Level V Evidence

Research Pearls: The Significance of Statistics and Perils of Pooling. Part 1: Clinical Versus Statistical Significance

Joshua D. Harris, M.D., Jefferson C. Brand, M.D., Mark P. Cote, P.T., D.P.T., M.S.C.T.R., Scott C. Faucett, M.D., M.S., and Aman Dhawan, M.D.

Abstract: Patient-reported outcomes (PROs) are increasingly being used in today's rapidly evolving health care environment. The value of care provision emphasizes the highest quality of care at the lowest cost. Quality is in the eye of the beholder, with different stakeholders prioritizing different components of the value equation. At the center of the discussion are the patients and their quantification of outcome via PROs. There are hundreds of different PRO questionnaires that may ascertain an individual's overall general health, quality of life, activity level, or determine a body part-, joint-, or disease-specific outcome. As providers and patients increasingly measure outcomes, there exists greater potential to identify significant differences across time points due to an intervention. In other words, if you compare groups enough, you are bound to eventually detect a significant difference. However, the characterization of significance is not purely dichotomous, as a statistically significant outcome may not be clinically relevant. Statistical significance is the direct result of a mathematical equation, irrelevant to the patient experience. In clinical research, despite detecting statistically significant pre- and post-treatment differences, patients may or may not be able to perceive those differences. Thresholds exist to delineate whether those differences are clinically important or relevant to patients. PROs are unique, with distinct parameters of clinical importance for each outcome score. This review highlights the most common PROs in clinical research and discusses the salient pearls and pitfalls. In particular, it stresses the difference between statistical and clinical relevance and the concepts of minimal clinically important difference and patient acceptable symptom state. Researchers and clinicians should consider clinical importance in addition to statistical significance when interpreting and reporting investigation results.

From the Houston Methodist Orthopedics & Sports Medicine, Institute of Academic Medicine, Houston Methodist Research Institute (J.D.H.), Houston, Texas; Weill Cornell Medical College (J.D.H.), New York, New York; Heartland Orthopedic Specialists (J.C.B.), Alexandria, Minnesota; UConn Musculoskeletal Institute, Human Soft Tissue Research Laboratory UConn Health (M.P.C.), Farmington, Connecticut; The Centers for Advanced Orthopaedics, The Orthopaedic Center (S.C.F.), Rockville, Washington DC; and Penn State Hershey Bone and Joint Institute (A.D.), Hershey, Pennsylvania, U.S.A.

The authors report the following potential conflicts of interest or sources of funding: J.D.H. receives support from Smith & Nephew, NIA Magellan, Depuy Synthes, and SLACK; and is a board member of Arthroscopy, *AANA Research Committee, AOSSM Self-Assessment Committee, and AAOS OA Performance and Function Workgroup. J.C.B. is a board member of* Arthroscopy. *S.C.F. receives support from Smith & Nephew, Ceterix, Synthes, and Ossur; and is a board member of* Arthroscopy, *AOSSM, and ISAKOS. A.D. receives support from Biomet and Smith & Nephew; and is a board member of* Arthroscopy, *AANA, and OJSM.*

Received October 28, 2016; accepted January 23, 2017.

Address correspondence to Joshua D. Harris, M.D., Houston Methodist Orthopedics & Sports Medicine, 6550 Fannin Street, Smith Tower, Suite 2500, Houston, TX 77030, U.S.A. E-mail: cristyhayes@comcast.ne

0749-8063/161051/$36.00

http://dx.doi.org/10.1016/j.arthro.2017.01.053

Clinical studies are increasingly using patient-reported outcome (PRO) measures to quantitatively capture the effect of an intervention. Statistical analysis quantifies the size and precision of the differences in PRO scores before and after an intervention or between groups of patients. Statistical methods are blind to the clinical relevance. In other words, "patients don't know what their *P* value is, nor do they care." Consequently, statistical significance may or may not reflect a clinically meaningful change. A patient should be able to perceive the effect of the intervention (i.e., treatment) as "better" or "improved" if it is clinically meaningful. Furthermore, this difference should meet a minimum threshold of satisfaction ("I'm happy," "I'm satisfied," or "I'd do the intervention all over again").

A recent Level 1 evidence randomized controlled trial comparing autologous chondrocyte implantation (ACI) and microfracture in the knee at 3-year follow-up has been published by Saris et al.,[1] which highlights the difference between statistical significance and clinical relevance. The authors showed statistically significant

($P < .05$) differences favoring ACI in the overall Knee Injury and Osteoarthritis Outcome Score (KOOS) and 2 KOOS subscores (score range, 0-100). The "overall" KOOS difference between the 2 groups was only 2.3 points (77.6 vs 75.3). However, can a patient and his/her physician detect a difference between a KOOS of 77.6 and 75.3? After anterior cruciate ligament reconstruction, a change in a KOOS of 8 to 10 has been suggested to denote the minimal clinically important difference (MCID) detectable by a patient.[2] This indicates that a patient is unlikely to perceive a difference of 2.3 points, despite the statistically significant difference suggesting such. Thus, it is the responsibility of authors, journal editors, and readers themselves to ensure that the interpretation and clinical translation of an investigation's conclusions meet not only statistical but also clinically meaningful thresholds. This requires all participants in peer review to critically analyze study results and conclusions.

Statistical Basis for Clinical Relevance

Distinguishing between clinical and statistical significance requires an understanding of the role of a random error. A random error is variability around the true mean in the outcome being measured. The amount of posterior tibial slope, dimensions of the supraspinatus insertion, and preoperative pain levels are examples of objective measures that differ between individuals due to inherent biological variation. It is impossible to know the true mean of the population as that would require measuring every single individual. Thus, samples of the population are used and summarized statistically with means and measures of variation (standard deviation) around the mean. Sampling, however, could include individuals who are away from the mean, "skewing" the data toward an erroneous mean. Increasing sample size captures a larger portion of the population, improving precision in estimating the true mean and reducing the effect of the random error. In this regard, all study results are impacted by the random error to some degree. This is an important point as the random error has historically been treated as either present or absent.

Specifically, null hypothesis testing and the corresponding probability or "*P*" values dichotomize the effect of the random error to "significant" or "not significant." This fundamentally flawed approach can lead to the conclusion that a particular finding that does not cross the threshold of 0.05 is neither real nor important. The size of the difference and how precisely it has been estimated should be of interest rather than a yes or no decision as to whether a difference exists. Confidence intervals are a useful measure for determining how precisely a difference has been estimated and are a preferred measurement of the random error by *Arthroscopy*. The upper and lower limits of the interval and how wide or narrow they are provide a measure of precision. Narrow limits reflect greater precision in estimating the difference, thereby reducing the random error.

Type I Error and Type II Error

Interpretation of an investigation's results requires more judgment than simply determining the presence or absence of a statistically significant difference. The terms "type I error" and "type II error" are often used to discuss the risk of misinterpreting the results of a study. These errors are the result of testing the null hypothesis and need to be considered. For example, when comparing 2 or more samples, researchers first designate a null hypothesis. The null hypothesis generally states that the samples or groups being studied are not different from each other, in a superiority design study. If the results of the study have means that are different enough from each other in their size and variance, then comparing these samples will result in "statistical significance."

A type I error occurs when the null hypothesis is discarded despite it being true. In other words, the difference observed between groups is assumed and reported to be true when, in fact, the difference does not actually exist, particularly a problem in investigations with a large "n," such as "big data" research. When the results of a study are statistically significant, a type I error should be considered. Conversely, researchers can also make an error by stating that no difference exists between groups when there truly is a difference. This is a type II error. A type II error is much more common than type I and is a risk with small sample size studies that failed to detect significant differences between groups. In this situation, the alternative hypothesis is true rather than the null hypothesis.

The investigation's power is crucial to determine the impact of both significant and nonsignificant results. The ability of a study to detect a difference when one truly exists is referred to as statistical power, defined as 1 minus the type II error rate. Traditionally, investigators aim to enroll enough subjects to reach at least 80% power. This means that if a difference truly exists, the study will detect it 80% of the time (20% type II error rate). Some investigations may be underpowered to detect a difference between groups if one existed. Online power calculators are readily and freely available. The calculation of an adequate sample size requires the type I error rate (typically 0.05), desired power (typically 80%), the size of the difference the investigators are trying to detect (quantitatively, should be at least the MCID—optimizes collinearity of statistical significance and clinical importance), and an estimate of variation (standard deviation) to calculate the required sample size. Some calculators require effect

size, which is the size of difference to be detected divided by the standard deviation of the baseline scores.

The size of the difference and standard deviation both affect the required sample size. Smaller differences (distributions close together) and measures with a larger standard deviation (wide distributions) require larger sample sizes to show that 2 means overlap by less than 5%. As these calculations are based on estimating what difference may exist and how much it may vary, it is important to compare the size of difference and the standard deviation used in the sample size calculation with what was actually observed. A study may have observed a clinically meaningful difference but fail to reach statistical significance because the difference used in the power analysis was larger than what was observed. Similarly, a study may have enrolled enough subjects to meet their sample size calculation but fail to reach statistical significance if the standard deviation observed is greater than what was used in the calculation.

Quantifying Clinical Relevance

Although confidence intervals aid in determining the effect of the random error, the clinical relevance of a change in score needs to be considered. Contrary to discrete continuous data measures, such as infection rate or range of motion where a meaningful difference is easily derived, the relevance of a change in score on PRO measures is not always obvious. Several different measures exist to define thresholds of patient perception of "meaningful" or "relevant" (Table 1).[3] Three groups are typically used, in nomenclature, to define the threshold of relevance: (1) the minimum difference in an outcome score below which cannot be distinguished from the random error in the measurement (e.g., minimum detectable change [MDC], smallest detectable difference, minimal detectable difference); (2) the minimum difference in an outcome score measured before and after an intervention perceived as good or bad by the patient (e.g., MCID, minimal clinically significant difference, minimal clinically important change, minimal clinically important improvement, minimal perceptible clinical improvement); and (3) the difference in an outcome score that is considered clinically "relevant," "important," or "meaningful" (e.g., clinically important difference). A full discussion of each of these psychometric analytic terms, relevant to every body part, joint, system, or disease, is beyond the scope of this primer and the authors direct the readers to other comprehensive reviews.[3-17] The MCID is the most commonly used measure of clinical relevance in arthroscopic and related surgery.[17-20]

Although the MCID is a simple and straightforward concept, several factors impact the utility of this metric. A number of different methods have been developed to determine the MCID resulting in a range of values for a single PRO measure. MCID values are also dependent on the sample in which they were derived creating problems with applicability. However, the MCID can reflect either an improvement or a worsening. In addition to detection of a magnitude of difference, patients assess and appreciate whether or not their outcome meets their individual definition or perception of "satisfaction" or "happiness" and/or "undergo the intervention again." The latter concept refers to the patient acceptable symptom state (PASS) and substantial clinical benefit (SCB) (Table 1).[16,20,21] MCID and PASS values for some of the most commonly used PRO questionnaires in arthroscopic and related surgery are listed in Tables 2 to 5. Tables 2 to 5 present a highly valuable resource for authors, reviewers, and editors to reference for several aspects of study design, conduct, reporting, and outcome interpretation. Power analysis calculation, psychometric property (proper outcome score development, validity, reliability, and responsiveness testing) evaluation, and manuscript peer review illustrate some examples of the utility of Tables 2 to 5. Although Tables 2 to 5 present a large volume of data, these values are constantly being modified and optimized with the rapid growth and evolution of research in arthroscopic and related surgery.

Levy et al.[16] published a recent *Arthroscopy* meta-analysis investigating a large number (n = 81) of studies (9,317 hips) that reported outcomes of subjects undergoing primary hip arthroscopy. The authors showed that subjects met the MCID 97%, 90%, and 93% of the time for the modified Harris hip score, hip outcome score—activities of daily living, and hip outcome score—sport-specific subscale, respectively.

Table 1. Measures of Clinical Relevance

Measure of Change (Δ)	Satisfaction Threshold
• MCID (minimal clinically important difference) • MIC (minimal important change) • MCIC (minimal clinically important change) • MCII (minimal clinically important improvement) • MCSD (minimal clinically significant difference) • MPCI (minimal perceptible clinical improvement) • CID (clinically important difference) • MID (minimal important difference) • MDC (minimum detectable change) • MDD (minimal detectable difference) • SDD (smallest detectable difference)	• PASS (patient acceptable symptom state) • SCB (substantial clinical benefit)

Table 2. Characteristics of Common Patient-Reported Outcome Scores for Patients With Hip Pain

PRO	Subscores/Domains	Number of Questions	Range	MCID	PASS	Condition
mHHS	Pain	1	0-91 (scaled 0-100)	MDC 12 (individual); 2 (group)	74	Hip arthroscopy patients[20,22]
	Function	7		MIC 8		
HOS	ADL	19	0-68 (scaled 0-100)	MCID 9 (individual); 1 (group)	87	Patients aged 13-66 yr with hip dysfunction; measures function, not symptoms[20,23-25]
	SSS	9	0-36 (scaled 0-100)	MCID 6 (individual); 2 (group)	75	
HOOS	Pain	10	0-100 per subscore	MDC 10 (individual); 1 (group)	n/r	Patients with hip and/or groin disability due to arthritic and nonarthritic conditions[26-29]
	Symptoms	5	Only subscores	MDC 14 (individual); 2 (group)		
	ADL	17	No overall score	MDC 9 (individual); 1 (group)		
	Sports	4		MDC 17 (individual); 2 (group)		
	Quality of life	4		MDC 15 (individual); 2 (group)		
HAGOS	Pain	10	0-100 per subscore	SDC 19 (individual); 2.8 (group)	n/r	Physically active patients with long-standing hip and/or groin pain[30]
	Symptoms	7	Only subscores	SDC 18 (individual); 2.7 (group)		
	ADL	5	No overall score	SDC 20 (individual); 3.0 (group)		
	Sports	8		SDC 22 (individual); 3.3 (group)		
	Physical activities	2		SDC 34 (individual); 5.2 (group)		
	Quality of life	5		SDC 18 (individual); 2.7 (group)		
iHOT-12	Symptoms	4	0-100 per question	n/r	n/r	Young (18-60 yr), active patients with hip disorders[31,32]
	Sports	3				
	Job related	1				
	Social, emotional	4				
iHOT-33	Symptoms	16	0-100 per question	MCID 6.1	n/r	Young (18-60 yr), active patients with hip disorders[31]
	Sports	6				
	Job related	4				
	Social, emotional	7				
NAHS	Pain	5	0-80 (scaled 0-100)	MDC 10 (individual); 2 (group)	n/r	Young active nonarthritic patients with hip problems[33]
	Mechanical	4				
	Functional	5				
	Activity	6				

ADL, activities of daily living; HAGOS, The Copenhagen Hip and Groin Outcome Score; HOOS, hip dysfunction and osteoarthritis outcome score; HOS, hip outcome score; iHOT-12, International Hip Outcome Tool-12; iHOT-33, International Hip Outcome Tool-33; MCID, minimal clinically important difference; MDC, minimal detectable change; mHHS, modified Harris Hip score; MIC, minimal important change; NAHS, nonarthritic hip score; n/r, not yet reported in the literature; PASS, patient acceptable symptom state; SDC, smallest detectable change; SSS, sport-specific subscore.

Table 3. Characteristics of Common Patient-Reported Outcome Scores for Patients With Knee Pain

PRO	Subscores/Domains	Number of Questions	Range	MCID	PASS	Condition
IKDC-SKF	Symptoms	7	0-100	MCID 6.3 (6 mo)	75.9	Patients with knee conditions including ligament, meniscal injury, articular cartilage lesions, patellofemoral pain, osteoarthritis[34,35]
	Function	2		MCID 16.7 (12 mo)		
	Sports	10				
KOOS	Pain	9	0-100 per subscore	SDC 16.6; MIC 16.7	88.9	Young, middle-aged, elderly patients with knee injury and/or osteoarthritis[34,36,37]
	Symptoms	7	Only subscores	SDC 17.4; MIC 10.7	57.1	
	ADL	17	No overall score	SDC 15.7; MIC 18.4	100	
	Sports	5		SDC 25.1; MIC 12.5	75.0	
	Quality of life	4		SDC 18.8; MIC 15.6	62.5	
Lysholm	Pain, swelling, limp, squatting, instability, stairs, support, locking	8	0-100	MDC 8.9	n/r	Patients with many knee conditions, especially knee ligaments, including ACL injury, ACLR, meniscal injury[38]
Cincinnati	Sports	1	0-100	MCID 14.0 (6 mo)	n/r	Patients with many knee conditions including ACI, ACLR, other injuries[35,39]
	Function	6		MCID 26.0 (12 mo)		
	Pain, swelling, giving way	1				
WOMAC	Pain	5	2 scoring methods	MCID 17.5 (6 mo); 7.5 (12 mo)	n/r	Patients with osteoarthritis of the knee[40-42]
	Stiffness	2	0-4 Likert	MCID 6.3 (6 mo); 18.8 (12 mo)		
	Physical function	17	0-100 VAS	MCID 8.1 (6 mo); 5.9 (12 mo)		
	Total	24	Higher score = more limitation	MCID 11.5 (6 mo); 11.5 (12 mo)		
Tegner activity	Sport level	1	0-10	MDC 1	n/r	Patients with many knee conditions, especially knee ligaments, including ACL injury, ACLR, meniscal injury[38]

ACI, autologous chondrocyte implantation; ACL, anterior cruciate ligament; ACLR, anterior cruciate ligament reconstruction; ADL, activities of daily living; IKDC-SKF, International Knee Documentation Committee Subjective Knee Form; KOOS, knee injury and osteoarthritis outcome score; MCID, minimal clinically important difference; MDC, minimal detectable change; MIC, minimal important change; n/r, not yet reported in the literature; PASS, patient acceptable symptom state; SDC, smallest detectable change; VAS, visual analog scale; WOMAC, Western Ontario and McMaster Universities Arthritis Index.

Table 4. Characteristics of Common Patient-Reported Outcome Scores for Patients With Shoulder Pain

PRO	Subscores/Domains	Number of Questions	Range	MCID	PASS	Condition
WOSI	Symptoms	10	100-0 per question	MCID 220 (10.4%)	n/r	Patients with shoulder instability[43,44]
	Sports/work	4	2,100-0 overall			
	Lifestyle	4				
	Emotions	3				
WORC	Pain	6	100-0 per question	MCID 245.26 (11.7%)	n/r	Patients with rotator cuff problems[45-47]
	Sports	4	2,100-0 overall			
	Work	4				
	Social	4				
	Emotional	3				
WOOS	Pain	6	100-0 per question	n/r	n/r	Patients with shoulder osteoarthritis[48]
	Sports	5	1,900-0 overall			
	Lifestyle	5				
	Emotional	3				
ASES	Pain	1	0-100	MCID 6.4	n/r	Patients with shoulder instability,
	Function	10		MDC 9.7		rotator cuff disease, glenohumeral arthritis[49-51]
Constant	Pain	1	0-100	MCIC 11	44	Patients with many shoulder conditions,
	ADL	3		MCID 10.4		including total shoulder arthroplasty, rotator
	Mobility	4				cuff repair, adhesive capsulitis, and proximal
	Power/strength	1				humerus fractures[18,52,53]
SST	Physical function	12	0-12 (0-100 scaled)	SDC 2.8	n/r	Functional limitations of the affected shoulder in
				MIC 2.2		patients with shoulder dysfunction[50,54]
DASH	Physical function	21	100-0	MCID 10.2	43	Physical disability and symptoms of the upper
	Pain	5		MDC 6.6-12.2		extremities in people with upper extremity
	Emotional, social	4		SDC 16.3		disorders (hand, wrist, elbow, shoulder)[50,53-55]
				MIC 12.4		
SPADI	Pain	5	0-11 VAS/question	MCID 8-13.2	41	Pain and disability associated with shoulder
	Disability	8	0-100 scaled	MDC 18		pathology from musculoskeletal, neurogenic,
						or other origin[50,53,56]

ADL, activities of daily living; ASES, American Shoulder and Elbow Surgeons Score; DASH, Disabilities of the Arm, Shoulder, and Hand; MCIC, minimal clinically important change; MCID, minimal clinically important difference; MDC, minimal detectable change; MIC, minimal important change; SDC, smallest detectable change; n/r, not yet reported in the literature; PASS, patient acceptable symptom state; SPADI, Shoulder Pain and Disability Index; SST, simple shoulder test; VAS, visual analog scale; WOOS, Western Ontario Osteoarthritis of the Shoulder Index; WORC, Western Ontario Rotator Cuff Index; WOSI, Western Ontario Shoulder Instability Index.

Table 5. Characteristics of Common Patient-Reported Outcome Scores for Patients With Wrist and Elbow or Foot and Ankle Pain

PRO	Subscores/Domains	Number of Questions	Range	MCID	PASS	Condition
DASH	Physical function	21	100-0	MCID 10-13.5	n/r	Wrist conditions, ulnar impaction; tendonitis, arthritis, or nerve compression syndromes from forearm to hand[11,57]
	Pain	5		MDC 9.3		
	Emotional, social	4				
PRWE	Pain	5	100-0	MCID 14-17	n/r	Wrist conditions, ulnar impaction; tendonitis, arthritis, or nerve compression syndromes from forearm to hand[11,57]
	Function	10		MDC 7.7		
FAAM	ADL	21	0-84	MCID 8 (ADL)	n/r	Patients with chronic ankle instability[58]
	Sports	8		MCID 9 (sports)		
				MDC 5.7 (ADL)		
				MDC 12.3 (sports)		
MOXFQ	Pain	5	100-0	MCID 12 (pain)	n/r	Patients with hallux valgus[59]
	Walking/standing	7		MCID 16 (walking/standing)		
	Social	4		MCID 24 (social)		

ADL, activities of daily living; DASH, Disabilities of the Arm, Shoulder, and Hand; FAAM, Functional Ankle Ability Measure; MCID, minimal clinically important difference; MDC, minimal detectable change; MOXFQ, Manchester-Oxford Foot Questionnaire; n/r, not yet reported in the literature; PASS, patient acceptable symptom state; PRWE, patient-rated wrist evaluation.

However, only 88%, 25%, and 30% of subjects met PASS, respectively. This shows that although patients are able to detect their change as an improvement, the absolute value of their "level of satisfaction" did not meet their definition of "satisfaction." MCID and PASS values can be calculated using different methods: consensus methods, anchor-based methods, and distribution-based methods.

Consensus Methods

As the name implies, a consensus method requires a panel of experts in the topic of the outcome measure to meet, discuss the outcome, individually select MCID, discuss differences in value selection, then achieve consensus based on the individual estimates. A limitation of the consensus method is that it is based on the perspective of the examiner, clinician, or researcher, and not the patient or the subject.

Anchor-Based System

The anchor-based method for MCID calculation compares change in scores with an "anchor" as a reference. The anchor must have face validity representing the outcome of interest.[3] A popular anchor is the anchor question at a specific point in time after treatment the patient might be asked: "Do you feel that you are improved by your treatment?" Answers to anchor questions could vary from a binary "yes" or "no" or a Likert-type scale (e.g., "much better," "slightly better," "about the same," "somewhat worse," and "much worse"). The latter set of answers are part of the Medical Outcomes Study Short Form-36, which is a tool often used for assessing the MCID. Therefore, the Short Form-36 can be used with another PRO to determine the MCID. The disadvantage of the anchor system is that the limited number of responses can reduce the accuracy of estimating the MCID.

Distribution-Based Methods

Distribution-based methods use statistics to measure the variation of a variable of interest and determine what degree of change in that variable is clinically important.[3] The use of standard error of measurement (SEM) illustrates the inexactness of the measurement. An MCID below the SEM does not represent a true difference or change. Another distribution method uses the MDC, which is the MDC above that of the measurement error. The MCID is then the upper limit on the 95% confidence interval for the average change in nonresponders. The basis of SEM as MCID is relegated on the fact that MCID values are usually 1/2 standard deviation of the mean.

Interpretation of MCID Values

The MCID is a metric for "within-individual" change. It is a value that represents a change in a single patient

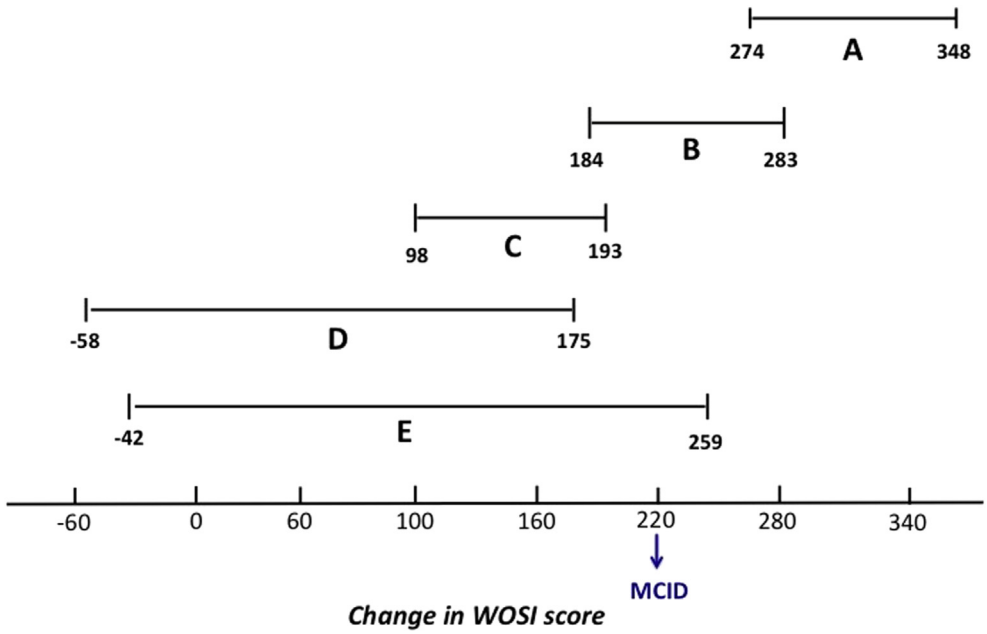

Fig 1. Confidence intervals for change scores on the Western Ontario Shoulder Instability Index (WOSI). The minimal clinically important difference (MCID) on the WOSI is 220. A: the difference observed is statistically significant and the lower limit of the confidence interval is greater than that of the MCID and is therefore clinically relevant; B: the difference is statistical significant but the lower limit is below the MCID and the higher limit is above the MCID indicating a clinically relevant change may exist; C: statistical significant difference; however, the upper limit of the confidence interval is less than the MCID indicating that the result is not clinically relevant; D: the difference is not statistically significant and the upper limit of the confidence interval is less than the MCID indicating a true negative finding; E: difference is not statistically significant, but the confidence interval overlaps the MCID indicating that the outcome is inconclusive.

over time and is intended to reflect the threshold for a clinically meaningful difference within an individual. This is an important point to be aware of as MCIDs are often applied on the group level rather than the individual level. Consider a study comparing the difference in outcome scores between 2 groups. One treatment group may have several patients whose change in score meets the MCID but due to a few outliers reporting little change, the group mean may fall short of the MCID. Conversely, a few outliers with large changes in scores may result in a group mean that meets the MCID despite the most of the group experiencing little to no change. In these situations, the proportion of patients in each group whose change in score crossed the threshold of the MCID should be compared rather than the mean change in score for each group. In the randomized trial published by Saris et al.,[1] who compared ACI to microfracture using KOOS as the primary clinical outcome measure, subjects were declared dichotomously as "treatment responders" if their overall KOOS increased at least 10 percentage points from baseline (or "nonresponder" if did not) and/or if 3 or more KOOS subdomains increased at least 10 percentage points from baseline (or "nonresponder" if did not). The proportion of treatment responders in each group could then be determined and compared, helping to reduce the effect of outliers on interpretation of the clinical importance of the difference observed from the intervention. Although Saris et al.[1] did not use MCID as their threshold definition of "treatment responder," their arbitrary value was not too dissimilar from the MIC (10.7 to 18.4; Table 3).

Confidence intervals are very useful for determining whether clinically meaningful differences truly exist, especially in studies that do not report the proportion of patients meeting the MCID. Figure 1 displays 95% confidence intervals different MCID finding using the Western Ontario Shoulder Instability Index.[43] If the difference observed is both statistically significant and the lower limit of the confidence interval is greater than that of the MCID, the result is clinically relevant (Fig 1A).[3] If statistical significance is achieved, but the confidence interval overlaps the MCID, then it is possibly clinically relevant (Fig 1B). If statistical significance is achieved, but the upper limit of the confidence interval is less than the MCID, then the result is not clinically relevant (Fig 1C). If the difference observed is not statistically significant and the upper limit of the confidence interval is less than the MCID, then the result represents a true negative finding (Fig 1D). If the result is not statistically significant, but the confidence interval overlaps the MCID, then the actual outcome is inconclusive (Fig 1E).

Conclusions

PROs are increasingly being used in today's rapidly evolving health care environment. To determine if the conclusion of a study is truly clinically relevant, the statistical analysis and magnitude of improvement must be perceived by the patient as significant and achieve a threshold of satisfaction. Psychometric measures of MCID and PASS represent these perceived differences and thresholds, respectively. These values are unique to each individual joint, body part, system, and disease and have values that vary based on the individual PRO used, the population in which it was studied, and the method on which it was calculated. In clinical research reporting outcomes of subjects undergoing arthroscopic and related surgery, researchers and clinicians must consider clinical importance in addition to statistical significance when interpreting and reporting investigation results.

References

1. Saris DB, Vanlauwe J, Victor J, et al. Treatment of symptomatic cartilage defects of the knee: Characterized chondrocyte implantation results in better clinical outcome at 36 months in a randomized trial compared to microfracture. *Am J Sports Med* 2009;37:10s-19s (suppl 1).
2. Roos EM, Lohmander LS. The knee injury and osteoarthritis outcome score (KOOS): From joint injury to osteoarthritis. *Health Qual Life Outcomes* 2003;1:64.
3. Katz NP, Paillard FC, Ekman E. Determining the clinical importance of treatment benefits for interventions for painful orthopedic conditions. *J Orthop Surg Res* 2015;10:24.
4. Mithoefer K, Acuna M. Clinical outcomes assessment for articular cartilage restoration. *J Knee Surg* 2013;26:31-40.
5. Berliner JL, Brodke DJ, Chan V, SooHoo NF, Bozic KJ. Can preoperative patient-reported outcome measures be used to predict meaningful improvement in function after TKA? *Clin Orthop Relat Res* 2017;475:149-157.
6. Ebert JR, Smith A, Wood DJ, Ackland TR. A comparison of the responsiveness of 4 commonly used patient-reported outcome instruments at 5 years after matrix-induced autologous chondrocyte implantation. *Am J Sports Med* 2013;41:2791-2799.
7. Lee WC, Kwan YH, Chong HC, Yeo SJ. The minimal clinically important difference for Knee Society Clinical Rating System after total knee arthroplasty for primary osteoarthritis [published online June 21, 2016]. *Knee Surg Sports Traumatol Arthrosc*. doi:10.1007/s00167-016-4208-9.
8. Franchignoni F, Vercelli S, Giordano A, Sartorio F, Bravini E, Ferriero G. Minimal clinically important difference of the disabilities of the arm, shoulder and hand outcome measure (DASH) and its shortened version (QuickDASH). *J Orthop Sports Phys Ther* 2014;44:30-39.
9. Kemp KA, Sheps DM, Beaupre LA, Styles-Tripp F, Luciak-Corea C, Balyk R. An evaluation of the responsiveness and discriminant validity of shoulder questionnaires among patients receiving surgical correction of shoulder instability. *ScientificWorldJournal* 2012;2012:410125.
10. Malay S, Chung KC. The minimal clinically important difference after simple decompression for ulnar neuropathy at the elbow. *J Hand Surg* 2013;38:652-659.
11. Sorensen AA, Howard D, Tan WH, Ketchersid J, Calfee RP. Minimal clinically important differences of 3 patient-rated outcomes instruments. *J Hand Surg* 2013;38:641-649.
12. Tashjian RZ, Hung M, Keener JD, et al. Determining the minimal clinically important difference for the American Shoulder and Elbow Surgeons score, Simple Shoulder Test, and visual analog scale measuring pain after shoulder arthroplasty. *J Shoulder Elbow Surg* 2017;26:144-148.
13. Torrens C, Guirro P, Santana F. The minimal clinically important difference for function and strength in patients undergoing reverse shoulder arthroplasty. *J Shoulder Elbow Surg* 2016;25:262-268.
14. Quintana JM, Escobar A, Bilbao A, Arostegui I, Lafuente I, Vidaurreta I. Responsiveness and clinically important differences for the WOMAC and SF-36 after hip joint replacement. *Osteoarthritis Cartilage* 2005;13:1076-1083.
15. Salaffi F, Stancati A, Silvestri CA, Ciapetti A, Grassi W. Minimal clinically important changes in chronic musculoskeletal pain intensity measured on a numerical rating scale. *Eur J Pain* 2004;8:283-291.
16. Levy DM, Kuhns BD, Chahal J, Philippon MJ, Kelly BT, Nho SJ. Hip arthroscopy outcomes with respect to patient acceptable symptomatic state and minimal clinically important difference. *Arthroscopy* 2016;32:1877-1886.
17. Cvetanovich GL, Weber AE, Kuhns BD, et al. Clinically meaningful improvements after hip arthroscopy for femoroacetabular impingement in adolescent and young adult patients regardless of gender [published online August 29, 2016]. *J Pediatr Orthop*. doi:10.1097/BPO.0000000000000852.
18. Kukkonen J, Kauko T, Vahlberg T, Joukainen A, Aarimaa V. Investigating minimal clinically important difference for Constant score in patients undergoing rotator cuff surgery. *J Shoulder Elbow Surg* 2013;22:1650-1655.
19. Coe MP, Sutherland JM, Penner MJ, Younger A, Wing KJ. Minimal clinically important difference and the effect of clinical variables on the ankle osteoarthritis scale in surgically treated end-stage ankle arthritis. *J Bone Joint Surg Am* 2015;97:818-823.
20. Chahal J, Van Thiel GS, Mather RC III, et al. The patient acceptable symptomatic state for the modified Harris hip score and hip outcome score among patients undergoing surgical treatment for femoroacetabular impingement. *Am J Sports Med* 2015;43:1844-1849.
21. Glassman SD, Copay AG, Berven SH, Polly DW, Subach BR, Carreon LY. Defining substantial clinical benefit following lumbar spine arthrodesis. *J Bone Joint Surg Am* 2008;90:1839-1847.
22. Byrd JW, Jones KS. Hip arthroscopy in the presence of dysplasia. *Arthroscopy* 2003;19:1055-1060.
23. Martin RL, Kelly BT, Philippon MJ. Evidence of validity for the hip outcome score. *Arthroscopy* 2006;22:1304-1311.
24. Martin RL, Philippon MJ. Evidence of validity for the hip outcome score in hip arthroscopy. *Arthroscopy* 2007;23:822-826.
25. Martin RL, Philippon MJ. Evidence of reliability and responsiveness for the hip outcome score. *Arthroscopy* 2008;24:676-682.

26. Thorborg K, Roos EM, Bartels EM, Petersen J, Holmich P. Validity, reliability and responsiveness of patient-reported outcome questionnaires when assessing hip and groin disability: A systematic review. *Br J Sports Med* 2010;44:1186-1196.
27. Hinman RS, Dobson F, Takla A, O'Donnell J, Bennell KL. Which is the most useful patient-reported outcome in femoroacetabular impingement? Test-retest reliability of six questionnaires. *Br J Sports Med* 2014;48:458-463.
28. Kemp JL, Collins NJ, Roos EM, Crossley KM. Psychometric properties of patient-reported outcome measures for hip arthroscopic surgery. *Am J Sports Med* 2013;41:2065-2073.
29. Klassbo M, Larsson E, Mannevik E. Hip disability and osteoarthritis outcome score. An extension of the Western Ontario and McMaster Universities Osteoarthritis Index. *Scand J Rheumatol* 2003;32:46-51.
30. Thorborg K, Holmich P, Christensen R, Petersen J, Roos EM. The Copenhagen Hip and Groin Outcome Score (HAGOS): Development and validation according to the COSMIN checklist. *Br J Sports Med* 2011;45:478-491.
31. Mohtadi NG, Griffin DR, Pedersen ME, et al. The development and validation of a self-administered quality-of-life outcome measure for young, active patients with symptomatic hip disease: The International Hip Outcome Tool (iHOT-33). *Arthroscopy* 2012;28:595-605; quiz 6-10.e1.
32. Griffin DR, Parsons N, Mohtadi NG, Safran MR. A short version of the International Hip Outcome Tool (iHOT-12) for use in routine clinical practice. *Arthroscopy* 2012;28:611-616; quiz 6-8.
33. Christensen CP, Althausen PL, Mittleman MA, Lee JA, McCarthy JC. The nonarthritic hip score: Reliable and validated. *Clin Orthop Relat Res* 2003;(406):75-83.
34. Muller B, Yabroudi MA, Lynch A, et al. Defining thresholds for the patient acceptable symptom state for the IKDC Subjective Knee Form and KOOS for patients who underwent ACL reconstruction. *Am J Sports Med* 2016;44:2820-2826.
35. Greco NJ, Anderson AF, Mann BJ, et al. Responsiveness of the International Knee Documentation Committee Subjective Knee Form in comparison to the Western Ontario and McMaster Universities Osteoarthritis Index, modified Cincinnati Knee Rating System, and Short Form 36 in patients with focal articular cartilage defects. *Am J Sports Med* 2010;38:891-902.
36. Monticone M, Ferrante S, Salvaderi S, Motta L, Cerri C. Responsiveness and minimal important changes for the knee injury and osteoarthritis outcome score in subjects undergoing rehabilitation after total knee arthroplasty. *Am J Phys Med Rehabil* 2013;92:864-870.
37. Collins NJ, Prinsen CA, Christensen R, Bartels EM, Terwee CB, Roos EM. Knee injury and osteoarthritis outcome score (KOOS): Systematic review and meta-analysis of measurement properties. *Osteoarthritis Cartilage* 2016;24:1317-1329.
38. Briggs KK, Lysholm J, Tegner Y, Rodkey WG, Kocher MS, Steadman JR. The reliability, validity, and responsiveness of the Lysholm score and Tegner activity scale for anterior cruciate ligament injuries of the knee: 25 years later. *Am J Sports Med* 2009;37:890-897.
39. Noyes FR, Barber SD, Mooar LA. A rationale for assessing sports activity levels and limitations in knee disorders. *Clin Orthop Relat Res* 1989;(246):238-249.
40. Davies AP. Rating systems for total knee replacement. *Knee* 2002;9:261-266.
41. Patt JC, Mauerhan DR. Outcomes research in total joint replacement: A critical review and commentary. *Am J Orthop (Belle Mead NJ)* 2005;34:167-172.
42. Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. Validation study of WOMAC: A health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. *J Rheumatol* 1988;15:1833-1840.
43. Kirkley A, Griffin S, McLintock H, Ng L. The development and evaluation of a disease-specific quality of life measurement tool for shoulder instability. The Western Ontario Shoulder Instability Index (WOSI). *Am J Sports Med* 1998;26:764-772.
44. Kirkley A, Griffin S, Dainty K. Scoring systems for the functional assessment of the shoulder. *Arthroscopy* 2003;19:1109-1120.
45. Kirkley A, Alvarez C, Griffin S. The development and evaluation of a disease-specific quality-of-life questionnaire for disorders of the rotator cuff: The Western Ontario Rotator Cuff Index. *Clin J Sport Med* 2003;13:84-92.
46. MacDermid JC, Drosdowech D, Faber K. Responsiveness of self-report scales in patients recovering from rotator cuff surgery. *J Shoulder Elbow Surg* 2006;15:407-414.
47. Tashjian RZ, Deloach J, Porucznik CA, Powell AP. Minimal clinically important differences (MCID) and patient acceptable symptomatic state (PASS) for visual analog scales (VAS) measuring pain in patients treated for rotator cuff disease. *J Shoulder Elbow Surg* 2009;18:927-932.
48. Lo IK, Griffin S, Kirkley A. The development of a disease-specific quality of life measurement tool for osteoarthritis of the shoulder: The Western Ontario Osteoarthritis of the Shoulder (WOOS) index. *Osteoarthritis Cartilage* 2001;9:771-778.
49. Kocher MS, Horan MP, Briggs KK, Richardson TR, O'Holleran J, Hawkins RJ. Reliability, validity, and responsiveness of the American Shoulder and Elbow Surgeons subjective shoulder scale in patients with shoulder instability, rotator cuff disease, and glenohumeral arthritis. *J Bone Joint Surg Am* 2005;87:2006-2011.
50. Roy JS, MacDermid JC, Woodhouse LJ. Measuring shoulder function: A systematic review of four questionnaires. *Arthritis Rheum* 2009;61:623-632.
51. Michener LA, McClure PW, Sennett BJ. American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form, patient self-report section: Reliability, validity, and responsiveness. *J Shoulder Elbow Surg* 2002;11:587-594.
52. Christiansen DH, Frost P, Falla D, Haahr JP, Frich LH, Svendsen SW. Responsiveness and minimal clinically important change: A comparison between 2 shoulder outcome measures. *J Orthop Sports Phys Ther* 2015;45:620-625.

53. Christie A, Dagfinrud H, Garratt AM, Ringen Osnes H, Hagen KB. Identification of shoulder-specific patient acceptable symptom state in patients with rheumatic diseases undergoing shoulder surgery. *J Hand Ther* 2011;24:53-60. quiz 1.
54. van Kampen DA, Willems WJ, van Beers LW, Castelein RM, Scholtes VA, Terwee CB. Determination and comparison of the smallest detectable change (SDC) and the minimal important change (MIC) of four-shoulder patient-reported outcome measures (PROMs). *J Orthop Surg Res* 2013;8:40.
55. Hudak PL, Amadio PC, Bombardier C. Development of an upper extremity outcome measure: the DASH (disabilities of the arm, shoulder and hand) [corrected]. The Upper Extremity Collaborative Group (UECG). *Am J Ind Med* 1996;29:602-608.
56. Paul A, Lewis M, Shadforth MF, Croft PR, Van Der Windt DA, Hay EM. A comparison of four shoulder-specific questionnaires in primary care. *Ann Rheum Dis* 2004;63:1293-1299.
57. Kim JK, Park ES. Comparative responsiveness and minimal clinically important differences for idiopathic ulnar impaction syndrome. *Clin Orthop Relat Res* 2013;471:1406-1411.
58. Eechaute C, Vaes P, Van Aerschot L, Asman S, Duquet W. The clinimetric qualities of patient-assessed instruments for measuring chronic ankle instability: A systematic review. *BMC Musculoskelet Disord* 2007;8:6.
59. Dawson J, Doll H, Coffey J, Jenkinson C. Responsiveness and minimally important change for the Manchester-Oxford foot questionnaire (MOXFQ) compared with AOFAS and SF-36 assessments following surgery for hallux valgus. *Osteoarthritis Cartilage* 2007;15:918-931.