater than +5 on the GRCS would improve the likelihood that the anchor value would reflect tangible and marked improvements in the patient's condition.

To determine the threshold levels associated with our a priori definition of MCII, receiver operating characteristic (ROC) curves were used to discriminate between patients with major improvement and those with unimportant change, defining the sensitivity- and specificity-based approach.[6] Using this approach, the MCII is based on the concepts of sensitivity and specificity and the ability to correctly classify patients as improved or nonimproved. The MCII was determined to be the magnitude of change associated with the upper-left corner of the curve, where both sensitivity and 1 minus specificity are maximized.[4] Area under the ROC curve (AUC) estimates and their associated 95% CIs were also provided. AUC can be interpreted as the probability that a randomly chosen patient score showing major improvement will have a higher score than a randomly chosen patient score showing unimportant change.[6,30] An AUC between 0.7 and 0.8 is considered to be acceptable and 0.8 to 0.9 to be excellent.[6]

The within-patients change score was calculated as the mean change score (posttreatment minus baseline score) for each of the 4 physical performance measures, that corresponded to patients who were defined as having shown major improvement (that is, those with a GRCS greater than +5).[10,16] Using this approach, the MCII represents the mean change in scores of the major-improvement patients. The between-patients change score was calculated as the difference in the change score of the major-improvement and unimportant-change patients.[6] Using this approach, the MCII is defined as the difference in change scores between 2 adjacent levels of a scale. A Student *t* test was performed to determine statistical significance in the change scores for the 2 groups.

To assess the extent of patients' changes after interventions detected by the physical performance measures, the proportions of patients with change scores exceeding the values of the MCII estimates were examined. We also reported the sensitivity, specificity, and positive likelihood ratios associated with each of the MCII scores to evaluate the effect of method on reported MCII.

TABLE 2 Interrater Reliability for the 4 Outcome Measures

Measure	$ICC_{2,1}$ (95% CI)	SEM
Timed up-and-go test, s	0.87 (0.74, 0.94)	0.84
40-m self-paced walk test, m/s	0.95 (0.90, 0.98)	1.00
30-s chair stand, n repetitions	0.81 (0.63, 0.91)	1.27
20-cm step test, n repetitions	0.91 (0.82, 0.96)	5.80

Abbreviations: CI, confidence interval; ICC, intraclass correlation coefficient.

RESULTS

Twenty-nine people were recruited for the reliability study. **TABLE 2** provides a summary of the reliability analyses and estimates of SEM. All of the $ICC_{2,1}$ values were greater than 0.80.

A total of 70 patients were randomized into a treatment group and completed the baseline examination. Sixty-five (24 male, 41 female) of the 70 patients (93%) completed the 9-week follow-up examination. Patients ranged in age from 41 to 85 years, with a mean ± SD age of 66.5 ± 9.4 years. At the 9-week follow-up, 9 of the 65 patients (14%) were classified as major improvement and 56 (86%) were classified as unimportant change, based on the GRCS. Patients classified as unimportant change (n = 56) did not differ from the entire sample in terms of age, body mass index, or mean TUG, 40-m SPWT, 30 CST, and 20-cm step test scores (P>.05) at baseline and the 9-week follow-up. The clinical characteristics of the study sample are summarized in **TABLE 3.**

Results of the MCII estimates using the 3 methodologies and associated responsiveness characteristics for the 4 physical performance measures are given in **TABLE 4.** One extreme outlier was identified for the TUG and 40-m SPWT in the unimportant change group and was removed from analysis to reduce error and improve the accuracy of estimates. The outlier was identified based on the results of histogram testing for normal distribution of the data.

Both subgroups demonstrated mean improvement in change scores for the TUG, 40-m SPWT, and 30 CST; however, those classified as major improvement showed greater and statistically significant improvements in the 40-m SPWT, 30 CST, and 20-cm step test (P<.05). Patients classified as unimportant change demonstrated a statistically significant improvement on the 40-m SPWT (P<.001), a small, nonstatistically significant improvement for the TUG and 30 CST (P>.05), and a statistically significant (P = .04) worsening in the number of steps performed during the 20-cm step test.

DISCUSSION

Our study investigated differences in findings of 3 common methods of MCII anchor-based measures for 4 commonly used physical performance measures. All of the physical performance measures demonstrated good reliability for time and/or count.[27] Significant variability in MCII values was found among the 3 different approaches.

Each of the 3 methods involved a finding derived from a group level, and all 3 methods provided consistent evidence that the physical performance measures were responsive to change following physiotherapy treatment: time decreased for the TUG, walking speed increased for the 40-m SPWT, and number completed increased for the 30 CST and 20-cm step test.

TABLE 3 Clinical Characteristics of the Study Sample*

Measure	Overall (n = 65)[†]	Major Improvement (n = 9)	Unimportant Change (n = 56)[‡]
Timed up-and-go test, s			
Baseline scores	7.1 ± 2.3 (6.5, 7.6)	7.5 ± 2.5 (6.9, 8.1)	7.0 ± 2.3 (6.4, 7.5)
9-wk scores	6.7 ± 2.0 (6.2, 7.2)	6.1 ± 1.1 (5.8, 6.3)	6.8 ± 2.1 (6.3, 7.3)
Change scores	−0.4 ± 1.6 (−0.8, 0.0)	−1.4 ± 1.9 (−1.9, −0.9)	−0.2 ± 1.5 (−0.6, 0.2)
P value		.06	.28
40-m self-paced walk test, m/s			
Baseline scores	1.3 ± 0.3 (1.3, 1.4)	1.2 ± 0.3 (1.2, 1.8)	1.3 ± 0.3 (1.3, 1.4)
9-wk scores	1.4 ± 0.3 (1.3, 1.5)	1.5 ± 0.2 (1.4, 1.6)	1.4 ± 0.3 (1.3, 1.5)
Change scores	0.1 ± 0.1 (0.1, 0.1)	0.3 ± 0.1 (0.2, 0.3)	0.1 ± 0.1 (0.1, 0.1)
P value		.0002	<.001
30-s chair stand, n repetitions			
Baseline scores	10.1 ± 4.4 (9, 11.2)	8.4 ± 4.2 (7.4, 9.5)	10.3 ± 4.4 (9.3, 11.4)
9-wk scores	10.9 ± 5.5 (9.5, 12.2)	11.0 ± 3.8 (10.1, 11.9)	10.8 ± 5.7 (9.4, 12.2)
Change scores	0.8 ± 3.0 (0.0, 1.5)	2.6 ± 2.2 (2.0, 3.1)	0.5 ± 3.1 (−0.3, 1.2)
P value		.008	.24
20-cm step test, n repetitions			
Baseline scores	34.3 ± 20.0 (29.4, 39.1)	29.9 ±19.6 (25.1, 34.7)	35.0 ± 20.1 (30.0, 39.9)
9-wk scores	32.9 ± 19.2 (30.2, 39.6)	42.7 ± 14.6 (39.1, 46.2)	31.3 ± 19.5 (26.5, 36.1)
Change scores	−1.4 ± 14.6 (−5.0, 2.2)	12.8 ± 16.4 (8.8, 16.8)	−3.7 ± 13.0 (−6.9, −0.5)
P value		.047	.04

**Data are mean ± SD (95% confidence interval), except where otherwise indicated.*
[†]In the overall group, n = 64 for the timed up-and-go test and the 40-m self-paced walk test.
[‡]In the unimportant change group, n = 55 for the timed up-and-go test and the 40-m self-paced walk test.

The percent correctly classified (**TABLE 4**) can be interpreted as indicating both the sensitivity and specificity of a given measure. The larger the percent, the better the measure's ability to distinguish between patients who have experienced a clinically important change and those who have not. While the percent correctly classified is fairly similar for the 3 methods, it is worth noting that greater sensitivity was found using the sensitivity- and specificity-based approach (method 1). This indicates the need to consider more than 1 level of analysis when interpreting the clinical significance of change on these tests. In general, all of the physical performance measures demonstrated low sensitivity, in comparison to high specificity, suggesting that classifying improvement on MCII values alone may lead to a misclassification of patients as not having improved when, in fact, they did improve. On the other hand, the high specificity suggests that clinicians can be somewhat confident that, if a patient does meet the MCII on these measures, the patient has, in fact, improved. Also, those patients classified as nonimproved by the GRCS are unlikely to attain the MCII.

MCII values reported as a product of mean change scores are difficult to interpret, as they lack associated confidence intervals to account for the distribution among patient scores. Turner et al[39] have compared the strengths and weaknesses of using the mean score or an ROC curve for the anchor-based approach. In their discussion, they highlight that mean change scores are a poor descriptor of data that are not distributed normally and can be susceptible to data outliers.[39] The authors also mention that patients who score lower than the mean but higher than the cutoff of the next category on the GRCS may be misclassified as not having experienced important change, when, in fact, they have. The ROC approach addresses the limitations of mean change scores, as the entire cohort is dichotomized into 2 categories that correspond to the boundaries on the GRCS. The ROC approach can accommodate skewed data, is not vulnerable to a small number of values within a category, and maximizes the number of individuals correctly classified.[39] This finding was further highlighted in the current hip OA population.

Regardless of the methodological approach, more patients exceeded the values of the MCII for the 30 CST than values of the TUG, 40-m SPWT, or 20-cm step test; the 30 CST appears to be more responsive to detect change compared to the other measures. The strength of the responsiveness of the 30 CST may be related to the nature of the test, which suggests that rising from sitting may be more meaningful to this patient population.

In our attempt to define a major clinically important improvement, a cut-

TABLE 4 Responsiveness Characteristics for Physical Performance Measures at 9-Week Follow-up Using 3 Methods*

Measure/Method	MCII	Sensitivity[†]	Specificity[‡]	Percent Correctly Classified	Positive Likelihood Ratio[§]	AUC[‖]
Timed up-and-go test, s (n = 64)						
Method 1[‖]	–0.8	55.6 (28.4, 79.7)	78.2 (73.7, 82.1)	75.0 (67.3, 81.8)	2.6 (1.1, 4.5)	0.69 (0.48, 0.90)
Method 2[¶]	–1.4	33.3 (12.8, 59.4)	87.2 (83.9, 91.5)	79.7 (73.9, 87.0)	2.6 (0.8, 7.0)	...
Method 3[#]	–1.2 (P = .03)	33.3 (12.8, 60.1)	85.5 (82.1, 89.8)	78.1 (72.3, 85.7)	2.3 (0.7, 5.9)	...
40-m self-paced walk test, m/s (n = 64)						
Method 1	0.2	66.7 (38.3, 86.9)	85.5 (80.8, 88.8)	82.8 (74.8, 88.5)	4.6 (2.0, 7.7)	0.89 (0.76, 1.00)
Method 2	0.3 (P = .0002)	55.6 (30.2, 74.2)	94.5 (90.4, 97.6)	89.0 (81.9, 94.3)	10.2 (3.1, 31.0)	...
Method 3	0.2 (P<.001)	66.7 (39.0, 86.1)	90.9 (86.4, 94.1)	87.5 (79.7, 93.0)	7.3 (2.9, 14.6)	...
30-s chair stand, n repetitions (n = 65)						
Method 1	2.0	66.7 (37.3, 87.4)	67.9 (63.1, 71.2)	67.7 (59.6, 73.4)	2.1 (1.0, 3.0)	0.73 (0.55,0.91)
Method 2	2.6	66.7 (37.3, 87.4)	67.9 (63.1, 71.2)	67.7 (59.6, 73.4)	2.1 (1.0, 3.0)	...
Method 3	2.1 (P = .06)	66.7 (37.3, 87.4)	67.9 (63.1, 71.2)	67.7 (59.6, 73.4)	2.1 (1.0, 3.0)	...
20-cm step test, n repetitions (n = 65)						
Method 1	5.0	55.6 (28.9, 78.7)	87.5 (83.2, 91.2)	83.1 (75.7, 89.5)	4.4 (1.7, 9.0)	0.78 (0.63,0.93)
Method 2	12.8	33.3 (13.7, 48.5)	96.4 (93.3, 98.9)	87.7 (82.2, 91.9)	9.3 (2.0, 43.1)	...
Method 3	16.4 (P = .001)	33.3 (13.7, 48.5)	96.4 (93.3, 98.9)	87.7 (82.2, 91.9)	9.3 (2.0, 43.1)	...

Abbreviations: AUC, area under the curve; MCII, major clinically important improvement.

All P values are in reference to statistical significance in the change scores between patients with major improvement and those with unimportant change.

[†] *Values are percent (95% CI): [number of true positives/(number of true positives + number of false negatives)].*

[‡] *Values are percent (95% CI): [number of true negatives/(number of true negatives + number of false positives)].*

[§] *Values are positive likelihood ratio [sensitivity/(1 – specificity)].*

[‖] *The sensitivity- and specificity-based approach.*

[¶] *Within-patients score change approach.*

[#] *Between-patients score change approach.*

point of greater than +5 was chosen, as consistent with other authors who have found this cut-point to be representative of a change in condition when patients are no longer seeking care.[31] We felt this was consistent with our focus on patient perception of major clinically important change, as the idea of self-report of significant improvement and a change in care-seeking behavior is specifically associated with the patient's satisfaction with current health state. A cut-point of greater than +5 corresponds to "a great deal better" on the GRCS, which reduces the chance for misclassification potentially associated with marginal improvements, while sacrificing the sensitivity associated with moderate improvements.

The major-improvement group demonstrated statistically significant improvements in scores on physical performance measures, with the exception of the TUG, whereas the unimportant-change group did not. In particular, the major-improvement group demonstrated an improvement on the 20-cm step test while the unimportant-change group demonstrated worsening performance (**TABLE 3**). These findings confirm the discriminant construct validity of 3 physical performance measures in capturing changes in function. The results also suggest evidence of sensitivity among the 3 physical performance measures in capturing changes in function. Given that the TUG demonstrated a close-to-significant *P* value of .06, it may show significant changes and greater responsiveness with a larger sample size.

Above, we discussed 2 broad methods of determining MCII: the anchor-based and distribution-based methods. The advantage of the anchor-based approach is that change is linked to a meaningful external anchor, taking into account patient perspective.[7] However, anchor-based methods have been criticized for the effect of recall bias on long-term responsiveness, in that patient report of change has been found to strongly reflect current health status rather than the amount of change from baseline.[5,23] Anchor-based methods using global ratings have also been criticized for their ability to take into account the measurement precision of the global instrument.

Distribution-based approaches to determining MCII are based on the statistical characteristics of the sample and, in turn, on statistically significant changes in relation to the probability that the change has occurred by chance. One advantage of distribution-based methods is that they are able to account for change beyond some level of random variation.[7] However, a weakness of distribution-based methods is that there are few agreed-upon benchmarks for establishing clinically significant improvement. Further, distribution-based methods do not address the question of clinical importance, which is distinctly different from statistical significance.[6,7] Another limitation is that distribution-based methods are sample specific, in that, given a large sample size with wide variability, MCII values can still be extracted by distribution alone rather than actual improvement.

While previous authors have outlined the strengths and weaknesses associated with the 2 methods, a single preferred standardized methodology for calculating MCID or MCII has yet to be determined.[6,7] While we acknowledge the strengths and weaknesses of both methods, given the current definition of clinically meaningful change based on patient perception, we feel that the anchor-based approach is the appropriate standard.[10] Methods failing to acknowledge clinical significance (distribution-based methods) raise a separate question unrelated to the definition of MCII. For this reason, the functional GRCS was chosen as our longitudinal anchor for evaluating the clinical significance of individual change. We felt that modifications of the GRCS to focus the patient specifically on function and activity limitation, versus global disability, strengthened this tool as our reference criterion. While it could be argued that other function-specific measures, such as the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) physical function subscale, might serve as a better gold standard, previous studies have shown that the WOMAC physical function subscale is largely influenced by non–functional-related variables (eg, pain). Thus we did not feel that this measure would accurately capture the construct of function in this patient population.[34,35]

Given that our MCII values were derived at the group level, they may not be meaningful to the individual patient, regardless of the method used. This has been highlighted as one of the problems associated with the calculation of MCII.[5] MCII scores reported as a single point estimate based upon the average score of the group lack associated confidence intervals representative of the wide distribution of actual change score values. This is in contrast to most statistical procedures whereby group means accompanied by 95% confidence intervals are recommended, identifying a range of values to be expected. At the individual level, reported MCII values might also misclassify people below the mean as not having experienced a clinically important change, when, in fact, they have.[3] While MCII scores cannot be defined to fall within a range, the 3 methods to calculate MCII were performed to identify several possible cut-off scores for defining MCII of the 4 physical performance measures in our patient population.

Our findings illustrate the importance of comparing methodologies and reporting a range of values associated with MCII, as such values may vary, depending upon the method chosen. This further supports the idea of recognizing the "elusive nature of the MCID."[3] Previous authors have suggested a move away from the retrospective approach to responsiveness and a focus on identifying baseline attribute scores prognostically stratified as predictors of response to treatment.[5,23]

Limitations

Our target sample size of 33 participants for the reliability study was not met, which might have threatened the statistical significance of our results. When reviewing our data, the patient sample demonstrated good heterogeneity in terms of age, severity of symptoms, and site of condition. As our preselected levels of agreement ($P_0 = .40$, $P_1 = .70$) were robust, we are confident in our interpretation of the results.

The sample size for the purposes of ROC analysis was only 65, with just 9 patients identified as showing major improvement, which might have affected our precision estimates. However, even with this small sample size, significant changes were detected. Nevertheless, all results should be interpreted with caution and as preliminary, given the small sample of those classified as showing major improvement.

Given that MCII may vary, depending upon the specific impairments and activity limitations relevant to a particular patient population, the extent to which these values may be applied to populations other than those with hip OA is unclear. Further, there are known weaknesses in the use of a GRCS as the anchor from which to derive an MCII value. Single-item instruments are assumed to be less reliable and valid than multi-item instruments. This limitation may lead to greater misclassification at the individual level and also explain lower correlations between the anchor and change score for the individual physical performance measures.[38]

CONCLUSION

The TUG, 40-m SPWT, 30 CST, and 20-cm step test demonstrate good level of agreement in patients seeking physiotherapy care for hip OA. The results of this study show the variability among reported MCIIs based upon choice of methodological approach. Given the extent to which MCII is dependent upon the methodological approach chosen, we recognize the elusive nature of the MCII and recommend additional concurrent comparison studies of methodologies used to calculate MCII. ◉

KEY POINTS

FINDINGS: MCII values vary depending on the methodological approach chosen, causing confusion surrounding selection of appropriate MCII values for determining success rates in response to treatment. Given the large variation in MCII values across different methods, caution is needed when interpreting and using reported MCII values in efforts to avoid misclassification of patient response to treatment.

IMPLICATIONS: Use of alternative outcome measures, such as those associated with care-terminating behavior or satisfaction of current health state, should be considered in determining patient response to treatment, given the inherit weaknesses associated with MCII values.

CAUTION: Additional concurrent comparison studies of methodologies used to calculate MCII are needed.

ACKNOWLEDGEMENTS: *This research was supported (in part) by a contract from the Health Research Council of New Zealand (Project numbers 07/199, 07/200). Drs Wright, Abbott, and Baxter were supported in part by the Health Research Council of New Zealand and by the Centre for Physiotherapy Research, University of Otago. Dr Wright was also supported (in part) by the University of Otago postgraduate publishing bursary.*

REFERENCES

1. Abbott JH, Robertson MC, McKenzie JE, Baxter GD, Theis JC, Campbell AJ. Exercise therapy, manual therapy, or both, for osteoarthritis of the hip or knee: a factorial randomised controlled trial protocol. *Trials*. 2009;10:11. http://dx.doi.org/10.1186/1745-6215-10-11
2. Altman R, Alarcon G, Appelrouth D, et al. The American College of Rheumatology criteria for the classification and reporting of osteoarthritis of the hip. *Arthritis Rheum*. 1991;34:505-514.
3. Beaton DE, Boers M, Wells GA. Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. *Curr Opin Rheumatol*. 2002;14:109-114.
4. Childs JD, Piva SR, Fritz JM. Responsiveness of the numeric pain rating scale in patients with low back pain. *Spine (Phila Pa 1976)*. 2005;30:1331-1334.
5. Cook CE. Clinimetrics corner: the minimal clinically important change score (MCID): a necessary pretense. *J Man Manip Ther*. 2008;16:E82-83.
6. 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;7:541-546. http://dx.doi.org/10.1016/j.spinee.2007.01.008
7. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. *J Clin Epidemiol*. 2003;56:395-407.
8. Gandhi R, Tsvetkov D, Davey JR, Syed KA, Mahomed NN. Relationship between self-reported and performance-based tests in a hip and knee joint replacement population. *Clin Rheumatol*. 2009;28:253-257. http://dx.doi.org/10.1007/s10067-008-1021-y
9. Hoeksma HL, Dekker J, Ronday HK, et al. Comparison of manual therapy and exercise therapy in osteoarthritis of the hip: a randomized clinical trial. *Arthritis Rheum*. 2004;51:722-729. http://dx.doi.org/10.1002/art.20685
10. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials*. 1989;10:407-415.
11. Jones C, Rikli R, Beam W. A 30-s chair-stand test as a measure of lower body strength in community-residing older adults. *Res Q Exerc Sport*. 1999;70:113-119.
12. Juhakoski R, Tenhonen S, Anttonen T, Kauppinen T, Arokoski JP. Factors affecting self-reported pain and physical function in patients with hip osteoarthritis. *Arch Phys Med Rehabil*. 2008;89:1066-1073. http://dx.doi.org/10.1016/j.apmr.2007.10.036
13. Kamper SJ, Maher CG, Mackay G. Global rating of change scales: a review of strengths and weaknesses and considerations for design. *J Man Manip Ther*. 2009;17:163-170.
14. Kennedy DM, Stratford PW, Wessel J, Gollish JD, Penney D. Assessing stability and change of four performance measures: a longitudinal study evaluating outcome following total hip and knee arthroplasty. *BMC Musculoskelet Disord*. 2005;6:3. http://dx.doi.org/10.1186/1471-2474-6-3
15. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159-174.
16. Lin KC, Hsieh YW, Wu CY, Chen CL, Jang Y, Liu JS. Minimal detectable change and clinically important difference of the Wolf Motor Function Test in stroke patients. *Neurorehabil Neural Repair*. 2009;23:429-434. http://dx.doi.org/10.1177/1545968308331144
17. Lin YC, Davey RC, Cochrane T. Tests for physical function of the elderly with knee and hip osteoarthritis. *Scand J Med Sci Sports*. 2001;11:280-286.
18. Lipsitz LA, Jonsson PV, Kelley MM, Koestner JS. Causes and correlates of recurrent falls in ambulatory frail elderly. *J Gerontol*. 1991;46:M114-122.
19. Ljungquist T, Harms-Ringdahl K, Nygren A, Jensen I. Intra- and inter-rater reliability of an 11-test package for assessing dysfunction due to back or neck pain. *Physiother Res Int*. 1999;4:214-232.
20. Ljungquist T, Jensen IB, Nygren A, Harms-Ringdahl K. Physical performance tests for people with long-term spinal pain: aspects of construct validity. *J Rehabil Med*. 2003;35:69-75.
21. Ljungquist T, Nygren A, Jensen I, Harms-Ringdahl K. Physical performance tests for people with spinal pain--sensitivity to change. *Disabil Rehabil*. 2003;25:856-866. http://dx.doi.org/10.1080/0963828031000090579
22. Maly MR, Costigan PA, Olney SJ. Determinants of self-report outcome measures in people with knee osteoarthritis. *Arch Phys Med Rehabil*. 2006;87:96-104. http://dx.doi.org/10.1016/j.apmr.2005.08.110
23. Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. *J Clin Epidemiol*. 1997;50:869-879.
24. Perillo M, Bulbulian R. Responsiveness of the Bournemouth and Oswestry questionnaires: a prospective pilot study. *J Manipulative Physiol Ther*. 2003;26:77-86. http://dx.doi.org/10.1067/mmt.2003.6
25. Pham T, van der Heijde D, Altman RD, et al. OMERACT-OARSI initiative: Osteoarthritis Research Society International set of responder criteria for osteoarthritis clinical trials revisited. *Osteoarthr Cartilage*. 2004;12:389-399. http://dx.doi.org/10.1016/j.joca.2004.02.001
26. Podsiadlo D, Richardson S. The timed "Up & Go": a test of basic functional mobility for frail elderly persons. *J Am Geriatr Soc*. 1991;39:142-148.
27. Portney LG, Watkins MP. *Foundations of Clinical Research: Applications to Practice*. 2nd ed. Upper Saddle River, NJ: Prentice Hall; 2000.
28. Rikli R, Jones C. Development and validation of a functional fitness test for community-residing older adults. *J Aging Phys Activity*. 1999;7:129.
29. Roach K. Measurement of health outcomes: reliability, validity, and responsiveness. *J Prosthet Orthot* 2006;18:8-12.
30. Salaffi F, Carotti M, Grassi W. Health-related quality of life in patients with hip or knee osteoarthritis: comparison of generic and disease-specific instruments. *Clin Rheumatol*. 2005;24:29-37. http://dx.doi.org/10.1007/s10067-004-0965-9
31. Stratford PW, Binkley J, Solomon P, Gill C, Finch E. Assessing change over time in patients with low back pain. *Phys Ther*. 1994;74:528-533.
32. Stratford PW, Goldsmith CH. Use of the standard error as a reliability index of interest: an applied example using elbow flexor strength data. *Phys Ther*. 1997;77:745-750.
33. Stratford PW, Kennedy D, Pagura SM, Gollish JD. The relationship between self-report and performance-related measures: questioning the content validity of timed tests. *Arthritis Rheum*. 2003;49:535-540. http://dx.doi.org/10.1002/art.11196
34. Stratford PW, Kennedy DM. Does parallel item content on WOMAC's pain and function subscales limit its ability to detect change in functional status? *BMC Musculoskelet Disord*. 2004;5:17. http://dx.doi.org/10.1186/1471-2474-5-17
35. Stratford PW, Kennedy DM. Performance measures were necessary to obtain a complete picture of osteoarthritic patients. *J Clin Epidemiol*. 2006;59:160-167. http://dx.doi.org/10.1016/j.jclinepi.2005.07.012
36. Stratford PW, Kennedy DM, Riddle DL. New study design evaluated the validity of measures to assess change after hip or knee arthroplasty. *J Clin Epidemiol*. 2009;62:347-352. http://dx.doi.org/10.1016/j.jclinepi.2008.06.008
37. Stratford PW, Kennedy DM, Woodhouse LJ. Performance measures provide assessments

of pain and function in people with advanced osteoarthritis of the hip or knee. *Phys Ther*. 2006;86:1489-1496. http://dx.doi.org/10.2522/ptj.20060002

38. Terwee CB, Roorda LD, Dekker J, et al. Mind the MIC: large variation among populations and methods. *J Clin Epidemiol*. 63:524-534. http://dx.doi.org/10.1016/j.jclinepi.2009.08.010

39. Turner D, Schunemann HJ, Griffith LE, et al. The minimal detectable change cannot reliably replace the minimal important difference. *J Clin Epidemiol*. 63:28-36. http://dx.doi.org/10.1016/j.jclinepi.2009.01.024

40. Watson CJ, Propps M, Ratner J, Zeigler DL, Horton P, Smith SS. Reliability and responsiveness of the lower extremity functional scale and the anterior knee pain scale in patients with anterior knee pain. *J Orthop Sports Phys Ther*. 2005;35:136-146.