Journal of Clinical Epidemiology 68 (2015) 1388–1396

Journal of Clinical Epidemiology

ORIGINAL ARTICLES

Minimal important change (MIC) based on a predictive modeling approach was more precise than MIC based on ROC analysis

Berend Terluin[a,*], Iris Eekhout[b], Caroline B. Terwee[b], Henrica C.W. de Vet[b]

[a]*Department of General Practice and Elderly Care Medicine, EMGO Institute for Health and Care Research, VU University Medical Center, Van der Boechorststraat 7, 1081 BT Amsterdam, The Netherlands*

[b]*Department of Epidemiology and Biostatistics, EMGO Institute for Health and Care Research, VU University Medical Center, Van der Boechorststraat 7, 1081 BT Amsterdam, The Netherlands*

Accepted 23 March 2015; Published online 28 March 2015

Abstract

Objectives: To present a new method to estimate a ''minimal important change'' (MIC) of health-related quality of life (HRQOL) scales, based on predictive modeling, and to compare its performance with the MIC based on receiver operating characteristic (ROC) analysis. To illustrate how the new method deals with variables that modify the MIC across subgroups.

Study Design and Setting: The new method uses logistic regression analysis and identifies the change score associated with a likelihood ratio of 1 as the MIC. Simulation studies were conducted to investigate under which distributional circumstances both methods produce concordant or discordant results and whether the methods differ in accuracy and precision.

Results: The ''predictive MIC'' and the ROC-based MIC were identical when the variances of the change scores in the improved and not-improved groups were equal and the distributions were normal or oppositely skewed. The predictive MIC turned out to be more precise than the ROC-based MIC. The predictive MIC allowed for the testing and estimation of modifying factors such as baseline severity.

Conclusion: In many situations, the newly described MIC based on predictive modeling yields the same value as the ROC-based MIC but with significantly greater precision. This advantage translates to increased statistical power in MIC studies.

Keywords: Change scores; Health-related quality of life; Minimal important change; ROC method; Predictive modeling; Likelihood ratio

1. Introduction

This article introduces a new method to estimate a ''minimal important change'' (MIC) of health-related quality of life (HRQOL) scales, which has some advantages over current methods. Ever since the introduction of HRQOL scales in research and clinical practice, investigators and clinicians face the challenge of making sense of changes in HRQOL scores [1]. It has been widely recognized that mean HRQOL changes may well reach statistical significance, whereas at the same time, the clinical relevance might be limited, if not completely absent. The ''minimal important change'' [MIC; also called ''minimal clinically important change'' or ''minimal (clinically) important difference''], loosely defined as the minimal amount of change in an HRQOL score that is worthwhile and perceived as ''important'' either by the patient or by a third party (e.g., the clinician), is an instantly appealing concept (see for a recent review: [2]). Various methods have been developed to determine MICs [3,4], which can generally be distinguished in two broad groups: anchor-based and distribution-based methods [3]. Distribution-based MICs are based on the distribution of HRQOL scores in various patient groups. The problem with distribution-based MICs is, however, that they do not relate to any judgment of what is deemed an important change [5]. Therefore, this article focuses on anchor-based MIC methods.

Anchor-based methods use external criteria to determine what constitutes an MIC [6]. This anchor is often a global rating of perceived change by the patient. The ''mean change method'' simply declares the mean change in HRQOL score within the group reporting a minimal important change according to the anchor, to be the MIC [1]. Another method, that has become increasingly popular, originated from diagnostic test methodology where the goal

Conflict of interest: None.
Funding: None.

* Corresponding author. Tel.: (31) 20-4448199; fax: (31) 20-4448361.
E-mail address: b.terluin@vumc.nl (B. Terluin).

http://dx.doi.org/10.1016/j.jclinepi.2015.03.015
0895-4356/

What is new?

Key findings

- The predictive minimal important change (MIC) equals the receiver operating characteristic (ROC)-based MIC when the improved and not-improved groups have the same change score variance and their distributions are normal or skewed with the restriction that the skewness is oppositely directed in both groups.
- The estimation of the predictive MIC is more precise than the ROC-based MIC.

What this study adds to what was known?

- The performance of the predictive MIC has never been studied and compared with the ROC-based MIC before.
- Effect-modifying factors can more easily be included in the estimation of the MIC by using the predictive MIC.

What is the implication and what should change now?

- The predictive MIC should be used more often, especially when the MIC needs to be corrected for external factors (e.g., baseline severity).

is to ''diagnose'' important change vs. not important change [7]. This method uses receiver operating characteristic (ROC) analysis to obtain the change score that is optimally discriminating between importantly changed and not importantly changed patients. The ROC method contrasts two groups and, thus, can only analyze change in one direction at the time. The optimal ROC cutoff point, for which the sum of sensitivity and specificity reaches its maximum (the Youden criterion [8]), assures the smallest overall chance of misclassification of importantly improved patients and not-improved patients. Therefore, this optimal ROC cutoff point is generally denoted the MIC. For the present, we will limit our discussion to the MIC for improvement, but we will address the MIC for deterioration in the Discussion section.

There are a few drawbacks attached to the ROC-based MIC, which we will further denote as MIC_{ROC}. First, the MIC_{ROC} is very sensitive to random sampling variation, especially in relatively small samples. Second, as ROC analysis is a nonparametric method, obtaining confidence intervals (CIs) around the MIC_{ROC} necessitates nonparametric bootstrapping. Third, the ROC method does not allow the accommodation of external factors acting on the MIC as effect modifiers. For instance, it has repeatedly been demonstrated that the MIC sometimes depends on the severity of baseline scores [9]. The ROC method can only examine this in subgroups (except when the MIC is proportionate to baseline severity. In that case, using percentages change scores instead of the raw change scores allows for taking baseline severity into account, without having to split the sample into severity subgroups.), implying ever decreasing sample sizes.

In this article, we will introduce an alternative to the ROC-based MIC, based on predictive modeling (further denoted as ''predictive MIC'' or MIC_{pred}), which is able to overcome the drawbacks of the MIC_{ROC} mentioned above. This article is structured as follows. First, we will describe the predictive model MIC method and illustrate how MIC_{pred} and its CI are calculated. Second, we will explore and illustrate under which circumstances the ROC method and the predictive modeling method produce concordant or discordant results. Third, we will examine differences in accuracy and precision between MIC_{pred} and MIC_{ROC}. Finally, we will illustrate how the predictive MIC method is capable of accounting for effect modification.

2. Predictive model MIC method

To illustrate how the predictive MIC is calculated, we have created a sample of 100 importantly improved patients (according to a hypothetical ''anchor'') and 100 not importantly improved patients. We have artificially created perfectly symmetrical, normally distributed HRQOL change scores with a mean of 7 in the improved group and 0 in the not-improved group. These change scores can be thought of as resulting from the comparison of two measurements, a baseline measurement (T1) and a follow-up measurement (T2) with some time in between, using an HRQOL instrument with a score range of, for example, 0 to 60, where higher scores indicate better HRQOL. However, in the present example, there is no added value in simulating the underlying HRQOL scores. Therefore, we have simulated the change scores directly by generating two separate change score distributions, one for each of the ''anchor'' defined groups. Fig. 1 displays the data in an ''anchor-based MIC distribution'' format [10]. Note that the change scores consist of integer numbers. The optimal ROC cutoff point (i.e., MIC_{ROC}) that maximizes the sum of sensitivity and specificity (Youden criterion) is 3.5. Application of this cutoff point assures that the sum of percentages misclassified in both groups is as small as possible. In this case in both groups, 24% of the patients are misclassified (i.e., sensitivity and specificity are equal).

When it comes to ''predicting'' the group (either improved or not-improved) to which a patient belongs, an alternative method is to estimate a predictive statistical model with group membership according to the anchor as the outcome and change score on the HRQOL instrument as the predictor variable [11]. As the outcome (improved vs. not-improved) is dichotomous, the appropriate method

Point biserial corr.

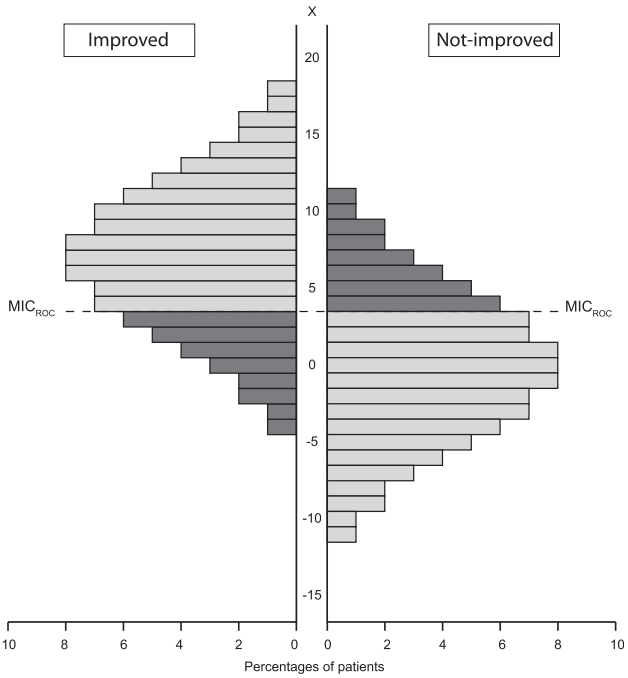

Fig. 1. Anchor-based MIC distribution of 100 improved and 100 not-improved patients. The vertical axis depicts the HRQOL change score *X*. MIC_{ROC} represents the minimal important change based on the optimal ROC cutoff point characterized by maximization of the sum of proportions correctly classified patients of both groups. The light shaded patients are correctly classified, whereas the dark shaded patients are misclassified. MIC, minimal important change; HRQOL, health-related quality of life; ROC, receiver operating characteristic.

is logistic regression analysis. The logistic regression equation reads:

$$\ln(\text{odds}_{post}) = C + B_X \times X \quad (1)$$

in which odds_{post} represents the posttest odds (in short: postodds) of being improved given a certain test result *X*, $\ln(\text{odds}_{post})$ represents the natural logarithm of this postodds, *C* represents the intercept and B_X represents the regression coefficient of the change score *X*. Note that the odds equal the probability (*P*) of belonging to the improved group, divided by the complement of this probability [i.e., odds = $P/(1-P)$]. The odds are called "posttest" because they are conditional on a specified test result. The test actually constitutes the difference between two HRQOL measurements: the change score *X*. The postodds answer the question what the odds are of belonging to the improved group given a certain amount of change in the HRQOL score.

Logistic regression analysis (using SPSS 20) of the example data in Fig. 1 yields the following parameters:

$$C = -1.070 \quad se_C = 0.235$$

$$B_X = 0.306 \quad se_{B_X} = 0.044 \quad r_{C-B_X} = -0.653$$

in which se_C represents the standard error (se) of the intercept *C*, se_{B_X} represents the se of the regression coefficient B_X and r_{C-B_X} represents the correlation between *C* and B_X. Using Equation (1), we can now calculate the $\ln(\text{odds}_{post})$ as a function of *X*.

Next, the postodds are obtained by taking the exponent of $\ln(\text{odds}_{post})$:

$$\text{odds}_{post} = \text{EXP}(\ln(\text{odds}_{post})) \quad (2)$$

The likelihood ratio (LR) of a certain test result *X* is defined as the probability of obtaining that result in the group of interest (the improved patients) divided by the probability of obtaining that result in the comparison group (the not-improved patients) [12]. An interesting characteristic of the LR is that it is also the ratio of the posttest odds of belonging to the group of interest (i.e., the improved patients) to the pretest odds of belonging to that group [12]. We use this latter characteristic to calculate the LR as a function of the change score *X*:

$$LR = \text{odds}_{post} / \text{odds}_{pre} \quad (3)$$

in which odds_{pre} represents the pretest odds (in short: preodds) of being improved. The preodds are determined independently of the test of interest. An analogy borrowed from medical diagnostic research may help to understand the nature of preodds. Consider a study into the diagnostic value of lung auscultation for detecting pneumonia. The researchers would probably recruit a sample of patients with possible pneumonia and perform two investigations independently. One investigation constitutes a diagnostic procedure (perhaps a chest X-ray and sputum culture) to determine the diagnosis pneumonia; this represents the reference or "gold standard" diagnosis. The other investigation involves lung auscultation and scoring the abnormality of the lung sounds; this represents the test of interest. Now, the pretest probability of having pneumonia in the study sample is given by the results of the independent diagnostic procedure. If *y* percent is diagnosed with pneumonia, the pretest odds are $y/(100-y)$.

In our present example, the simulated anchor provides the reference or gold standard "diagnosis" of improvement, whereas the HRQOL change score represents the test of interest. As the prevalence of being improved, based on the anchor, is 50%, we can say that the probability for every patient in the sample of being improved is 0.5. The preodds thus are $0.5/(1-0.5)$, or the prevalence of improvement, divided by (1 minus the prevalence).

Importantly, LRs, like postodds, are by definition linked to specific test results (i.e., to specific HRQOL change scores in our example). An interesting change score is the value of *X* for which the LR equals 1. Change scores with LRs >1 indicate that the posttest probability of belonging to the improved group is greater than the pretest probability, whereas change scores with LRs <1 indicate that the posttest probability of belonging to the improved group is smaller than the pretest probability. Patients with a change score with LR = 1 have the same posttest probability as

their pretest probability (i.e., 50% in this example). Note that, when it comes to predicting which patients probably have improved and which patients have not, the LR = 1 change score separates patients with a relatively large probability of belonging to the improved group from patients with a relatively small probability of belonging to the improved group (relative to the pretest probability). Therefore, in this article, we propose the change score for which LR = 1 to represent the MIC based on predictive modeling. The calculation of this MIC_{pred} can be performed using Equations (1) and (3), worked around.

$$LR = odds_{post} / odds_{pre} \quad (3)$$

As LR = 1:

$$odds_{post} = odds_{pre}$$

$$\ln(odds_{post}) = \ln(odds_{pre})$$

Substituting $\ln(odds_{post}) = C + B_X \times X$ yields: (1)

$$C + B_X \times X = \ln(odds_{pre})$$

$$X = (\ln(odds_{pre}) - C)/B_X \quad (4)$$

As $odds_{pre}$ in this example are 1, $\ln(odds_{pre}) = 0$, then MIC_{pred} is obtained by solving:

$$X = (0 + 1.070)/0.306 = 3.50$$

In this example, MIC_{pred} turns out to be exactly identical to MIC_{ROC}: 3.5. Whether this is always true will be addressed in Section 3.

Fig. 2 displays the LR and its 95% CI as a function of the change score *X*. We refer to the Appendix at www.jclinepi.com for the calculation of 95% CIs around the $\ln(odds_{post})$, $odds_{post}$, and LR, based on the values of se_C, se_{B_X}, and r_{C-B_X}.

It is important to note that LRs, unlike postodds and posttest probabilities, are largely independent of the prevalence [12]. Therefore, LRs calculated in a particular sample with a particular prevalence of improvement also apply to samples with different improvement rates. The 95% confidence limits of MIC_{pred} can also be observed in Fig. 2: the lower limit of MIC_{pred} is the *X*-value where the upper limit of the LR equals 1, whereas the upper limit of MIC_{pred} is

Not entirely true - see Terluin 2017

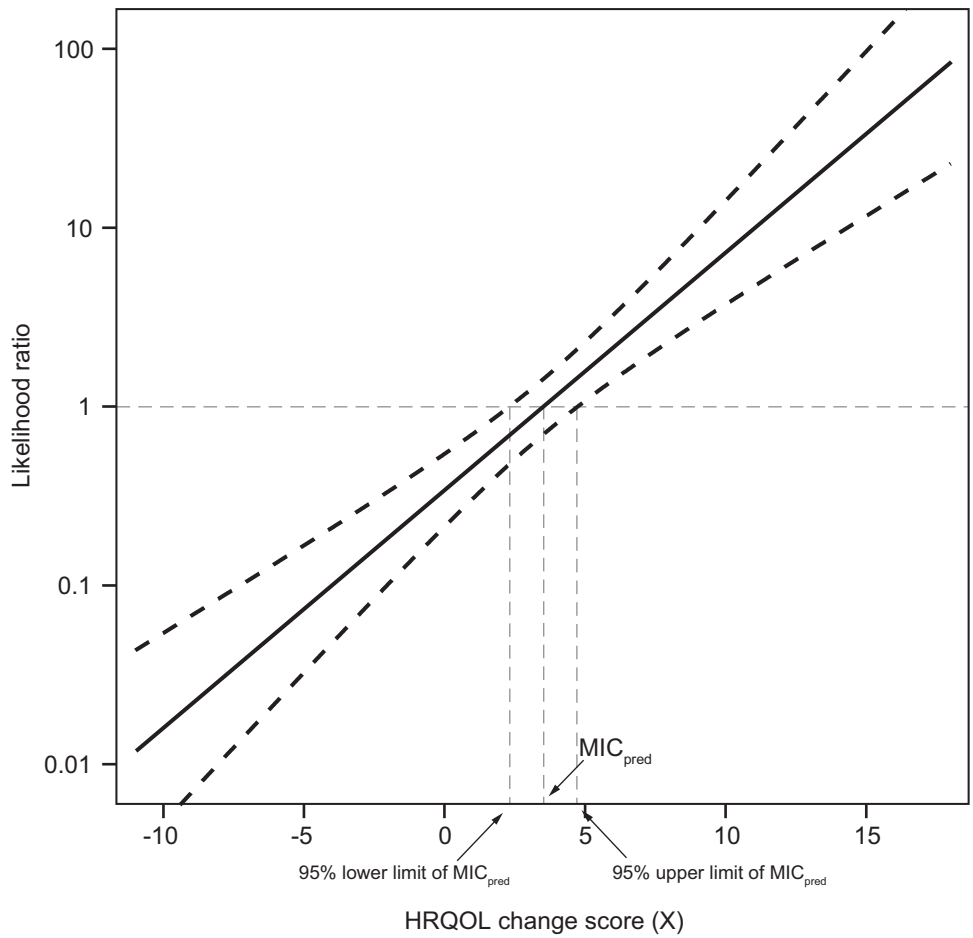

Fig. 2. Likelihood ratio (LR) as a function of the change score (*X*) (solid black line) with 95% confidence limits (dashed black lines). The dashed gray lines indicate the change score corresponding with an LR of 1 (denoted as MIC_{pred}) and its confidence limits. MIC, minimal important change; HRQOL, health-related quality of life.

the *X*-value where the lower limit of the LR equals 1. How these values are calculated is explained in the Appendix at www.jclinepi.com. We have incorporated the relevant equations into an Excel worksheet (which can be downloaded from the Journal's Web site). Entering the output of the regression analysis into the worksheet yields 2.31 and 4.69 as the 95% confidence limits of MIC_{pred}.

3. When are MIC_{ROC} and MIC_{pred} concordant and when are they discordant?

In the example used in the previous section, we have found that MIC_{ROC} and MIC_{pred} were equal. The question is whether MIC_{ROC} and MIC_{pred} are always the same, and if not, under which circumstances the estimates are the same and under which circumstances do they differ? We have explored this question using a series of simulations with different distributional parameters for the improved and not-improved patients. We used the statistical software program R [13] to simulate 10 distinct situations, differing in the prevalence of improvement (situations 1A and 1B), unequal variances of improved and not-improved groups (situations 2A and 2B), skewness of the change scores (situations 3A, 3B, and 3C), the presence of a ceiling effect (situations 4A, 4B, and 4C), and presence of a floor effect (situations 5A, 5B, and 5C). For each of the specified situations, we simulated 100 samples of 2,000 patients with change scores rounded to integers. In each of the 100 simulated samples, we calculated MIC_{pred} using Equation (4) and MIC_{ROC} using ROC analysis as implemented in the R-package "pROC" [14] and using the Youden criterion [8] for the best cutoff. We compared the mean MIC_{ROC} and MIC_{pred} over the 100 simulated samples per situation using paired *t*-tests (the R code for the simulation and analyses can be obtained from the first author).

Results are shown in Table 1. The first situation (1A) was akin to the Section 2 example: equal group sizes (i.e., prevalence = 0.5) and normally distributed change scores with mean scores of 0 and 7, respectively, and equal standard deviations (4.7). The concordance of MIC_{ROC} and MIC_{pred} under these circumstances is confirmed (difference between average MICs = 0.06; $P = 0.252$). In situation 1B, the groups have unequal group sizes (70% improved and 30% not-improved), but this does not affect the concordance between MIC_{ROC} and MIC_{pred} (difference between average MICs = 0.06; $P = 0.265$). Apparently, changing the prevalence of improvement, keeping all other distributional parameters constant, does not affect the estimation of MIC_{ROC} and MIC_{pred} and, consequently, does not affect the concordance between MIC_{ROC} and MIC_{pred}. This may not come as a surprise as the ROC analysis deals with proportions correctly classified within the improved and not-improved groups (regardless of their absolute sizes), whereas the LR aims at probabilities of certain change scores within these groups (also regardless of their sizes). Changing the relative sizes of the groups does not change the proportions improved within the groups, neither the probabilities of certain change scores within the groups.

Situations 2A and 2B evaluate the effect of unequal variance across the groups. In the simulated situations, the variance of change scores in the improved group is (much) greater than in the not-improved group. This can happen when improvement triggers a self-reinforcing process of further improvement. This has clearly different effects on the estimation of MIC_{ROC} and MIC_{pred}, MIC_{ROC} being increased whereas MIC_{pred} is being decreased, and the difference is highly significant ($P < 0.001$). Note that the reversed situation (i.e., the variance in the improved group smaller than in the not-improved group) would result in reversed discordance because the analysis is completely symmetrical with respect to improved vs. not-improved.

The next situations (3A–3C) evaluate the effect of skewness of the change score distributions. In situation 3A, the improved group is positively skewed, whereas the

Table 1. Concordance and discordance of MIC_{ROC} and MIC_{pred} under different distributional circumstances

			Change scores improved patients			Change scores not-improved patients						
Situation	Variant	Prevalence[a]	Mean[a]	SD[a]	Skewness[a]	Mean[a]	SD[a]	Skewness[a]	MIC_{ROC}[a]	MIC_{pred}[a]	*t* (df = 99)[a]	*P*
1A	Reference	0.50	7.01	4.72	−0.00	0.00	4.70	−0.00	3.57	3.51	1.153	0.252
1B	Prevalence	0.70	7.03	4.71	0.01	0.04	4.69	0.00	3.47	3.53	−1.121	0.265
2A	Variance	0.50	7.00	6.62	0.01	0.01	4.71	−0.01	4.39	3.28	17.666	0.000
2B	Variance	0.50	6.98	9.47	−0.00	−0.00	4.70	0.00	5.59	3.17	31.726	0.000
3A	Skewness	0.50	6.98	4.71	0.98	−0.00	4.72	−0.99	3.61	3.49	0.547	0.586
3B	Skewness	0.50	6.98	4.71	−0.97	0.02	4.72	0.97	3.57	3.50	1.198	0.234
3C	Skewness	0.50	7.00	4.70	−0.97	0.00	4.72	−0.98	6.29	3.77	61.107	0.000
4A	Ceiling = 36	0.50	5.44	4.25	0.11	−0.21	4.40	−0.09	2.62	2.62	−0.075	0.940
4B	Ceiling = 35	0.50	5.14	4.19	0.17	−0.25	4.33	−0.11	2.56	2.44	−2.008	0.047
4C	Ceiling = 34	0.50	4.85	4.15	0.23	−0.27	4.24	−0.12	2.37	2.27	−1.853	0.067
5A	Floor = 14	0.50	6.89	4.69	0.06	0.20	4.41	0.10	3.36	3.46	1.959	0.053
5B	Floor = 18	0.50	6.58	4.71	0.16	0.39	4.08	0.18	3.09	3.31	3.537	0.001
5C	Floor = 22	0.50	5.93	4.73	0.32	−0.52	3.57	0.36	2.76	2.93	2.454	0.016

Abbreviations: MIC, minimal important change; ROC, receiver operating characteristic; SD, standard deviation; df, degrees of freedom.
[a] Mean values over 100 simulated samples per specified situation.

not-improved group is negatively skewed. Consequently, the forms of the distributions in both groups mirror each other. This situation could occur when the anchor and the HRQOL change score are highly correlated and measurement error is relatively small. In situation 3B, the skewness of the groups is reversed, but the forms of the distributions mirror each other just the same. In situation 3C, both groups have the same direction of skewness, and therefore, the distributions do not mirror each other. As long as the skewed distributions of the groups mirror each other (3A and 3B), the estimation of MIC_{ROC} and MIC_{pred} is in agreement ($P > 0.05$). However, when the skewed distributions do not mirror each other (3C), MIC_{ROC} and MIC_{pred} are discordant ($P < 0.001$). The increase in MIC_{ROC} values in situations 2A, 2B, and 3C probably results from the flattened left tail of the change score distribution curve of the improved patients, combined with a much steeper right tail of the change score distribution curve of the not-improved patients in these situations.

The final six situations (4A–5C) deal with ceiling and floor effects of the HRQOL scale. Ceiling and floor effects indicate that the scale fails to measure HRQOL levels beyond the range of the scale. As ceiling and floor effects occur in the HRQOL measurements and only indirectly affect change scores, we had to simulate the HRQOL scores underlying the change scores. First, we generated a normally distributed baseline HRQOL score (T1) with a mean of 25 and a standard deviation of 7, without taking ceiling or floor effects into account yet. Second, we defined a random half of the sample as not-improved and the other half as improved (the anchor). Third, for each of these anchor-defined groups, we generated normally distributed change scores with means of 0 and 7, respectively, and equal standard deviation of 4.7, much as we did in situation 1A. Fourth, we calculated the HRQOL follow-up score (T2) by summing the T1 and change scores. Note that the scores simulated so far represent ''virtual'' scores that could be observed in the absence of ceiling or floor effects. However, in the fifth step, we introduced ceiling or floor effects by trimming the T1 and T2 scores to a specified maximum (i.e., the ceiling) or minimum (i.e., the floor). In the final step, we recalculated the ''observed'' change scores by subtracting the trimmed T1 score from the trimmed T2 score.

Ceiling effects are simulated in situations 4A–4C. The percentage observations in the highest scores vary between 7% and 12% at T1 and between 20% and 30% at T2. Apparently, ceiling effects tend to push the estimated MIC values downward, MIC_{pred} a little more so than MIC_{ROC} causing discordance between MIC_{ROC} and MIC_{pred}. Note that the existence of ceiling effects also affects the variance and skewness of the change score distributions to some extent. Floor effects are simulated in situations 5A–5C. The percentage observations in the lowest scores vary between 7% and 35% at T1 and between 7% and 25% at T2. Floor effects appear to exert a much stronger effect on the change score distributions of the not-improved patients than on the improved patients, resulting in unequal variances and some nonmirroring skewness. This clearly results in discordance between MIC_{ROC} and MIC_{pred}.

To summarize, MIC_{ROC} and MIC_{pred} are concordant when the improved and not-improved groups have the same variance and preferably normal distributions or at least distributions that mirror each other in form. Unequal variances or nonmirroring distributions yield discordant results for MIC_{ROC} and MIC_{pred}.

4. Accuracy and precision

This section examines possible differences in accuracy and precision between MIC_{pred} and MIC_{ROC}. We will do so by simulating a large number of samples drawn from a hypothetical population in which, akin to the Section 2 example, the prevalence of improvement is 50% and the change scores in the improved and not-improved groups are normally distributed and have the same variance. Thus, the given ''true'' population MIC (i.e., the gold standard in this section), by either method, is 3.5, as we have seen in Section 2. In Section 3, we have seen that MIC_{pred} and MIC_{ROC} yield similar average results over a number of repetitions in large sample sizes. Here, we will examine the variability of MIC_{pred} and MIC_{ROC} in samples more closely mimicking the size of samples usually encountered in research practice.

We have generated 1,000 samples of 200 patients with the following parameters of the population distribution: prevalence of improvement: 50%, mean [and standard deviation (SD)] of the change score in the improved group: 7 (4.7), mean (and SD) of the change score in the not-improved group: 0 (4.7), and change score distribution in both groups was normal. Note that any deviations from the given parameters in the individual samples must be ascribed to random sampling variation. Scores were rounded to integer values. In each of the 1,000 samples, we calculated MIC_{pred} and MIC_{ROC} and their 95% confidence limits to determine their accuracy and precision. Logistic regression was performed, and MIC_{pred} was calculated using Equation (4). The 95% confidence limits of MIC_{pred} were calculated using the equations from the Appendix at www.jclinepi.com. ROC analysis and the estimation of MIC_{ROC} (i.e., best Youden cutoff) were performed using the package ''pROC'' [14]. To calculate the 95% confidence limits of MIC_{ROC} in each of the 1,000 simulated samples, nonparametric bootstrapping was used (drawing 2,000 bootstrap samples; the R code for the simulation and analyses can be obtained from the first author).

The simulation results are presented in Table 2. The mean estimates of MIC_{ROC} and MIC_{pred} were both not significantly different from the given population MIC

Table 2. Accuracy and precision of MIC_{ROC} and MIC_{pred} over 1,000 simulations[a] (true MIC value = 3.5)

Statistic	MIC_{ROC}	MIC_{pred}	Test statistic	df	*P*
Mean MIC estimate	3.47	3.50	$F = 0.952$	1; 1998	0.329
Variance MIC estimate	1.305	0.119	Levene = 598.9	1; 1998	0.000
Bias (mean MIC estimate −3.5)	−0.035	0.002	$F = 0.952$	1; 1998	0.329
Mean square error (MSE = $bias^2$ + variance)	1.306	0.119			
Root mean square error (RMSE)	1.143	0.345			
Coverage[b]	99.1%	100%	$\chi^{2c} = 7.143$	1	0.008
Mean 95% CI length	3.90	2.32	$F = 1292.4$	1; 1998	0.000

Abbreviations: MIC, minimal important change; ROC, receiver operating characteristic; df, degrees of freedom; CI, confidence intervals.

[a] Sample size = 200, prevalence of improvement = 0.5, change scores normally distributed with equal variances in improved and not-improved patients.

[b] Coverage = percentage of times the 95% CI of the MIC estimate includes the true MIC value of 3.5.

[c] With continuity correction.

and not significantly different from each other, indicating that both methods to determine the MIC are equally accurate (i.e., unbiased) under the specific circumstances in this simulation (i.e., change scores with normal distributions and equal variances of the groups). However, the variance of MIC_{ROC} across 1,000 samples was much larger than the variance of MIC_{pred}, resulting in a large difference in mean square error (MSE). The square root of the MSE suggests that the MIC_{ROC} estimate was on average 1.14 points off the true population MIC value, whereas the MIC_{pred} estimate was on average much closer to the true population value. The average length of the CI of MIC_{pred} was significantly smaller than the average length of the CI of MIC_{ROC}, suggesting greater precision of MIC_{pred} over MIC_{ROC}. The coverage (i.e., the percentage of times the true population value was included in the 95% CI) was very good in both methods.

To summarize, under the specification of normal distributions and equal variances of the change scores in the improved and not-improved groups, MIC_{ROC} and MIC_{pred} are both accurate estimates of the given population MIC value, but MIC_{pred} constitutes a more precise estimate than MIC_{ROC}.

5. Accounting for MIC modifying factors

This section illustrates how the predictive MIC method is capable of accounting for effect modification by a third variable such as the severity of baseline scores. In a number of cases, the MIC depends on baseline severity [9]. Understandably, patients with a more severely affected quality of life need greater improvement to perceive change as (minimally) important. The predictive MIC method allows for a direct testing of subgroup effects. To illustrate the testing of baseline severity as an effect modifier, building on the data of the Section 2 example, we split the improved and not-improved groups into a high baseline severity group and a low baseline severity group, effectively creating four groups of patients: low baseline and not-improved patients, low baseline and improved patients, high baseline and not-improved patients, and high baseline and improved patients. To simulate the effect of baseline severity on the MIC, we increase the change scores of the high baseline and improved patients by five points, effectively increasing the mean change score in improved high baseline severity patients from 7 to 12. In the other groups, the change scores remain as in the Section 2 example. Given that the mean change score of the not-improved high baseline severity patients remains 0, this creates a situation in which the MIC for high baseline patients is 6 [i.e., (7 + 5)/2] instead of 3.5. Note that the MIC for low baseline severity patients remains the same, that is, 3.5. Logistic regression analysis, including baseline severity as effect modifier, yields the following output.

$C = -1.089$ $\quad se_C = 0.335$

$B_X = 0.306$ $\quad se_{B_X} = 0.062$ $\quad r_{C-B_X} = -0.659$

$B_S = -2.136$ $\quad se_{B_S} = 0.833$ $\quad r_{C-B_S} = -0.402$ $\quad r_{B_X-B_S} = 0.265$

$B_{XS} = 0.240$ $\quad se_{B_{XS}} = 0.130$ $\quad r_{C-B_{XS}} = 0.316$ $\quad r_{B_X-B_{XS}} = -0.479$ $\quad r_{B_S-B_{XS}} = -0.832$

In which B_S and B_{XS} are the regression coefficients of severity and the interaction of the change score and severity, respectively. The se_{B_S} and $se_{B_{XS}}$ are the standard errors of these coefficients, and r_{C-B_S}, $r_{C-B_{XS}}$, $r_{B_X-B_S}$, $r_{B_X-B_{XS}}$, and $r_{B_S-B_{XS}}$ are the correlations between C and B_S, C and B_{XS}, B_X and B_S, B_X and B_{XS}, and B_S and B_{XS}, respectively.

The regression equation of the model looks like:

$$\ln(\text{odds}_{\text{post}}) = -1.089 + 0.306 \times X - 2.136 \times \text{Sev} + 0.240 \times X \times \text{Sev}$$

in which Sev represents severity (coded 1 for high and 0 for low severity).

Note that for patients with low baseline severity (severity = 0), the equation simplifies into:

$$\ln(\text{odds}_{\text{post}}) = -1.089 + 0.306 \times X$$

Entering the C and B_X coefficients together with the correlation coefficient r_{C-B_X} in the Excel sheet yields a MIC_{pred} of 3.56 and 95% confidence limits 1.79 and 5.34. Note that the MIC_{pred} is slightly different from the original 3.5 because the sample could not be divided into two exactly identical samples of low and high severity. Note also that the CI is substantially wider than in the original example because presently two extra parameters (B_S and B_{XS}) must be estimated, resulting in an increase of the standard errors.

The easiest way to calculate the MIC_{pred} for the high severity group is to reverse the severity coding (1 = low baseline severity and 0 = high baseline severity) and repeat the analysis. The output now yields:

$C = -3.225$ $\quad se_C = 0.762$

$B_X = 0.546$ $\quad se_{B_X} = 0.114$ $\quad r_{C-B_X} = -0.877$

$B_S = 2.136$ $\quad se_{B_S} = 0.833$ $\quad r_{C-B_S} = -0.916$ $\quad r_{B_X-B_S} = 0.803$

$B_{XS} = -0.240$ $\quad se_{B_{XS}} = 0.130$ $\quad r_{C-B_{XS}} = 0.770$ $\quad r_{B_X-B_{XS}} = -0.878$ $\quad r_{B_S-B_{XS}} = -0.832$

in which B_S represents the regression coefficient of severity reversed.

For patients with high baseline severity (severity reversed = 0), the regression equation becomes:

$$\ln(\text{odds}_{\text{post}}) = -3.225 + 0.546 \times X$$

Entering the C and B_X values together with the correlation between C and B_X in the Excel sheet yields a MIC_{pred} = 5.91 (95% CI: 4.48, 7.36). Note that this outcome is slightly different from the expected MIC_{pred} = 6 because the sample could not be divided into two exactly identical samples of low and high severity.

In sum, the predictive model method to calculate the MIC provides the opportunity to test possible modifying factors of the MIC and calculate the MIC for relevant subgroups without having to split the sample into subgroups.

6. Discussion

In this article, we have presented a new approach to the calculation of the MIC of HRQOL scales. The resulting ''predictive MIC'' (MIC_{pred}) represents the change score characterized by an LR of 1, indicating that the probability of belonging to the improved group (as opposed to the not-improved group) equals the average probability of being improved in the sample. Change scores above the MIC_{pred} indicate that the likelihood of being improved is greater than the average probability of being improved, whereas change scores below the MIC_{pred} suggest a smaller than average probability of belonging to the improved group. We have shown that the MIC_{pred} is identical to the MIC calculated by ROC analysis (MIC_{ROC}), provided that the change score variance is the same in the improved and not-improved groups and the change score distribution is normal in both groups or, when the distributions are not-normal, the forms of the distributions in the groups mirror each other. Under these conditions, MIC_{pred} might be preferred over MIC_{ROC} as it is a much more precise estimate of the MIC, which translates in gain in efficiency and statistical power in research situations. Fortunately, in practice, change scores often follow a fairly normal distribution.

However, in situations characterized by gross variance inequality across the improved and not-improved groups or when the forms of the change score distributions do not mirror each other, we have observed that MIC_{pred} and MIC_{ROC} estimates significantly diverge. We feel that the meaning of MIC_{pred}, with its emphasis on the predictive value of change scores (expressed in the LR), is intuitively more appealing than MIC_{ROC}. However, further research should clarify the relative merits of both methods under various circumstances.

The predictive framework, dealing with probabilities, odds, and LRs, provides more information to individual change scores over establishing whether the probability of being improved is greater or smaller than the average probability in the sample. How much the change score is below or above the MIC can be translated into score-specific LRs and posttest probabilities. Consider, for example, a patient from the Section 2 example with a change score of 8. This score is above the MIC of 3.5, so the posterior probability

of being improved must be greater than the prior probability (which was 50%). However, using the predictive MIC method, the probability can be quantified with more precision. Entering the results of the logistic regression analysis in the Excel sheet, and entering a change score of 8, yields an LR of 3.97 (95% CI: 2.35, 6.69) and a posttest probability of 0.80 (95% CI: 0.70, 0.87). So, the patient with a change score of 8 has a probability of being improved of 80% (95% CI: 70%, 87%), which is a much more precise estimate than >50%. The predictive framework provides an elegant way to determine the meaning of specific change scores.

So far, we have dealt only with the MIC for improvement. The MIC for deterioration, whenever of interest, can be determined in the same way as the MIC for improvement by dividing the total sample in deteriorated patients and not-deteriorated patients. The Excel sheet can be used in the same way to determine MIC_{pred} and its 95% CI.

We must acknowledge some limitations of our study. First, this study used simulated data only. However, the simulated data were characterized by situations that are often seen in practice. Consequently, the methods were studied for the most common situations. Second, we did not cover all possible distributional particularities in our simulations. For example, we did not examine bimodal distributions of change scores. Third, in those cases where MIC_{pred} and MIC_{ROC} are discordant, we are unable to decide which of these MICs should be preferred in practice. Future studies of real or simulated data should clarify the relative merits of MIC_{pred} and MIC_{ROC} under different circumstances.

In conclusion, we feel that we have identified the two crucial distributional characteristics that determine the concordance of MIC_{pred} and MIC_{ROC}: equal variances of the improved and not-improved groups and normal distributions of change scores or not normal (e.g., skewed) distributions in the groups that mirror each other with respect to the forms of the distributions. In these particular situations (which can be empirically tested), we recommend that MIC_{pred} be used instead of MIC_{ROC} because MIC_{pred} offers greater precision as it is less sensitive to random sampling variation. In addition, MIC_{pred} offers the opportunity to test and estimate the effect of possible modifying variables such as the severity of baseline scores.

Supplementary data

Supplementary data related to this article can be found online at http://dx.doi.org/10.1016/j.jclinepi.2015.03.015.

References

[1] Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 1989;10:407−15.
[2] King MT. A point of minimal important difference (MID): a critique of terminology and methods. Expert Rev Pharmacoecon Outcomes Res 2011;11:171−84.
[3] Crosby RD, Kolotkinc RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol 2003;56:395−407.
[4] Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol 2008;61:102−9.
[5] de Vet HC, Terwee CB, Ostelo RW, Beckerman H, Knol DL, Bouter LM. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. Health Qual Life Outcomes 2006;4:54.
[6] Turner D, Schünemann HJ, Griffith LE, Beaton DE, Griffiths AM, Critch JN, et al. The minimal detectable change cannot reliably replace the minimal important difference. J Clin Epidemiol 2010;63:28−36.
[7] Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. J Chron Dis 1986;39:897−906.
[8] Youden WJ. Index for rating diagnostic tests. Cancer 1950;3:32−5.
[9] 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.
[10] de Vet HC, Ostelo RW, van der Roer N, Knol DL, Beckerman H, Boers M, et al. Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. Qual Life Res 2007;16:131−42.
[11] Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. New York: Springer; 2009.
[12] Deeks JJ, Altman DG. Diagnostic tests 4. Likelihood ratios. BMJ 2004;329:168−9.
[13] R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2012.
[14] Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12:77.