Journal of Clinical Epidemiology 83 (2017) 90–100

Journal of Clinical Epidemiology

The anchor-based minimal important change, based on receiver operating characteristic analysis or predictive modeling, may need to be adjusted for the proportion of improved patients

Berend Terluin[a,*], Iris Eekhout[b,c], Caroline B. Terwee[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, Amsterdam 1081 BT, The Netherlands

[b]Department of Epidemiology and Biostatistics, EMGO Institute for Health and Care Research, VU University Medical Center, Amsterdam, The Netherlands

[c]Department of Child Health, Netherlands Organisation for Applied Scientific Research (TNO), Schipholweg 77-89, Leiden 2316 ZL, The Netherlands

Accepted 16 December 2016; Published online 14 January 2017

Abstract

Objectives: Patients have their individual minimal important changes (iMICs) as their personal benchmarks to determine whether a perceived health-related quality of life (HRQOL) change constitutes a (minimally) important change for them. We denote the mean iMIC in a group of patients as the "genuine MIC" (gMIC). The aims of this paper are (1) to examine the relationship between the gMIC and the anchor-based minimal important change (MIC), determined by receiver operating characteristic analysis or by predictive modeling; (2) to examine the impact of the proportion of improved patients on these MICs; and (3) to explore the possibility to adjust the MIC for the influence of the proportion of improved patients.

Study Design and Setting: Multiple simulations of patient samples involved in anchor-based MIC studies with different characteristics of HRQOL (change) scores and distributions of iMICs. In addition, a real data set is analyzed for illustration.

Results: The receiver operating characteristic–based and predictive modeling MICs equal the gMIC when the proportion of improved patients equals 0.5. The MIC is estimated higher than the gMIC when the proportion improved is greater than 0.5, and the MIC is estimated lower than the gMIC when the proportion improved is less than 0.5. Using an equation including the predictive modeling MIC, the log-odds of improvement, the standard deviation of the HRQOL change score, and the correlation between the HRQOL change score and the anchor results in an adjusted MIC reflecting the gMIC irrespective of the proportion of improved patients.

Conclusion: Adjusting the predictive modeling MIC for the proportion of improved patients assures that the adjusted MIC reflects the gMIC. Limitations: We assumed normal distributions and global perceived change scores that were independent on the follow-up score. Additionally, floor and ceiling effects were not taken into account.

Keywords: Minimal important change; Receiver operating characteristics; Predictive modeling; Proportion improved patients; Adjusted minimal important change; Present state bias

1. Introduction

Health-related quality of life (HRQOL) has become an important outcome in current studies evaluating the benefit and harm of treatments for various medical conditions. However, questionnaires designed to measure HRQOL provide scores that lack intrinsic meaning. Similarly, changes in such scores offer no obvious interpretation of the importance of those changes. This is where the concept of the "minimal important change" (MIC; also called "minimal clinically important change" or "minimal (clinically) important difference") comes in [1]. The MIC is defined as the minimal amount of change in an HRQOL score that is perceived as "important" [2,3].

MICs can be determined in different ways. Two broad approaches include distribution-based methods and anchor-based methods [4]. Distribution-based methods relate HRQOL change scores to the distribution of change scores or the probability that a change score might be attributed to measurement error. Anchor-based methods relate HRQOL change scores to an external criterion (the anchor) of what constitutes the smallest HRQOL change that is deemed important [5]. Often, the patient's "global perceived change" (GPC) is used as such an anchor—and

* 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.2016.12.015
0895-4356/

What is new?

Key findings

- Patients have their individual minimal important changes as their personal benchmarks of what constitutes a (minimally) important change in health-related quality of life (HRQOL). We denote the mean individual MIC in a patient sample as the "genuine MIC" (gMIC).
- Based on multiple simulation studies, we found that the anchor-based minimal important change (MIC), determined by the receiver operating characteristic (ROC) method or by the predictive modeling method, equaled the gMIC when the proportion of improved patients was 0.5. HRQOL (change) scores were assumed to be normally distributed.
- When the proportion of improved patients was either greater or smaller than 0.5, the MIC was a biased estimate of the gMIC. However, we were able to derive an empirical formula to correct this bias for the predictive modeling MIC. The ROC-based MIC was too imprecise to allow meaningful adjustment.

What this adds to what was known?

- Both the bias of the MIC by the proportion improved and the adjustment of this bias have not been described before. Future research can produce MIC values more accurately approximating the gMIC in a given patient sample.What is the implication and what should change now?

What is the implication and what should change now?

- We propose to substitute the predictive modeling MIC for the ROC-based MIC. In addition, we propose to consider adjusting the MIC for the proportion improved, especially when the HRQOL (change) scores appear to be normally distributed.

for good reasons. By providing the anchor for the MIC, patients directly provide the standard by which to measure the benefits and harms of their treatments.

HRQOL changes can be important in two opposite directions: improvement and deterioration. For the sake of simplicity, however, we will limit our discussion to the direction of improvement. Everything that is true for the MIC for improvement applies—although reciprocally—to the MIC for deterioration. In the Section 8, we will offer specific suggestions on how to apply the methodology elaborated in this paper to the MIC for deterioration.

Furthermore, we will focus on two specific anchor-based approaches: the receiver operating characteristic (ROC) method and the predictive modeling method. A popular anchor-based MIC method uses ROC analysis to determine the change score that is optimally discriminating between importantly improved patients and not importantly improved patients [6]. Recently, we have introduced a novel method to determine an anchor-based MIC based on predictive modeling [7]. In this approach, the MIC is related to the change score-specific likelihood ratio (i.e., the ratio of the likelihood of a specific change score in the importantly improved group to the likelihood of that specific change score in the not importantly improved group). The predictive modeling-based MIC (in short: "predictive MIC") is defined as the change score for which the likelihood ratio equals 1 [7]. We have demonstrated that the predictive MIC method and the ROC-based method provide identical MIC values when the HRQOL change scores are normally distributed, and their variances are equal across the improved and not improved groups. In addition, the predictive MIC was more precise than the ROC-based MIC, and its 95% confidence interval (CI) was easier to calculate [7].

The ROC-based or predictive MIC is based on the analysis of HRQOL change scores in relation to patient-rated GPC scores. Apparently, patients are able to rate their GPC in HRQOL. We assume that they do so by comparing their perceived HRQOL improvement with their individual minimal important change (iMIC). If patients' perceived HRQOL improvement exceeds their personal iMIC, they will rate themselves as "improved," and otherwise, they will rate themselves as "not improved." We assume that each patient has their own iMIC as a personal benchmark of what constitutes a minimal important HRQOL improvement. As the MIC is intended to reflect the minimal improvement that is important to (a group of) patients, it seems desirable that the MIC reflects the mean of the iMICs in a group of patients. So, the mean iMIC should constitute the gold standard to compare the MIC with. We propose to denote the mean iMIC as the "genuine MIC" (gMIC) because it is the amount of improvement that the "average" patient values as (minimally) important. The gMIC (i.e., the mean iMIC) of a group of patients is, assumingly, relatively invariant. Unfortunately, iMICs and their mean, the gMIC, are not directly observable and measurable. The gMIC is a theoretical construct, put in place to explain how patients respond to GPC questions in an anchor-based MIC study.

Global perceived change and iMIC

Genuine MIC = avg. iMIC

So, how does the ROC-based or predictive MIC relate to the gMIC? Although iMICs escape direct observation, the relationship between the MIC and a set of iMICs can be studied using simulation techniques. Through simulations, unobservable variables can be created and controlled to enable the exploration of relationships between hidden phenomena (e.g., iMICs) and observable variables (e.g., MICs). Moreover, it is easy to simulate large samples with specified characteristics. The leading question in this paper is ‘‘How does the MIC, as determined by ROC analysis or predictive modeling, compare to the gMIC?’’.

We have extensively explored the relationship between ROC based or predictive MICs and gMICs using simulations, and we will present some interesting findings. In particular, we will examine:

- To what extent the ROC based or predictive MIC equals the gMIC when the proportion of improved patients (hereafter: ‘‘proportion improved’’) is 0.5 (Section 3);
- To what extent the MIC equals the gMIC when the proportion improved is smaller or greater than 0.5 (Section 4);
- If it is possible to adjust the MIC for the proportion improved bias (Sections 5 and 6);
- The application of the MIC adjustment in a real data set (Section 7).

Before we continue with these findings, we will describe our simulation method in the next section.

2. Simulating an anchor-based MIC study

We have conducted our simulations from the perspective of classical test theory (CTT) that assumes that observed test scores consist of ‘‘true’’ scores and error scores (i.e., measurement error). Considering observed HRQOL scores, CTT assumes that these scores constitute the sum of the patients’ experience of various levels of HRQOL (i.e., the true scores) and errors that occur when patients answer the questions (items) of an HRQOL questionnaire (the error scores). Importantly, just like perceived levels of HRQOL can be captured in true scores, so can perceived changes in those levels be captured in true change scores. CTT assumes further that the measurement errors are distributed equally and randomly across all patients and all measurements irrespective of their true score levels. True scores, true change scores and error scores can all be expressed in the scale of the HRQOL questionnaire. CTT thus provides a language to distinguish between what patients experience (but what cannot be measured directly) and what is obtained through HRQOL assessments (Fig. 1).

We started by simulating the HRQOL experience of a sample of patients at baseline (T1) (Fig. 1, box ①). To avoid too much sample-to-sample fluctuations, we chose a

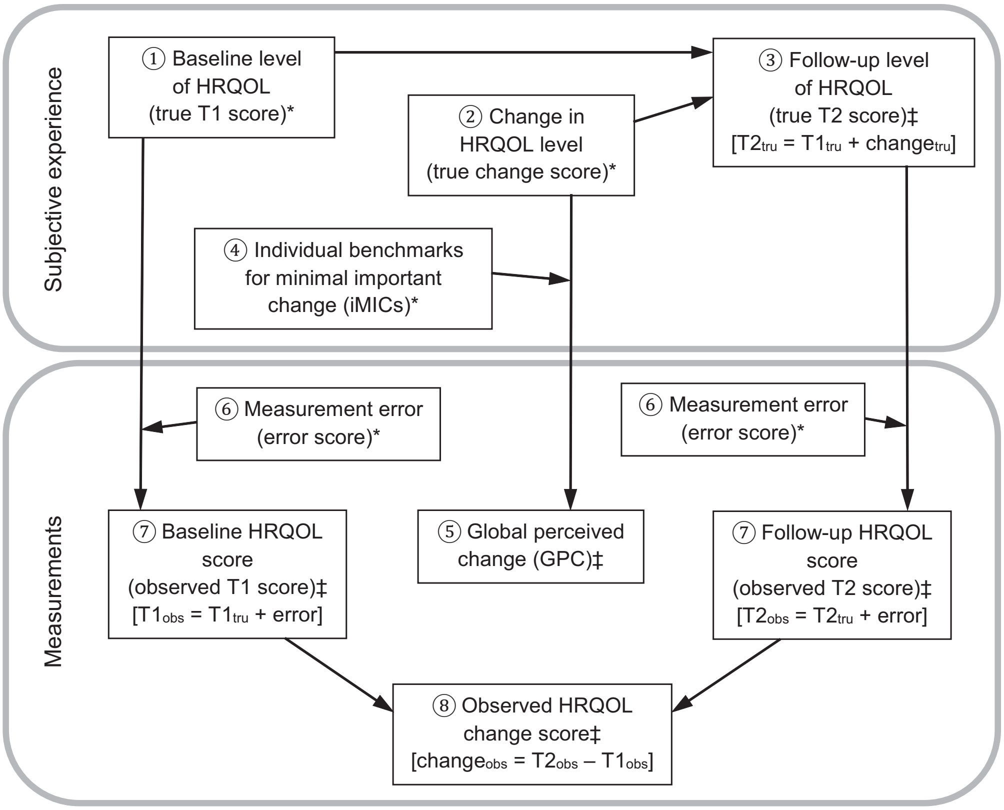
Subjective experience
① Baseline level of HRQOL (true T1 score)*
② Change in HRQOL level (true change score)*
③ Follow-up level of HRQOL (true T2 score)‡ [T2tru = T1tru + changetru]
④ Individual benchmarks for minimal important change (iMICs)*
Measurements
⑥ Measurement error (error score)*
⑥ Measurement error (error score)*
⑦ Baseline HRQOL score (observed T1 score)‡ [T1obs = T1tru + error]
⑤ Global perceived change (GPC)‡
⑦ Follow-up HRQOL score (observed T2 score)‡ [T2obs = T2tru + error]
⑧ Observed HRQOL change score‡ [changeobs = T2obs – T1obs]

Fig. 1. Graphical representation of the simulation process of an anchor-based MIC study. The upper panel displays variables confined to the subjective experience of patients. The lower panel shows the measurement process through which the subjective experience is being captured in observed HRQOL (change) scores. Simulated variables are indicated by an asterisk (*), whereas variables that are derived from other variables are indicated by a two-barred cross (‡). The formula for deriving the latter variables is shown in brackets. The GPC was derived by comparing the true change score with the iMIC: when the true change score exceeded the iMIC the GPC was coded ‘‘improved,’’ otherwise the GPC was coded ‘‘not improved.’’ HRQOL, health-related quality of life.

relatively large sample size ($n = 2000$). We generated a variable representing the true baseline HRQOL score at T1, a normally distributed score with an arbitrary mean of 40 and a standard deviation (SD) of 10 on a scale between 0 and 100, higher scores representing better HRQOL (Fig. 2A). Next, we assumed that the mean perceived HRQOL improved between T1 and T2 (Fig. 1, box ②). We simulated the improvement by generating a normally distributed true HRQOL change score with a mean of 15 and an SD of 10 (Fig. 2B). Note that the mean change score of 15 corresponded to a group-level (mean) improvement with an effect size of 1.5. The SD of 10 assured a range of change scores between −15 and 45 (Fig. 2B). The change in perceived HRQOL resulted in a new HRQOL situation at T2 (Fig. 1, box ③). Therefore, the true T2 score was derived by summing the true T1 score and the true change score (Fig. 2A).

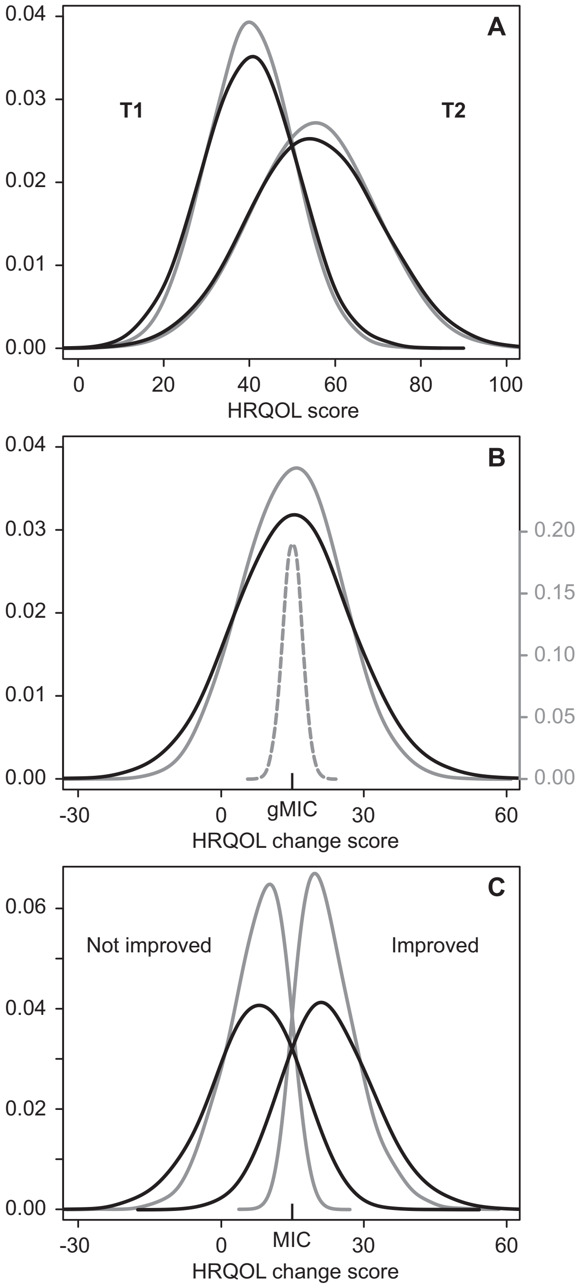

Fig. 2. Density plots of simulated variables in an anchor-based MIC study. (A) presents the distributions of true (gray curves) and observed (black curves) HRQOL scores at baseline (time 1, T1) and follow-up (time 2, T2). (B) shows the distributions of true (solid gray curve) and observed (black curve) HRQOL change scores (left y-axis) and the individual minimal important changes (iMICs; dashed gray curve, right y-axis). The mean iMIC represents the "genuine MIC" (gMIC). (C) displays the distributions of true (gray curves) and observed (black curves) HRQOL change scores of the improved and not improved patients according to the anchor whether or not the true HRQOL change score exceeded the iMIC. The MIC is also indicated in (C).

Next, we simulated the patients' iMICs (Fig. 1, box ④). Assuming that, like other human attributes (such as height), the iMICs followed a normal distribution, we generated a normally distributed variable with a mean of 15 and an SD of 2 to represent the iMICs' distribution (Fig. 2B). Importantly, the difference between the mean change score and the mean iMIC (the gMIC) would determine the proportion of improved patients according to their GPC. If the mean change score equaled the gMIC (as in Fig. 2B), half of the patients were expected to have an HRQOL improvement that was greater than their iMIC and, consequently, the proportion improved would be 0.5. If the mean change score would have been smaller than the gMIC, the proportion improved would be less than 0.5. Similarly, if the mean change score would have been greater than the gMIC, the proportion improved would be greater than 0.5.

For now, we chose the gMIC to equal the mean true change score to obtain equally sized groups of improved and not improved patients. The iMICs' SD of 2 was chosen as to prevent very small or even negative iMICs, which would not be credible as individual benchmarks of important improvement. Next, we derived the patients' GPC responses by comparing their true change scores with the values of their iMICs (Fig. 1, box ⑤). The GPC score served as the anchor to define the improved and not improved groups in the simulation (Fig. 2C).

Then, we obtained "observed" HRQOL scores. So far, we had been simulating subjectively experienced true levels of (or changes in) HRQOL. However, when measuring HRQOL, measurement error needs to be taken into account (Fig. 1, box ⑥). Note that measurement error, by definition, occurs during measurements, hence at T1 and T2. Therefore, we generated two independent variables with means of 0 and SDs of 5 representing the measurement errors of the HRQOL measurements. As measurement error is random, the errors' mean is typically 0. The observed HRQOL scores at T1 and T2 were then derived by summing the true HRQOL scores and their respective measurement errors and then rounded to the nearest integer as most HRQOL scales use integer scales (Fig. 1, box ⑦). Note that the errors' SD was chosen to be 5 as to obtain a reliability of the observed T1 score of 0.80. Finally, we obtained observed change scores by subtracting the observed T1 scores from the observed T2 scores (Fig. 1, box ⑧).

Once we had created an observed change score and a GPC anchor, an MIC could be determined. We used two different methods: the ROC-based method [6] and the predictive modeling method [7]. The ROC-based MIC was calculated using the statistical package pROC version 1.8 [8] in the statistical program R, version 3.2.0 [9]. The best ROC cutoff according to the Youden criterion is the cutoff that maximizes the sum of sensitivity and specificity [10]. The predictive MIC was determined by logistic regression analysis with the observed change score as independent variable and the GPC anchor as dependent variable. The change score associated with a likelihood ratio of 1 represents the predictive MIC [7]. All simulations were performed in R 3.2.0 [9] (all R codes are provided in Supplementary File 1 at www.jclinepi.com on the Journal's Web site).

3. MIC equals gMIC if proportion improved is 0.5

In the present section, we will examine to what extent the ROC-based and predictive MICs equal the gMIC when the improved and not improved groups are equally sized.

We examined this in a large number of different simulated anchor-based MIC studies using various distributional parameters regarding (1) the mean and SD of the T1 scores, (2) the reliability of the scores, (3) the mean and SD of the change scores, and (4) the gMIC and the SD of the iMICs. We can think of this as simulating the performance of a number of different HRQOL scales (hence different scale parameters) in different patient populations (hence different iMIC distributions). However, in each simulated study, the proportion improved was constrained to 0.5 by setting the gMIC (i.e., the mean iMIC) equal to the mean change score. Table 1 lists the parameters varying across the samples. Some parameters were defined as ratios of that parameter to another parameter. For example, the mean true change score was defined as a ratio of the mean true change score to the SD of the true T1 score. Thus, the mean true change score ranged from 0.5 × 0.2 × 30 = 3 to 1.5 × 0.3 × 50 = 22.5. All variables were assumed to be normally distributed.

The variations in parameters resulted in 486 different combinations. For each combination, five simulated samples were created. For each of the 2,430 simulated samples, we estimated ROC based and predictive MICs. The mean proportions improved across the samples was 0.50 (range 0.48−0.52). The gMICs ranged from 2.9−23.4. Fig. 3 shows that the relationship between the MICs and the gMICs across the samples was perfectly linear, apart from sample fluctuation. There was no evidence that any of the parameters in which the samples varied, impacted on the MIC−gMIC relationship. So, when the proportion improved is 0.5, the MIC equals the gMIC irrespective of other sample characteristics. In other words, when the proportion improved is 0.5, the following equation is true:

$$\mathrm{MIC} = \mathrm{gMIC} = \text{mean change score}$$

Consequently, when the proportion improved is 0.5, there is no need to use any statistical method to calculate the MIC. In this situation, the MIC equals the mean change score and reflects the gMIC. This is true irrespective of sample characteristics such as the reliability of the scores, the mean change score relative to the SD of the T1 scores, and the variability of the change scores.

4. Proportion improved affects the MIC

The present section examines what happens to the ROC based and predictive MICs, relative to the gMIC, when the proportion improved varies between 0.2 and 0.8. We took the example simulation in Section 2 as starting point and generated samples with the same values for the mean, SD and reliability of the T1 scores, for the SD of the true change scores, and for the mean and SD of the iMICs. However, to create different proportions improved, the mean change score was made to vary between 7 and 23, whereas the gMIC was fixed to 15. We can think of these simulated samples as studies of the same HRQOL instrument (hence the same scale parameters) in similar patients (hence the same gMIC in all samples). Because of treatments with different effectiveness, the study samples experienced different degrees of mean improvement in HRQOL, resulting in different proportions improved across the samples. We generated 150 samples for each of the 17 different mean change scores.

Fig. 4 (left panels) shows the relationship between the MICs and the proportions improved. Fitted regression lines are also shown. The estimated MIC values decreased as the proportions improved decreased, and, conversely, the MIC values increased as the proportions improved increased. The likely explanation of the MIC shifting away from the gMIC when the proportion improved shifts away from 0.5 is that changes in the proportion improved are associated with shifts in the mean change score in the direction of the largest group (see Supplementary File 2 at www.jclinepi.com for a more detailed explanation).

Close inspection of Fig. 4 (upper left panel) reveals a cubic relationship between the proportion improved and the predictive MIC. A log-odds transformation of the proportion improved, however, demonstrated a linear relationship with the predictive MIC (Fig. 4, upper right panel). The log-odds of improvement (further denoted as "log-odds(imp)") is the natural logarithm of $p/(1 - p)$ in which p represents the proportion improved. It should be noted that $p = 0.5$ corresponds with log-odds(imp) $= 0$.

5. Adjusting the MIC for the proportion improved

In this section, we will examine how the MIC can be adjusted for the effect of the proportion improved. As the predictive MIC is a much more precise function of the

Table 1. Varying parameters across 2,430 simulated samples (Section 3)

Parameter	Values	Explanation
Mean true T1 score	30, 40, 50	Arbitrary values
SD true T1 score	0.20, 0.25, 0.30	Values are ratios of the SD of the true T1 score to the mean true T1 score
Mean true change score	0.5, 1, 1.5	Values are ratios of the mean true change score to the SD of the true T1 score
SD true change score	0.5, 0.75, 1	Values are ratios of the SD of the true change score to the mean true change score
Reliability of the T1 score	0.60, 0.70, 0.80	Values are reliability coefficients
gMIC (mean of the iMICs)	Equal to mean true change score	gMIC was set equal to the mean true change score to obtain proportions improved of 0.5
SD of the iMICs	0.10, 0.15	Values are ratios of the SD of the iMICs to the gMIC (mean of the iMICs)

Abbreviations: SD, standard deviation; iMIC, individual MIC; gMIC, genuine MIC (= mean iMIC).

gMIC (Fig. 3) and the log-odds(imp) than the ROC-based MIC (Fig. 4), we will focus on adjusting the predictive MIC, leaving the ROC-based MIC aside.

The linear relationship between the MICs and the log-odds(imp), given a certain gMIC (Fig. 4, upper right panel), can be described by the following regression equation:

$$\mathrm{MIC} = \mathrm{gMIC} + S \times \log - \mathrm{odds(imp)} \quad (1)$$

When the log-odds(imp) equals 0, the MIC equals the gMIC. For one unit increase in log-odds(imp), the MIC increases with S, the slope coefficient. If the value of S were known, adjusting the MIC to approximate the gMIC would be easy. Substituting "adjusted MIC" (aMIC) for gMIC offers the following equation for the adjusted MIC:

$$\mathrm{aMIC} = \mathrm{MIC} - S \times \text{log-odds(imp)} \quad (2)$$

If we could "predict" the value of S with sufficient precision from observable/measurable sample characteristics, we would be able to calculate the aMIC using Equation (2). Therefore, we examined S in a large set of simulated samples (denoted as "exploration set") in which all parameters, used so far in Sections 3 and 4, were allowed to vary. This time, we also varied the proportion improved by shifting the mean change score relative to the gMIC. Table 2 presents an overview of the parameters.

There were 2,430 possible combinations of sample parameters and for every combination we simulated two samples. Across these samples, we then tried to "predict" S from the sample characteristics. First we needed to calculate S using a rewritten version of Equation (1):

$$S = (\mathrm{MIC} - \mathrm{gMIC})/\text{log-odds(imp)}$$

Because the log-odds(imp) can be 0, and to avoid division by 0, we temporarily proceeded with a subset of samples in which log-odds(imp) was either < -0.1 or > 0.1. Next, we explored possible predictors of S, notably the mean and SD of the observed T1, T2 and change scores, the reliability of these scores, and the point biserial correlation between the change score and the GPC anchor. A point biserial correlation is the product moment correlation between a continuous variable and a dichotomous variable. By far the best predictor of S turned out to be the SD of the observed change score (R^2 adjusted: 0.66). Together with the correlation between the observed change score and the anchor, these predictors explained 69% of the variance of S. The final model for the prediction of S was:

$$S = 0.090 \times \mathrm{SD}_{\mathrm{change}} + 0.103 \times \mathrm{SD}_{\mathrm{change}} \times Cor \quad (3)$$

where $\mathrm{SD}_{\mathrm{change}}$ represents the SD of the observed change score and Cor represents the correlation between the observed change score and the GPC anchor. Next, aMIC values in all samples of the exploration set were calculated using Equations (2) and (3). Fig. 5 shows the relationships

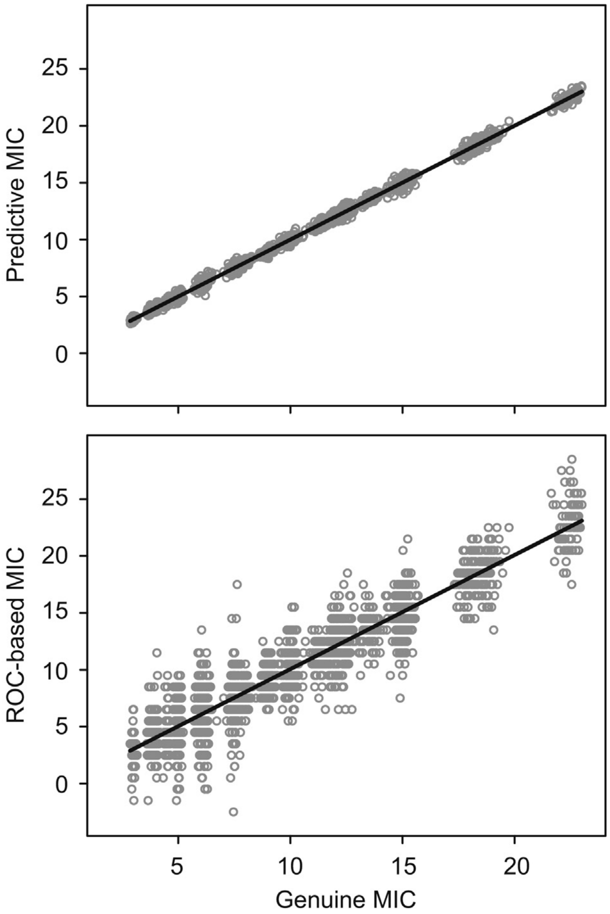

Fig. 3. Associations of the predictive MIC and the ROC-based MIC with the genuine MIC across 2,430 simulated samples in which the proportion of improved patients was constrained to 0.5. ROC, receiver operating characteristic; MIC, minimal important change.

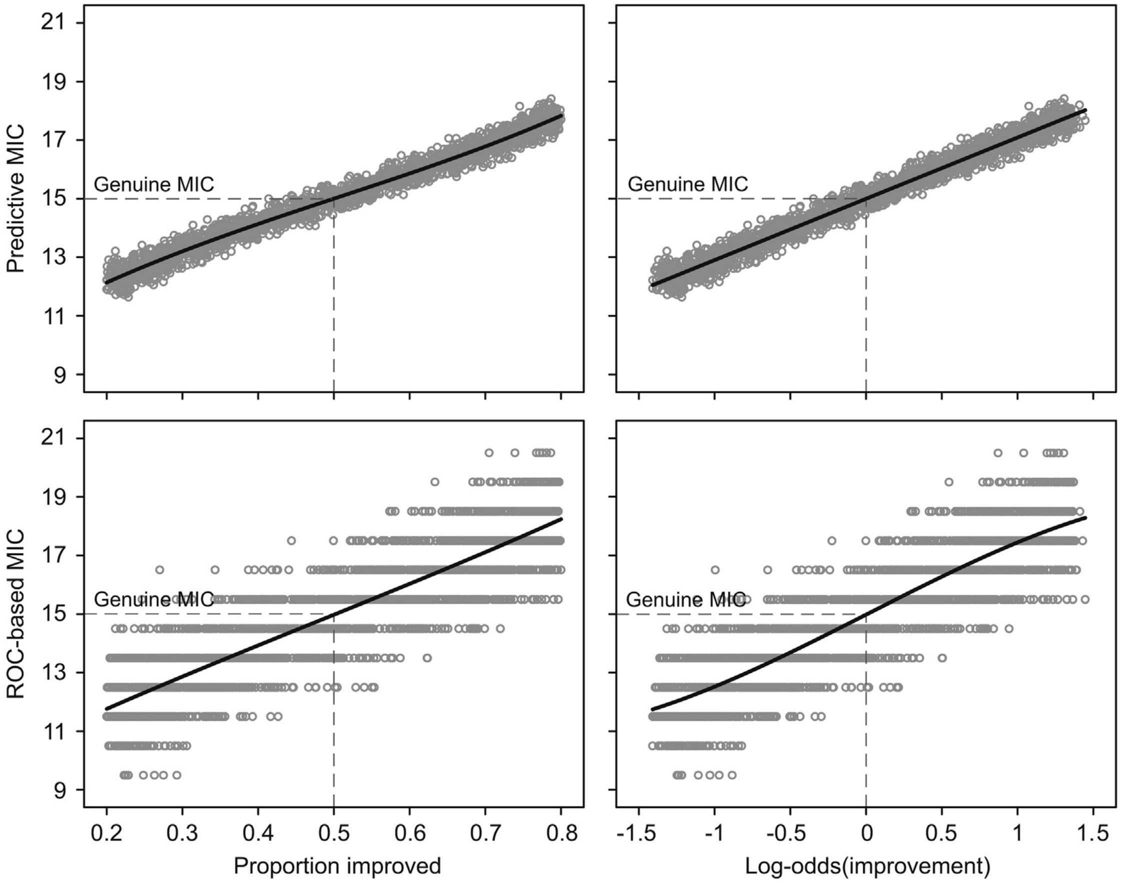

Fig. 4. Associations of the predictive MIC (upper panels) and ROC-based MIC (lower panels) with the proportion of improved patients (left hand panels) and the log-odds of improvement (right hand panels). Results of 2,550 simulated samples in which the genuine MIC was constrained to 15. The dashed lines indicate that, on average, the MICs equaled the genuine MIC when the proportion improved was 0.5 and the log-odds of improvement was 0. ROC, receiver operating characteristic; MIC, minimal important change.

between the MIC and the gMIC (upper panel) and between the aMIC and the gMIC (lower panel). The correlation between the MIC and the gMIC was 0.847, whereas the correlation between de aMIC and the gMIC was 0.994. Clearly, the aMIC was a much better approximation of the gMIC than the MIC. Table 3 (column "exploration set") shows the residual statistics of gMIC minus aMIC. The mean difference between the gMIC and the aMIC was practically 0, indicating that the aMIC was an unbiased estimate of the gMIC. The aMIC was in some samples a little off the mark, but in over 95% of the samples, this was less than 1 point (on a scale of approximately 100 points). The root mean square error indicated that the average deviation of the aMIC (either above or below) the gMIC across all samples was 0.37 points. So, the aMIC turned out to be an unbiased and precise approximation of the gMIC.

6. Validation of the adjusted MIC formula

In this section, we will validate the formula for adjusting the MIC, as developed in the previous section, in a new set of simulated data, denoted the "validation set." Simulating different proportions improved can be accomplished by

Table 2. Varying parameters across 4,860 simulated samples in the exploration set (Section 5)

Parameter	Values	Explanation
Mean true T1 score	30, 40, 50	Arbitrary values
SD true T1 score	0.20, 0.25, 0.30	Values are ratios of the SD of the true T1 score to the mean true T1 score
gMIC (mean of the iMICs)	0.5, 0.8, 1.1	Values are ratios of the gMIC to the SD of the true T1 score
SD of the iMICs	0.10, 0.15	Values are ratios of the SD of the iMICs to the gMIC (mean of the iMICs)
SD true change score	0.5, 1, 1.5	Values are ratios of the SD of the true change score to the SD of the true T1 score
Mean true change score	−0.84, −0.38, 0, 0.38, 0.84	Values express the difference between the mean true change score and the gMIC in SD units of the true change score
Reliability of the T1 score	0.60, 0.70, 0.80	Values are reliability coefficients

Abbreviations: SD, standard deviation; iMIC, individual MIC; gMIC, genuine MIC (= mean iMIC).

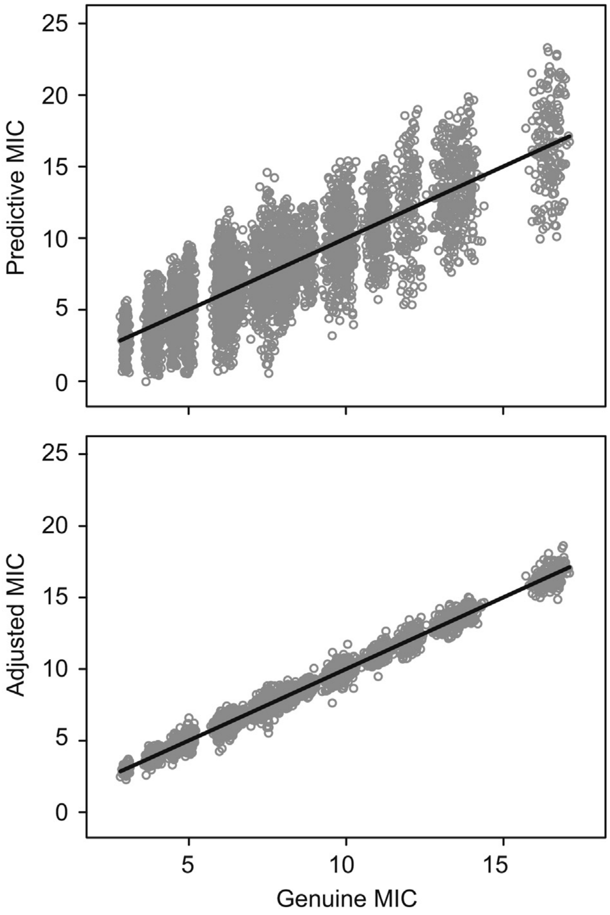

Fig. 5. The association of the predictive MIC (upper panel) and the adjusted MIC (lower panel) with the genuine MIC in the exploration set. MIC, minimal important change.

shifting the mean change score and the gMIC (i.e., the mean iMIC) relative to each other (see Section 2). In Section 5, we fixed the gMIC and then shifted the mean change score relative to the gMIC. Instead, in the present section, we shifted the gMIC relative to the mean change score. The parameters that were varied are presented in Table 4.

There were 4,374 possible combinations of sample parameters, and for every combination, we simulated two samples. Inspection of the samples revealed that in 533 samples, the gMIC was smaller than 0.2 times the SD of the observed T1 score, implying a gMIC value corresponding with less than a small effect in terms of effect size. Because we assumed that gMICs should represent at least a small effect, we deemed these very small gMIC values unrealistic. Therefore, these samples were removed. In the remaining 8,215 samples, we calculated the aMIC according to Equations (2) and (3). Fig. 6 demonstrates that the aMIC was a more accurate and a more precise estimate of the gMIC than the MIC. Note that, in the validation set, the smallest gMIC values belonged to samples with proportions improved >0.5, whereas the greatest gMIC values belonged to samples with proportions improved <0.5. The upward bias of the MIC in samples with proportions improved >0.5 and the downward bias of the MIC in samples with proportions improved <0.5 was evidenced by a clockwise rotation of the regression line away from the diagonal (indicated by a dashed line) in the upper panel. After adjustment, in the lower panel, the aMIC estimates were neatly aligned along the diagonal. Table 3 (column "validation set") shows the residual statistics of gMIC minus aMIC, confirming the validity of the approach to adjust the MIC for the proportion improved.

Table 3. Residual statistics for the adjusted MIC (aMIC) as predictor of the genuine MIC (gMIC)

Statistic	Exploration set	Validation set
Mean (bias)	−0.00	0.08
Minimum	−1.72	−1.89
Maximum	1.99	2.38
95% confidence interval	−0.74, 0.74	−0.66, 0.95
Variance	0.14	0.15
Mean squared residuals (bias^2 + variance)	0.14	0.16
Root mean squared residuals	0.37	0.40

As it is virtually impossible to simulate all possible combinations of parameters, we developed a "do-it-yourself validation tool" that researchers may use to test the proposed method to adjust the MIC specifying their own set of parameters. The tool is provided as Supplementary File 3 at www.jclinepi.com, and a short manual is provided as Supplementary File 4 at www.jclinepi.com. The tool consists of an R-code simulating a user-specified number of samples consisting of a user-specified number of patients. All parameters that we have varied can be specified by the researcher. In addition, correlations between the T1 score and the change score, and between the T1 score and the iMICs can be specified (we did not find that these correlations had any influence on the MIC or the adjusted MIC). Moreover, the tool offers the possibility to simulate (partial) dependency of the GPC on the T2 score (see Section 8 for the reason why). The researcher needs no knowledge of R to run the code and obtain results.

7. Real data example

We thankfully used the data from 442 low back pain patients providing baseline and 12-week follow-up HRQOL scores and GPC scores after 12 weeks of treatment [11]. Separate analyses were conducted for (sub)acute and chronic patients. The study included three HRQOL instruments and three methods to calculate an MIC, but we focused only on the Quebec Back Pain Disability Scale and the ROC-based method (which we compared with the predictive modeling method). The Quebec Back Pain Disability Scale is a 20-item instrument, measuring physical functioning in low back pain patients [12]. Each item

Table 4. Varying parameters across 8,215 simulated samples in the validation set (Section 6)

Parameter	Values	Explanation
Mean true T1 score	30, 40, 50	Arbitrary values
SD true T1 score	0.20, 0.25, 0.30	Values are ratios of the SD of the true T1 score to the mean true T1 score
Mean true change score	0.5, 1, 1.5	Values are ratios of the mean true change score to the SD of the true T1 score
SD true change score	0.5, 0.75, 1	Values are ratios of the SD of the true change score to the mean true change score
gMIC (mean of the iMICs)	−0.85, −0.65, −0.45, −0.25, 0, 0.25, 0.45, 0.65, 0.85	Values express the difference between the gMIC and the mean true change score in SD units of the true change score
SD of the iMICs	0.10, 0.15	Values are ratios of the SD of the iMICs to the gMIC (mean of the iMICs)
Reliability of the T1 score	0.60, 0.70, 0.80	Values are reliability coefficients

Abbreviations: SD, standard deviation; iMIC, individual MIC; gMIC, genuine MIC (= mean iMIC).

is scored on a 6-point scale ranging from 0 (no trouble) to 5 (unable to), and the total score ranges from 0 (no dysfunction) to 100 (maximum dysfunction). We reversed the scores to obtain a scale in which higher scores represented better HRQOL (i.e., less disability) and positive change scores represented improvement. The results of our reanalysis are presented in Table 5. The proportion improved was much greater in (sub)acute patients than in chronic patients (0.83 vs. 0.54). We found predictive MIC values of 15.6 (95% CI: 13.2, 17.9) and 9.0 (95% CI: 6.7, 11.2), respectively. Bootstrapping (1,000 samples) was used to calculate CIs. The predictive MIC values were not too much different from the ROC-based MICs reported by van der Roer et al. (17.5 and 8.5). After adjusting for the proportion improved, the adjusted MIC in the (sub)acute patients dropped substantially from 15.6 to 11.8 (95% CI 9.1, 14.4), whereas the adjusted MIC in the chronic patients decreased just a little bit. This example illustrates that adjusting for the proportion improved can have a major impact on the estimation of the gMIC.

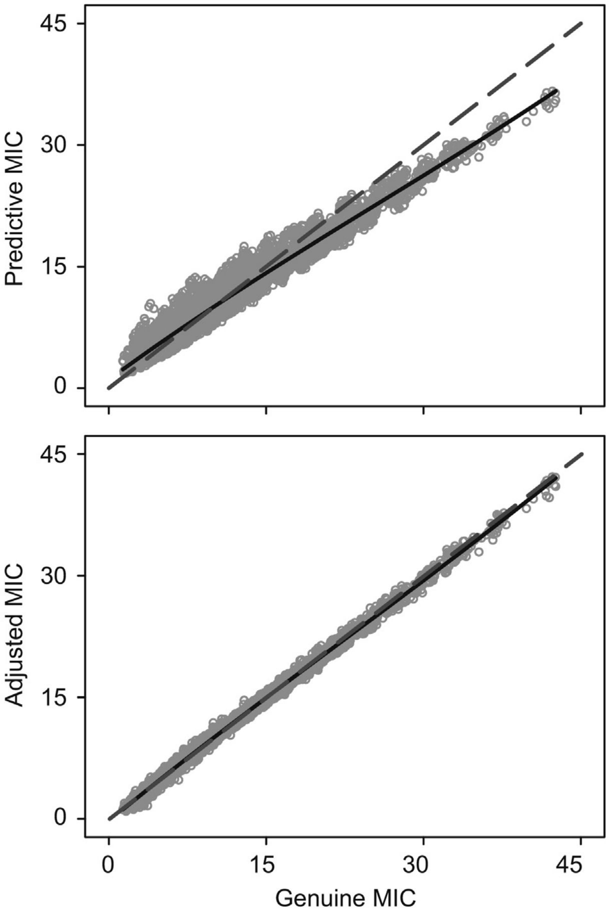

Fig. 6. The association of the predictive MIC (upper panel) and the adjusted MIC (lower panel) with the genuine MIC in the validation set. The dark-gray dashed lines indicate the diagonals along which the observations should be aligned when the MIC equals the genuine MIC. MIC, minimal important change.

8. Discussion

This is the first study to demonstrate that an ROC based or predictive MIC, as determined in a typical anchor-based MIC study, equals the "gMIC" in a patient sample provided that the improved and not improved groups are equally sized (i.e., when the proportion improved is 0.5). However, when the improved and not improved groups differ in size, the MIC will be biased. If less than 50% of the patients are improved, the MIC will be an underestimation of the gMIC, whereas if a majority of patients are improved, as in the (sub)acute low back pain patients in Section 7, the MIC will be an overestimation of the gMIC. We were able to empirically derive a formula to adjust the predictive MIC for the proportion improved (Box 1) to obtain an "adjusted MIC" (aMIC) as a more accurate estimate of the gMIC.

It is reasonable to assume that many ROC-based MIC values reported in the literature are biased estimates of the gMIC. Reported MIC values of a specific HRQOL scale may vary considerably from study to study because of

Table 5. Results of the reanalysis of a real data set

Statistic	(Sub)Acute (n = 303[a])	Chronic (n = 138)
Mean baseline score	57.8	64.8
SD baseline score	17.9	18.1
Mean change score	22.0	10.2
SD change score	19.3	16.0
Correlation anchor — change score	0.36	0.36
Proportion improved	0.83	0.54
Log-odds of improvement	1.55	0.15
ROC-based MIC[b]	17.5	8.5
Predictive MIC	15.6	9.0
95% confidence interval	13.2, 17.9	6.7, 11.2
Adjusted MIC	11.8	8.7
95% confidence interval	9.1, 14.4	6.2, 11.2

Abbreviations: SD, standard deviation; ROC, receiver operating characteristic; MIC, minimal important change.

[a] One patient had missing value(s).

[b] ROC-based MIC reported by van der Roer et al. [11].

differences in the proportions improved. Moreover, we have seen that the ROC-based MIC is a rather imprecise estimate even with sample sizes as large as 2,000. The predictive MIC is a much more precise estimate and can be adjusted for the proportion improved. The ROC-based MIC is so imprecise that it is virtually useless to try to adjust it.

The methodology for the ROC-based MIC, the predictive MIC, and the adjusted MIC for deterioration is reciprocal to the methodology for these MICs for improvement. Probably, the safest way to determine an (adjusted) MIC for deterioration is by reversing the HRQOL scores so that higher scores mean worse HRQOL and positive change scores represent deterioration. In addition, the GPC for deterioration should be substituted for the GPC for improvement, and the proportion deteriorated should replace the proportion improved. Next, the methods as described above can be applied, including the formula for adjusting the MIC. Finally, the results should be reversed back to the original scale.

Throughout our simulations, we have treated the GPC as a valid indicator of the change in HRQOL between T1 and T2 as perceived by the patients. However, some authors suggest that the GPC may be influenced more by the level of the HRQOL at follow-up (i.e., the T2 score) than by the change per se [13—15]. This so-called "present state bias" is thought to result from recall bias: patients remember their current state better than a previous state [16]. This may be a greater problem with longer T1—T2 intervals. From the literature, it is not clear to what extent present state bias is really a problem. Many MIC researchers take the validity of the GPC as a matter of course. However, we have asked ourselves how much effect this could have on our approach of the MIC and the aMIC. Supplementary File 5 at www.jclinepi.com describes a simulation study in which we progressively added more "dependency" of the GPC on the T2 score. The results indicated that T2 score dependency did not affect the estimation of the ROC based and predictive MICs when the proportion improved equaled 0.5. However, when the proportion improved was either smaller or greater than 0.5 (we tested 0.25 and 0.75), T2 score dependency resulted in additional bias on top of the proportion improved bias. The adjustment formula worked well for adjusting the proportion improved bias but not for the extra bias caused by the T2 score dependency. The effect, however, was relatively limited: even when the GPC was totally based on the T2 score, the adjusted MIC was still a more accurate estimate of the gMIC than the unadjusted MIC when no T2 score dependency was present. We provided the possibility to define various degrees of T2 score dependency (from 0% to 100%) in the "do-it-yourself" validation tool.

We need to acknowledge important limitations of the present work. First, we examined the relationship between MICs, gMICs, and proportions improved empirically using simulations, as a first step to elucidate the matter. We were not able to clarify the associations underlying the formula to adjust the MIC from a theoretical perspective. However, we encourage the scientific community to take on the challenge to approach the matter in a more theoretical manner. In the future, our formula for adjusting the MIC may be improved based on a more comprehensive theoretical understanding of the MIC—gMIC relationship or by additional empirical research. In any case, situations not fulfilling our present assumptions, such as skewed or bimodal (change) scores and the presence of floor or ceiling effects in the measurements, need further study. In addition, much more work needs to be done on the issue of T2 score dependency of the GPC.

In the meantime, we propose that researchers substitute the predictive MIC for the ROC-based MIC (because of its

Box 1 Formula for calculating the adjusted MIC

$$\mathrm{MIC}_{\mathrm{Adjusted}} = \mathrm{MIC}_{\mathrm{Predictive}} - (0.090 + 0.103 \times Cor) \times \mathrm{SD}_{\mathrm{change}} \times \text{log-odds(imp)}$$

Where $\mathrm{MIC}_{\mathrm{Adjusted}}$ = adjusted minimal important change; $\mathrm{MIC}_{\mathrm{Predictive}}$ = predictive minimal important change; *Cor* = point biserial correlation between the HRQOL change score and the anchor; $\mathrm{SD}_{\mathrm{change}}$ = standard deviation of the HRQOL change score; log-odds(imp) = log-odds of improvement = natural logarithm of [proportion improved/(1 − proportion improved)] greater precision) and consider to calculate the adjusted MIC especially when the baseline, follow-up, and change scores appear to be normally distributed.

9. Conclusions

When in an ROC-based or predictive MIC study, the (importantly) improved group and the not improved group are equally sized (i.e., balanced), the MIC reflects the mean of the individual MICs patients use as their personal benchmarks of a MIC (i.e., the gMIC). However, imbalance of the improved and not improved groups causes a shift in the estimated MIC away from the gMIC. When the (change) scores are normally distributed, it appears to be possible to adjust the (predictive) MIC using an equation including the MIC, the log-odds of improvement, the SD of the change score, and the correlation between the change score and the GPC anchor.

Acknowledgments

The authors thank Dr. Nicole van der Roer for sharing the real example data.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jclinepi.2016.12.015.

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] de Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine. A practical guide. Cambridge: Cambridge University Press; 2011.

[4] Crosby RD, Kolotkinc RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol 2003;56:395−407.

[5] 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.

[6] 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.

[7] Terluin B, Eekhout I, Terwee CB, de Vet HCW. Minimal important change (MIC) based on a predictive modeling approach was more precise than MIC based on ROC analysis. J Clin Epidemiol 2015;68:1388−96.

[8] 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.

[9] R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2012.

[10] Youden WJ. Index for rating diagnostic tests. Cancer 1950;3:32−5.

[11] van der Roer N, Ostelo RWJG, Bekkering GE, van Tulder MW, de Vet HCW. Minimal clinically important change for pain intensity, functional status, and general health status in patients with nonspecific low back pain. Spine 2006;31:578−82.

[12] Kopec JA, Esdaile JM, Abrahamowicz M, Abenhaim L, Wood-Dauphinee S, Lamping DL, et al. The Quebec back pain disability scale: conceptualization and development. J Clin Epidemiol 1996;49:151−61.

[13] Wyrwich KW, Tardino VM. Understanding global transition assessments. Qual Life Res 2006;15:995−1004.

[14] Knox SA, King MT. Validation and calibration of the SF-36 health transition question against an external criterion of clinical change in health status. Qual Life Res 2009;18:637−45.

[15] Kamper SJ, Ostelo RWJG, Knol DL, Maher CG, de Vet HCW, Hancock MJ. Global perceived effect scales provided reliable assessments of health transition in people with musculoskeletal disorders, but ratings are strongly influenced by current status. J Clin Epidemiol 2010;63:760−766.e1.

[16] Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. J Clin Epidemiol 1997;50:869−79.