Journal of Clinical Epidemiology 82 (2017) 128−136

Journal of Clinical Epidemiology

The minimal clinically important difference raised the significance of outcome effects above the statistical level, with methodological implications for future studies

Felix Angst[a,*], André Aeschlimann[a], Jules Angst[b]

[a]Rehabilitation Clinic ("RehaClinic"), Department of Research, Quellenstrasse 34, 5330 Bad Zurzach, Switzerland
[b]Department of Psychiatry, Psychotherapy and Psychosomatics, Psychiatric Hospital Burghölzli, University of Zurich, Lenggstrasse 31, 8008 Zurich, Switzerland

Accepted 30 November 2016; Published online 14 December 2016

Abstract

Objective: To illustrate and discuss current and proposed new concepts of effect size (ES) quantification and significance, with a focus on statistical and clinical/subjective interpretation and supported by empirical examples.

Study Design and Settings: Different methods for determining minimal clinically important differences (MCIDs) are reviewed, applied to practical examples (pain score differences in knee osteoarthritis), and further developed. Their characteristics, advantages, and disadvantages are illustrated and discussed.

Results: Empirical score differences between verum and placebo become statistically significant if sample sizes are sufficiently large. MCIDs, by contrast, are defined by patients' perceptions. MCIDs obtained by the most common "mean change method" can be expressed as absolute or relative scores, as different ES parameters, and as the optimal cutoff point on the receiver operating characteristic curve. They can further be modeled by linear and logistic regression, adjusting for potential confounders.

Conclusion: Absolute and relative MCIDs are easy to interpret and apply to data of investigative studies. MCIDs expressed as effect sizes reduce bias, which mainly results from dependency on the baseline score. Multivariate linear and logistic regression modeling further reduces bias. Anchor-based methods use clinical/subjective perception to define MCIDs and should be clearly differentiated from distribution-based methods that provide statistical significance only.

Keywords: Osteoarthritis; WOMAC; Outcome measurement; Effect size; Receiver operating characteristics curve; Statistics; Significance; Standardized mean difference; Minimal clinically important difference; Regression; Confounding

1. Introduction

Whether consciously or not, we are constantly measuring changes in health dimensions, such as pain, physical function, social function, or depression. Especially if affected by symptoms, we are alert to daily alterations in our state. In medicine, particularly in rheumatology, the last 30 years have seen considerable progress in the standardization of the measurements and methods of outcome effect quantification [1−4]. It should be possible for results that are qualitatively and quantitatively identical or very similar to mean the same across cultures and languages and to be similarly interpreted.

Human perceptions of most health dimensions are subjective and individual, posing an inherent problem for the standardization of assessment, interpretation, and comparison. Time, place, health state, and internal and external circumstances all affect perception and can lead to wide variations. Any changes measured may disappear in the noise of variability, hampering the use of conventional statistical, analytical methods.

Many studies, especially older pharmacological trials, confine themselves to quantifying the size and the significance of differences in health dimensions by conventional statistical methods, for example, by the *t*-test. However, statistically significant differences are mainly dependent on

Conflict of interest: None.

Funding: This study was supported by the Zurzach Rehabilitation Foundation SPA, Bad Zurzach, Switzerland.

* Corresponding author. Tel.: ++41-56-269-53-97; fax: ++41-56-269-51-72.

E-mail address: fangst@vtxmail.ch (F. Angst).

http://dx.doi.org/10.1016/j.jclinepi.2016.11.016

What is new?

Key findings

- This study provides an overview of the most important methods for determining minimal clinically important differences (MCIDs) and presents further developments.

What this adds to what is known?

- It illustrates the strengths and weaknesses of different MCID parameters and the relationship of MCIDs to statistically significant differences.

What is the implication and what should change now?

- Multivariate regression modeling of MCIDs may open up new prospects for less-biased estimates of MCIDs.

the number of persons examined (n), as will be demonstrated later [1,5–7].

Given these shortcomings, an alternative "significance" has been identified to characterize the sizes of effects [1]. Because medicine measures outcome in humans, who feel and communicate, the patient alone can define an identifiable difference. Responding to the relevant assessments, the patient can only assess an effect that is felt. This led to the development of the concept of the smallest subjectively perceptible effect that is "clinically" important, named the minimal clinically important difference (MCID) [1,8–10].

An instructive overview of the history, concepts, and characteristics of methods which estimate "clinical significance" is provided by Kamath et al. in their fundamental textbook "Methods and applications in clinical trials" [1]. "Anchor-based" methods use external criteria (the anchor) to quantify differences measured by an outcome instrument (example: the mean change method). "Distribution-based" methods define different statistical parameters to assess clinical significance (example: t-test). Kamath et al. include our previous investigation (2001) on statistically detectable differences and MCIDs as one of seven exemplary studies [5].

In this article, we will compare current concepts of effect size quantification and significance, with a focus on statistical and clinical/subjective meaning. Based on those models, new concepts for quantifying MCIDs will be developed, discussed, and illustrated by specific examples from empirical studies. The concepts of our earlier study are expanded and combined with new approaches which are particularly relevant for randomized controlled trials (RCTs).

2. Smallest detectable difference (SDD)—statistical definition for quantifying the significance of differences

In empirical outcome research, the t-test formula is used in many analyses of continuous parameters as a suitable approximation, although some of the necessary assumptions (normal distribution, homoscedasticity, etc.) are often not met by the data [5–7]. Choosing the simplest example of pairwise differences in one sample, for example, within-person differences between baseline and follow-up, the statistic to examine significance is [6] $t = \Delta/(s/\sqrt{n})$, where Δ = difference of the means = mean of the differences (baseline to follow-up), s = standard deviation of the differences, and $s/\sqrt{n}$ is the standard error of Δ, n = sample size.

To reach statistical significance, t has to be large. It can then be assumed with a low probability of error (type I error P) that the difference really exists. Δ and s are finite parameters, especially in closed scales, such as the visual analogue scale (VAS) for pain between 0 and 100 (mm). The sample size n may increase to high, almost infinite numbers. Therefore, t mainly grows by n, together with the probability of significance. A minimal Δ, whose corresponding t reaches a predefined significance level P, is defined as the smallest (statistically) detectable difference [5,6]. A common example of the principal application of the t-test is the measurement of effects in RCTs.

3. Standardized mean difference (SMD)—statistical parameter for quantifying differences in RCTs

The SMD is expressed as the difference Δ in the two mean score differences (baseline to follow-up) between the verum and placebo groups divided by the so-called "pooled" or "within" standard deviation s_Δ of the two groups [7].

$$\mathrm{SMD} = \frac{\Delta}{s_\Delta} \; \textit{and} \; s_\Delta = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Thus s_Δ is the square root of the mean of the variances (of the score differences), weighted by the number of subjects, of the verum group (index 1) and the placebo group (index 2). Where sample sizes are equal ($n_1 = n_2$), the pooled variance ${s_\Delta}^2$ is simply the mean of the two variances.

In other words, the SMD is the difference in mean pain relief between verum and placebo in number of pooled standard deviations and is dimensionless. Positive SMDs reflect the superiority of the verum, negative SMDs the superiority of the placebo. The larger the SMD, the greater the probability of attaining statistical significance to support the conclusion that verum is more effective than placebo. Nowadays, the SMD is the standard effect size parameter for RCTs [7]. In meta-analyses, the SMDs of different RCTs are themselves pooled to give a global effect size.

The 95% confidence interval (95% CI) of the SMD based on the standard error (se) is given by

$\mathrm{se(SMD)} = \sqrt{\frac{n_1+n_2}{n_1 n_2} + \frac{\mathrm{SMD}^2}{2(n_1+n_2)}}$. The 95% CI of the SMD is then [SMD $- t \times$ se(SMD), SMD + t × se(SMD)], and t is the t-value of the t-distribution with $(n_1 + n_2 - 2)$ degrees of freedom and the probability $(1 - 0.05/2)$ or one minus half of the type I error. In empirical studies with small sample sizes, the t-distribution has to be used. It approximates to the normal distribution with increasing n by the central limit theorem [6]. Thus, many authors use $t = z = 1.96$, that is, the value of the normal distribution, which is equal to t with infinite degrees of freedom.

The SMD and SE(SMD) have to be estimated on the basis of the empirical data, which leads to a correction, that is, multiplication by a correction factor J [7]: $J = 1 - \frac{3}{4(n_1+n_2-2)-1}$. J corrects the overestimation of the SMD and its standard error to avoid (a small) bias when the sample sizes are small [7]. Examples of different Js are shown in Table 1.

The formula for SE(SMD) clearly reveals that with an increase in the sample sizes n_1 and n_2, the term $n_1 + n_2$ grows moderately compared to the much faster growth of the term $n_1 n_2$. Thus, SE(SMD) diminishes with increasing sample sizes. The SE(SMD) becomes then increasingly dependent on the sample sizes if the SMD (and Δ) is constant.

This can be illustrated by a hypothetical example of pain relief (Table 1). On the VAS pain scale from 0 = no pain to 100 = maximum pain, the verum is expected to relieve pain by a mean of 15 score points, placebo by 10. The mean difference $\Delta = 15 - 10 = 5$ is then constant in all examples of Table 1. Both groups have standard deviations of 10 points, that is, the pooled s_Δ is also 10. With growing sample sizes, the 95% CIs of the SMDs diminish together with P. The constant effect difference SMD becomes statistically significant with 33 subjects at a level of $P < 0.050$ (exact: 0.0496) and highly significant ($P = 0.001$) for 100 test subjects per group.

4. MCID—"clinical" definition for quantifying the significance of differences

The concept was originated by Jaeschke et al. 1989 [8]: "The MCID can be defined as the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive costs, a change in the patient's management." At the final follow-up, patients were asked to globally rate the changes perceived in comparison to baseline. This "transition" item offered seven response options: (1) "almost the same, hardly any better at all", (2) "a little better", (3) somewhat better" (4) "moderately better", (5) "a good deal better", (6) "a great deal better", (7) "a very great deal better", and analogous options for "worse" [7]. The MCID was defined as the score differences (between baseline and follow-up) of the patients who rated the transition item as 1, 2, or 3.

In rheumatology, the concept was further developed by Redelmeier, who in 1993 introduced the specific self-assessment transition question that is currently most frequently used [9]: "How is your health today (at follow-up) compared to baseline?" This has five Likert response options: "much better", "slightly (or somewhat) better", "about the same", "slightly (or somewhat) worse", "much worse" (for an example: see Table 2). For the transition response categories, score differences (baseline to follow-up) measured by an outcome instrument were then compared. "Specifically, we classified the difference in perceptions reported as being about the same and being somewhat better as an important symptomatic difference" [9]. Thus, the MCID for improvement is equal to the mean score difference (baseline to follow-up) of the "slightly better" minus the "about the same" groups, and the MCID for worsening equal to that of the "slightly worse" minus the "about the same" groups. The purpose is "to adjust for possible bias in ratings, …" [9].

This concept is known as the "mean change method" [10]. The absolute difference between the two means of score differences is scaled positively to reflect an MCID for improvement (e.g., pain relief) and negatively to reflect an MCID for worsening. The MCIDs thus determined are valid for application at the group but not at the individual level.

Table 1. Dependence of the statistical significance on sample size (n_1, n_2) with a constant effect difference of 15 − 10 = 5 score points

Verum VAS difference			Placebo VAS difference			Boren-stein's			t-test
n_1	m_1	s_1	n_2	m_2	s_2	J	SMD	95%-CI	P
10	15	10	10	10	10	0.9577	0.479	−0.434, 1.392	0.285
20	15	10	20	10	10	0.9801	0.490	−0.147, 1.127	0.128
30	15	10	30	10	10	0.9870	0.494	−0.024, 1.011	0.061
32	15	10	32	10	10	0.9879	0.494	−0.007, 0.995	0.053
33	15	10	33	10	10	0.9882	0.494	0.001, 0.987	<0.050
34	15	10	34	10	10	0.9886	0.494	0.008, 0.980	0.046
50	15	10	50	10	10	0.9923	0.496	0.096, 0.896	0.016
100	15	10	100	10	10	0.9962	0.498	0.216, 0.780	0.001

Abbreviations: VAS, visual analogue scale of pain (scale: 0–100 score points), difference between baseline and follow-up; *n*, sample size; *m*, mean; *s*, standard deviation; *J*, correction factor by Borenstein; SMD, standardized mean difference; 95% CI, 95% confidence interval; *P*, statistical significance (type I error) that SMD is different from zero.

Table 2. Data from the evaluation study [11]: knee osteoarthritis before inpatient rehabilitation (baseline) and 3 months later (follow-up) on the WOMAC pain scale

Transition item	WOMAC pain: difference baseline to follow-up		
Pain at follow-up was	*n*	*m*	*s*
Much better	19	31.58	17.63
Slightly better	49	13.51	21.58
About the same	62	4.77	15.62
Slightly worse	44	−1.32	20.81
Much worse	16	−3.50	17.03
All: score difference	190	7.60	21.13
All: baseline score	190	50.93	21.48

Abbreviations: WOMAC, Western Ontario and McMaster Universities questionnaire; *n*, number of subjects; *m*, mean; *s*, standard deviation, both for the score differences (baseline to follow-up).

WOMAC pain scaling: 0 = maximal pain, 100 = no pain. A positive difference reflects pain relief and vice versa.

5. Parameters for quantifying MCIDs

5.1. Evaluation study

The purpose of an evaluation study is to quantify the MCID. In the interests of simplicity all concepts, demonstrated parameters and examples are limited to the MCID for improvement in the following analysis. The MCID for deterioration can be analogously determined, for example, by the mean change method, that is, the difference between the mean score difference of the "slightly worse" group and that of the "about the same" group [11,12].

To illustrate the different methods of quantifying MCIDs, we present unpublished results of a reanalysis of our "Zurzach Osteoarthritis Study," a longitudinal cohort study of hip and knee patients observed before and after a standardized 3-week rehabilitation intervention [11]. That study, published in 2011, is based on the data of a first report [12] enriched by data of new patients.

The findings from the complete study are presented in Table 2. This shows the score differences between baseline = start of rehabilitation and the follow-up 3 months later of $n = 190$ knee osteoarthritis patients. Patients assessed pain (5 items), stiffness (2 items), and function (17 items) on the Western Ontario and McMaster Universities questionnaire (WOMAC), one of the most thoroughly tested and widely used self-rating instruments for hip and knee osteoarthritis [4]. The five pain items are scaled in integers from 0 = no pain to 10 = maximum pain (numeric rating scale). The arithmetic mean of at least three of five valid items (missing rule for the WOMAC) multiplied by 10 results in the WOMAC pain scale (0 = no pain to 100 = maximum pain). As in many studies, the scale is transformed into a scale from 0 = maximum pain to 100 = no pain by subtracting the original score from 100 (100 − score). This is also the scaling method used in the most widely used outcome instrument, the Short Form 36 Health Survey; see examples in [11,12].

In our study [11], mean WOMAC pain relief was 13.51 points (standard deviation = 21.58) in the "slightly better" group ($n = 49$) and 4.77 (standard deviation = 15.62) in the "about the same group" ($n = 62$) between baseline and follow-up in response to the transition question "how is your pain today (=follow-up) compared to baseline?" (Table 2). Based on these data, different parameters can now be calculated to quantify the MCID for improvement (Tables 3 and 4).

5.2. Absolute MCID

The MCID according to Jaeschke was 13.51 and that according to Redelmeier 8.74 score points (Table 3). The Redelmeier MCID, the mean change method, is the method most commonly found in the literature [9–15]. It is easy to interpret and to apply in an investigative study if the scaling is the same

Table 3. MCID for improvement on the WOMAC pain scale: absolute, relative, and effect sizes

Method	Numerator	Denominator	MCID	95%-CI	*P*	Comment
Jaeschke [8]	13.51	—	13.51	7.25, 19.77	<0.001	Score difference of the "slightly better" group
Mean change meth. Redelmeier [9]	13.51–4.77	—	8.74	1.73, 15.74	0.015	Score difference of the "slightly better" group minus that of the "about the same" group
% baseline score	8.74	50.93	17.15%	6.86%, 27.39%	0.015	Mean change in % of the baseline score
% total score	8.74	100.00	8.74%	3.49%, 13.95%	0.015	Mean change in % of the maximal score
ES, Kazis [17]	8.74	21.48	0.407	0.024, 0.789	0.038	Mean change divided by the standard deviation of the group's baseline score
SRM, Liang [18]	8.74	21.13	0.413	0.031, 0.796	0.035	Mean change divided by the standard deviation of the group's score differences
SMD, Borenstein [7]	8.74*0.993	18.48	0.469	0.092, 0.847	0.016	Mean change*J divided by the pooled standard deviation of the two transition group's score differences

Abbreviations: MCID, minimal clinically important difference scaled in score points (scale: 0–100); WOMAC, Western Ontario and McMaster Universities questionnaire; ES, effect size (according to Kazis); SRM, standardized response mean (according to Liang); SMD, standardized mean difference (according to Borenstein); ES, SRM, SMD, dimensionless (scaled by number of standard deviations); 95% CI, 95% confidence interval; *P*, type I error of the test that the MCID is different from zero.

Table 4. MCID for improvement on the WOMAC pain scale: ROC and regression methods

Method	Dependent variable	Independent variables	MCID	95%-CI	*P*	Comment
ROC	Transition	Δ pain	15.00	8.74, 21.26	<0.001	Area under ROC: 0.637 (95% CI: 0.528 –0.747), sensitivity = 0.531, specificity = 0.871
Linear regression: bivariate	Δ pain	Transition	8.74	1.73, 15.74	0.015	Same result as by the mean change method
Linear regression: multivariate	Δ pain	Transition, sex, age, WOMAC pain baseline score	7.09	0.93, 13.25	0.024	Adjusted for the added potential confounders (independent variables)
Logistic regression: bivariate	Transition	Δ pain	OR: 1.026 (beta: 0.0261, se = 0.0110)	1.004, 1.049	0.018	Odds ratio: probability of being "slightly better" for 1.00 point pain relief
Logistic regression: multivariate	Transition	Δ pain, sex, age, WOMAC pain baseline score	OR: 1.029 (beta: 0.0286, se = 0.0128)	1.004, 1.055	0.025	Odds ratio, adjusted for the added potential confounders (independent variables)

Abbreviations: MCID, minimal clinically important difference scaled in score points (scale: 0–100); WOMAC, Western Ontario and McMaster Universities questionnaire; 95% CI, 95% confidence interval; *P*, type I error of the test that the MCID is larger than zero; ROC, receiver operating characteristic curve; Δ pain, WOMAC pain score difference baseline to follow-up; Transition item response, 0 = about the same, 1 = slightly better; Logistic regression, OR = odds ratio, beta = regression coefficient for Δ pain, se = standard error.

as that of the evaluation study. If the baseline score is close to the end of the scale, the size of the absolute MCID is limited and tends to move toward the scale mean according to the regression-to-the-mean phenomenon.

This phenomenon is observed in empirical data, with follow-up scores tending to be located toward the mean/ the middle of a scale because individual baseline scores near the end of a closed scale can change to a limited extent only and are more likely to move to more moderate levels than to more extreme levels. For example, on the closed scale from 0 to 100 = no pain, a pain baseline score of 95 can improve by a maximum of 5 points to 100 at the follow-up, whereas it can deteriorate by maximum of 95 points to 0. A severely affected patient is more likely to experience larger improvements under treatment than one with almost no pain. In any event, the absolute MCID depends on the baseline score [12,14,16], a phenomenon observed in almost every empirical outcome study.

5.3. Relative MCID

The absolute MCID can thus be related to the baseline score (mean of the whole group: 50.93, see Table 2) or to the maximum possible score: 17.15% or 8.74% (Table 3). These relative MCIDs are fairly easy to interpret and to transfer to the data of other studies. However, toward the end of the closed scale, this method again raises problems. For example, the same absolute MCID for improvement would be 8.74/10 = 87.4% at a baseline score of 10 and 8.74/90 = 9.7% at a baseline score of

Table 5. Application of an a priori evaluated MCID [11] to an RCT [19]

		Baseline		Follow-up		Difference	Pooled
WOMAC pain	*n*	*m*	*s*	*m*	*s*	*m*	*s*
Intervention	36	59.6	15.8	71.4	15.8	11.8	15.8
Placebo	35	60.2	17.0	60.4	21.6	0.2	19.4
Total	71	59.9	16.4	66.0	18.9	6.1	17.7

	Δ	*s*	Parameter	95%-CI	*P*
Empiric SMD	(11.8–0.2)*0.989	17.7	0.649	0.168, 1.129	0.008
MCID as SMD	8.74*0.989	17.7	0.489	0.013, 0.964	0.044
MCID as SRM	8.74	17.7	0.494	0.013, 0.975	0.044
MCID as ES	8.74	16.4	0.533	0.051, 1.015	0.031
Logistic regr.	exp(0.0261*11.6)	se = 0.0110	OR = 1.354	1.049, 1.746	0.021

Abbreviations: WOMAC pain, 0 = maximal pain, 100 = no pain; *n*, number of patients; *m*, mean; *s*, standard deviation; **Δ**, relevant difference of score differences; MCID, minimal clinically important difference (positively scaled to reflect improvement); ES, effect size according to Kazis; SRM, standardized response mean according to Liang; SMD, standardized mean difference according to Borenstein; 95% CI, 95% confidence interval; *P*, type I error of the test that the effect size is different from zero; OR, odds ratio; se, standard error.

90. Alternatively, the absolute MCID related to the maximum possible score, that is, 8.74/100 = 8.74% provides a constant level of absolute and relative MCID. Because the relative MCID remains dependent on the baseline score, the phenomenon of regression-to-the-mean persists.

5.4. MCID as effect size

Such problems of bias are reduced by quantification methods using standardized effect size parameters, a new approach which also improves the comparability of MCIDs across different studies. "Effect size" is also an umbrella term for all relative and dimensionless effect parameters [7]. The absolute MCID can be standardized into effect sizes, for example, by (1) the effect size (ES) according to Kazis (division by the baseline standard deviation of the whole sample: 21.48, see Table 2); (2) the standardized response mean (SRM) according to Liang (division by the standard deviation of the score differences of the whole sample: 21.13, see Table 2); or (3) the SMD according to Borenstein between the "slightly better" and "almost equal" transition items [7,17,18]: ES = 0.407, SRM = 0.413, SMD = 0.469 (Table 3). The Kazis ES and the Liang SRM are especially appropriate for paired data measuring intra-individual differences [7,17,18]. At the ends of the scale, the score differences (and the absolute MCID) tend to become smaller, together with the standard deviations. For that reason, the estimate of the MCID tends to be more constant and less biased over the scale range.

To our knowledge, no study so far has expressed the MCID as a SMD; the SMD method uses the (smaller) standard deviations of the transition category groups. Thus, the evaluated MCID expressed as the SMD results in higher figures than those obtained by the ES and SRM methods (see Table 3). It tends to overestimate the clinically relevant effect because the two transition groups tend to show homogeneous effects. This problem disappears, however, in practical application, as the specific example of Table 5 demonstrates [11,19].

5.5. MCID by receiver operating characteristic curve

Finally, the transition responses can be examined by the receiver operating characteristic (ROC) curve [14,20,21]. This analyzes the performance of the classification into the categories "somewhat better" and "about the same" on the basis of the individual score differences (baseline to follow-up). The ROC curve determines the optimal cutoff point, where the sum of the false positive (1 − sensitivity) and the false negative (1 − specificity) is smallest, that is, where the sum of sensitivity + specificity is maximized [14,21]: in our example, the MCID = 15.00 score points with 53.1% sensitivity and 87.1% specificity (Table 4).

5.6. MCID by linear regression

As another new quantification method, the score difference between baseline and follow-up can be further modeled on the basis of the transition item (independent variable), using bivariate linear regression. The transition item is coded as 1 for "slightly better" and 0 for "almost the same." In this simplest model, the linear coefficient for the transition item is identical to the MCID for improvement obtained by the mean change method: 8.74 score points (Table 4).

This model can then be extended by covariables that may have an influence on the dependent pain variable (confounding) and may not be equally distributed between the transition categories, for example, sex, age, or WOMAC pain baseline score. In the example, this MCID for improvement was 7.09 score points. The most important confounder was the WOMAC pain baseline score ($P < 0.001$ in the model).

5.7. MCID—increased probability by logistic regression

A novel alternative approach to the MCID is provided by logistic regression using the transition responses (0 = almost equal and 1 = slightly better) as the dependent variable, and the (WOMAC pain) score difference and further possible confounders as independent variables. The resulting odds ratios (ORs) reflect the relative probabilities of being in the "slightly better" rather than in the "about the same" group [22]. This method tests whether the WOMAC pain score difference is a statistically significant predictor for allocation to the "slightly better" as compared to "almost equal" group.

In our example, the OR was 1.026 (beta: 0.0261) in the bivariate model and 1.029 (beta: 0.0286) in the multivariate model (Table 4). This means that the probability of being "slightly better" was +2.6% (bivariate) and +2.9% (multivariate) greater for a +1.00 increase in the WOMAC pain difference (pain relief). Using the beta coefficient, the increased probability of being slightly better can be determined for every score difference and for the MCID. For example, taking the MCIDs of the linear models (8.74 and 7.09 score points), the increased probability of being "slightly better" was 25.6% (bivariate) and 22.5% (multivariate), as calculated by exp(beta*MCID) [22].

6. Application of the evaluation study MCID to the investigative study

The empirically determined MCID for improvement can now help to quantify the effects of other studies, for example, RCTs that examine treatment effects. In an exemplary RCT of Hinman et al., hip and knee osteoarthritis patients were treated by aquatic-based physiotherapy (verum or intervention: WOMAC pain score difference = 11.8) or "usual daily activities and medication" (control: WOMAC pain score difference = 0.2) [19]. The observed SMD for WOMAC pain relief was 11.6 × 0.989/

17.7 = 0.649, since $\Delta = 11.8 - 0.2 = 11.6$, $s_\Delta = 17.7$, and J = 0.989 (Table 5).

Application of the MCID for improvement (8.74 score points) from our study [11] results in an MCID expressed as SMD = 8.74 × 0.989/17.7 = 0.489 for the RCT. Hence, the observed effect in favor of aquatic exercise (SMD = 0.649) was not only statistically significant but also clinically important, that is, on average, subjectively perceptible. Applying logistic regression, we found an increased probability of +35.4% of feeling "slightly better" with the verum treatment (Tables 4 and 5).

7. Alternative anchors and terminology

Besides the subjective assessment of an effect by means of the transition item, other "clinically" relevant parameters may function as anchors, such as the varyingly objective measures obtained by an analytical apparatus, or external rating by the treating health professional. For example, in rheumatoid arthritis, the disappearance of CCP antibodies in the serum under pharmacological treatment can be related to outcome parameters such as pain or function.

The different anchoring possibilities have prompted some authors to propose renaming the MCID the minimal important difference (MID), as "clinical" does not accurately reflect the subjective rating by the transition item [23,24]. Removing the clinical reference, however, opens the MID to all kinds of interpretation of importance, including the statistical one.

8. Pitfalls and problems

The enthusiasm for using the transition rating as an anchor to give health changes "clinical" meaning is shared by many experts in the field of psychometrics and clinimetrics [25,26]. The resulting MCIDs are often confused with statistically based methods, where, for example, the standard error of measurement defines clinical importance [20,24,27−29]. This, however, merely produces SDDs. For example, a difference of 2.00 or 1.96 standard errors or 1.96 × $\sqrt{2}$ × standard error of measurement has been considered to be clinically relevant [20,27,29]. The "minimal detectable change with 95% confidence" is in fact a difference beyond the 95% confidence interval and consequently an SDD [4]. However, MCIDs smaller than a difference that the instrument can measure with accuracy [4,27,29]—which depend on validity and reliability—do not make sense.

MCIDs and SDDs are often confused [21,27−29] although distribution-based SDDs and anchor-based MCIDs are clearly differentiated in modern reviews [4,13,29]. Differences between the SDD, MDC95%, and absolute and relative MCIDs have been illustrated in five different studies [21]. A good overview and interpretation of the different concepts, including those of Kazis (ES) and Liang (SRM), is given by Revicki et al. [13].

Although many studies apply the original concept according to Redelmeier, that is, they reduce the score difference of the "slightly better" group by that of the "about the same" group, others prefer to quantify the MCID (for improvement) by the score difference of the "slightly better" group only, as described by Jaeschke [15,16,24]. This produces different results and can lead to confusion.

ROC analysis has been used in several studies to determine MCIDs [13,14,16,20,21,24]. Because the ROC curve is designed to identify the score difference that optimally separates and specifies "better" from "equal," the effect is in fact maximized to an "optimal clinically important difference," to the detriment of the concept of the "minimal CID." Moreover, the MCID determined by the ROC is a single point estimate of an empirical sample and sensitive to stochastic instability [10]. The ROC curve plots single points (score differences) of the population examined in the evaluation study. On this curve, the "optimal" point, that is, the score difference which separates "slightly better" from "almost equal" by maximizing the sum sensitivity + specificity is then taken as the MCID for improvement, independently of the sample size. This means that a single point representing one individual on the curve of the whole group determines a difference which will then be generalized to another sample, that is, used on the group level [10]. In our example, the MCID = 15.00 obtained is far from the results of the other methods. As further illustration of the methodological problems inherent in the ROC analysis, many studies also add the score differences of the "much better" to those of the "somewhat better" or compare the data of all who rated "better" to all who rated "equal or worse" [14,16,24].

The MCID based on the mean change method and expressed as SMD might therefore be the most appropriate concept for RCTs since that is the standard method for expressing effect differences in such trials. Finally, it remains a subject of methodological and philosophical debate whether an MCID determined in an evaluation study can be validly applied and the results generalized to a second investigative study, for example, an RCT.

9. The size and generalizability of the MCID

A review of various studies reveals an MCID in the range of 6−10% of the total score, consistent with the results of the present study [3−5,8,9,11,12,15,19,20]. This corresponds to an effect size (ES, SRM, SMD) of 0.30−0.50 (standard deviations) [13]. These levels facilitate basic assumptions and the classification of the assessment scores of empirical study effects. Bearing in mind the reservations regarding generalizability, they can be used as an approximate estimate for an MCID if an evaluation study is not available for reasons of cost, time, or other constraints. This effect size range can be used as a general estimate for different questionnaires and settings and applied across different studies (observational cohort studies, RCTs, meta-analyses) and outcome instruments.

10. Conclusion

In the quantification of effects, it is crucial to distinguish ‘‘clinical’’ or rather ‘‘subjective’’ significance from statistical significance. The final aim remains that of giving effects a different significance from the purely statistical one: ‘‘It is recommended that the patient’s perspective be given the most weight, because these are patient-related outcome measures, although the clinician’s perspective is considered important as well’’ [13]. It might indeed be more appropriate to talk of ‘‘minimal subjectively perceptible (important) differences’’ (MSPD) to avoid misunderstanding and to remain close to the original concepts of Jaeschke and Redelmeier [8,9,13]. MCIDs help to add further significance to statistically significant effects [10–12,19]. Despite this awareness, important current reviews still argue in terms of SDDs alone [30].

MCIDs depend on various factors, such as the underlying health disorder, the health dimension being measured, its baseline score, the time between baseline and follow-up (‘‘recall bias’’), and the intervention [3,4,10,12,13,15,20,24]. An MCID should ideally be estimated by an evaluation study similar in its preconditions to an investigative study that determines effects. As a crude guideline, effect sizes between 0.30 and 0.50 are considered to be minimally clinically/subjectively important. Determination of the MCID also plays a role in the assessment of sensitivity to change, that is, responsiveness [31].

Regression modeling as an extension of the mean change method opens up new horizons and has never before been presented in this form in the literature. The resulting MCID is adjusted or ‘‘controlled’’ by various confounding factors and is therefore less biased. This form of analysis is especially appropriate for observational cohort studies comparing two groups with different levels and statistical distributions of the characterizing parameter, which may have a confounding influence on the MCID.

Acknowledgments

The authors thank Ms. Joy Buchanan and Ms. Elizabeth Kyrke for the English editing.

References

[1] Kamath CC, Sloan JA, Cappelleri C. Clinical significance. In: BalakrishnanMethods and applications of statistics in clinical trials: Concepts principles, trials, and designs, Vol 1. Hoboken, New Jersey, USA: John Wiley & sons Inc; 2014:170–90. 15.

[2] Streiner DL, Norman GR, editors. Health measurement scales. 4th ed. Oxford, UK: Oxford University Press; 2008.

[3] Angst F, Stucki G, Aeschlimann A. Quality of life assessment in osteoarthritis. Expert Rev Pharmacoecon Outcomes Res 2003;3(5):623–36.

[4] Katz PP. Patient outcomes in rheumatology, 2011. A review of measures. Arthritis Care Res (hoboken) 2011;63 Suppl 11:S1–490.

[5] Angst F, Aeschlimann A, Stucki G. Smallest detectable and minimal clinically important differences of rehabilitation intervention with their implications for required sample sizes using WOMAC and SF-36 quality of life measurement instruments in patients with osteoarthritis of the lower extremities. Arthritis Rheum 2001;45:384–91.

[6] Bland M. Comparing the means of small samples (t-test). In: . 4th ed.. In: Bland M, editor. An introduction to medical statistics, 10 Oxford, UK: Oxford University press; 2015:131–41.

[7] Borenstein M. Effect sizes for continuous data. In: . 2nd ed.. In: Cooper H, Hedges LV, Valentine JC, editors. The Handbook of research synthesis and meta-analysis, 12 New York: Russell Sage Foundation; 2009:222–36.

[8] Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Controlled Clin Trials 1989;10:407–15.

[9] Redelmeier DA, Lorig K. Assessing the clinical importance of symptomatic improvements. An illustration in rheumatology. Arch Intern Med 1993;153:1337–42.

[10] Angst F. The minimal clinically important difference assigns significance to outcome effects. J Rheumatol 2016;43:258–9.

[11] Angst F, Verra ML, Lehmann S, Benz T, Aeschlimann A. Effects of inpatient rehabilitation in hip and knee osteoarthritis. A naturalistic prospective cohort study with intra-individual control of effects. Arch Phys Med Rehabil 2013;94:2139–45.

[12] Angst F, Aeschlimann A, Michel BA, Stucki G. Minimal clinically important rehabilitation effects in patients with osteoarthritis of the lower extremities. J Rheumatol 2002;29:131–8.

[13] Revicki D, Hays RD, Cella DE, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol 2008;61:102–9.

[14] Mills K, Naylor J, Eyles J, Roos E, Hunter D. Examining the minimal important difference of patient reported outcome measures for individuals with knee osteoarthritis: a model using the knee injury and osteoarthritis outcome score. J Rheumatol 2016;43:395–404.

[15] Bellamy N, Hochberg M, Tubach F, Martin-Mola E, Awanda H, Bombardier C, et al. Development of multinational definitions of minimal clinically important improvement and patient acceptable symptomatic state in osteoarthritis. Arthritis Care Res (Hoboken) 2015;67(7):972–80.

[16] Escobar A, Perez LG, Herrera-Espineira C, Aizpuru F, Sarasqueta C, Gonzalez M, et al. Total knee replacement; minimal clinically important differences and responders. Osteoarthritis Cartil 2013;21:2006–12.

[17] Kazis ES, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med Care 1989;27:S178–89.

[18] Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. Med Care 1990;28:632–42.

[19] Hinman RS, Heywood SE, Day AR. Aquatic physical therapy for hip and knee osteoarthritis: results of a single-blind randomized controlled trial. Phys Ther 2007;87:32–43.

[20] 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;14:109–14.

[21] De Vet HCW, Terluin B, Knol DL, Roorda LD, Mokkink LB, Ostelo RWJG, et al. Three ways to quantify uncertainty in

individually applied ‘‘minimally important change’’ values. J Clin Epidemiol 2010;63:37−45.

[22] Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression. 3rd ed. New York, USA: John Wiley & Sons; 2013.

[23] Schuenemann HJ, Guyatt GH. Commentary-goodbye M(C)ID! Hello MID, where do you come from? Health Serv Res 2005;40:593−7.

[24] Terwee CB, Roorda LD, Dekker J, Bierma-Zeinstra SM, Peat G, Jordan KP, et al. Mind the MIC: large variation among populations and methods. J Clin Epidemiol 2010;63:524−34.

[25] Fischer D, Steward AI, Bloch DA, Loring K, Laurent D, Holman H. Capturing the patient’s view of change as a clinical outcome measure. JAMA 1999;282:1157−62.

[26] Guyatt GH, Norman GR, Juniper EF, Griffith LE. A critical look at transition ratings. J Clin Epidemiol 2002;55:900−8.

[27] Wyrwich KW. Minimal important difference thresholds and the standard error of measurement: is there a connection? J Biopharmaceut Stat 2004;14(1):97−110.

[28] King MT. A point of minimal important difference (MID): a critique of terminology and methods. Expert Rev Pharmacoeconomics Outcomes Res 2011;11(2):171−84.

[29] 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.

[30] Bannuru RR, Schmid CH, Kent DM, Vaysbrot EE, Wong JB, McAllindon TE. Comparative effectiveness of pharmacologic interventions for knee osteoarthritis. A systematic review and network analysis. Ann Intern Med 2015;162:46−54.

[31] Stratford PW, Riddle DL. Assessing sensitivity to change: choosing the appropriate change coefficient. Health Qual Life Outcomes 2005;3:25.