Title
Between-group minimally important change versus individual treatment responders

Journal
Quality of Life Research, 30(10)

ISSN
0962-9343

Authors
Hays, Ron D
Peipert, John Devin

Publication Date
2021-10-01

DOI
10.1007/s11136-021-02897-z

Peer reviewed

Quality of Life Research
https://doi.org/10.1007/s11136-021-02897-z

COMMENTARY

Between-group minimally important change versus individual treatment responders

Ron D. Hays[1] · John Devin Peipert[2]

Accepted: 25 May 2021

Abstract

Purpose Estimates of the minimally important change (MIC) can be used to evaluate whether group-level differences are large enough to be important. But responders to treatment have been based upon group-level MIC thresholds, resulting in inaccurate classification of change over time. This article reviews options and provides suggestions about individual-level statistics to assess whether individuals have improved, stayed the same, or declined.

Methods Review of MIC estimation and an example of misapplication of MIC group-level estimates to assess individual change. Secondary data analysis to show how perceptions about meaningful change can be used along with significance of individual change.

Results MIC thresholds yield over-optimistic conclusions about responders to treatment because they classify those who have not changed as responders.

Conclusions Future studies need to evaluate the significance of individual change using appropriate individual-level statistics such as the reliable change index or the equivalent coefficient of repeatability. Supplementing individual statistical significance with retrospective assessments of change is desirable.

Keywords Meaningful change · Minimally important difference · Responder · Reliable change index

Introduction

In health-related quality of life (HRQOL) research, the significance of group-level change is evaluated to assess treatment efficacy and effectiveness. In addition, group-level minimally important change (MIC) thresholds are used because trivial mean change can be statistically significant if the sample size is large enough. The MIC indicates if statistically significant group mean differences are large enough to be important or meaningful to patients and clinicians. Identifying those who improve ("responders" to treatment) provides important supplemental information to group-level change. This paper reviews approaches for assessing MIC and estimating responders to treatment. We note that while group-level MIC thresholds have been used to identify responders to treatment in HRQOL studies [1, 2], other approaches are more appropriate.

Estimating the MIC

MIC estimates rely on anchors to provide an external indication of the level of underlying change. The variety of possible anchors makes a single MIC estimate problematic. It is advisable to use multiple anchors whenever possible, but the most used anchor is a retrospective rating of change question such as:

How is your health now compared to 6 weeks ago?

Much better
A little better
About the same
A little worse
Much worse

✉ Ron D. Hays
drhays@ucla.edu

[1] Department of Medicine, UCLA Division of General Internal Medicine and Health Services Research, 1100 Glendon Avenue, Suite 850, Los Angeles, CA 90024, USA

[2] Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL 60611, USA

Published online: 15 June 2021

This example item refers to change in "health". Depending on the context and measure being evaluated, the anchor might be worded more specifically such as "physical functioning", "pain", "getting along with family", etc. The choice of words is likely to result in different MIC estimates. In addition, there are known limitations of retrospective ratings of change, which include a tendency to reflect the patient's current state more than change, potentially due to recall bias [3, 4].

Change on the target measure should be correlated and have a monotonic association with change indicated on the anchor. The mean group change on the target measure should be larger for the subgroup of people who report they are much better on the anchor than mean change for the other subgroups. And those who report no change on the anchor should have no more than minimal change on the target measure [5]. The mean group changes on a HRQOL measure for those who report being "a little better" (improvement) or "a little worse" (decrement) are the basis for MIC estimates. But sometimes investigators fail to limit the MIC to those who changed a little and include all those with any change on the anchor. This was the case in a sample of 123 adult surgical patients with spinal deformity [6] and in a study of 223 patients with chronic low back pain [7]. Including all those who change rather than focusing on those with minimal but important change means that the MIC thresholds are too large.

Identifying responders to treatment

Individual-level variation and change can be estimated using simulation modeling for time series data, but it requires a minimum of 10 observation in the data stream [8]. Similarly, Moinpour et al. [9] estimated mixed effect models and noted that the PROMIS fatigue computer adaptive test would need 15 total assessments to obtain 0.90 reliability of change. Because of limits on research budgets and concerns about respondent burden, nearly all longitudinal HRQOL studies are limited to a few waves of assessment (e.g., two time points). Guidance for identifying responders to treatment for this environment are needed. Hence, we review approaches for estimating individual change from baseline to a single post-baseline assessment.

Table 1 lists several formulae previously proposed for estimating the significance of individual change that are analogous to t-tests for within group change [10, 11]. All the formulae include individual change in the numerator and "error" in the denominator. The different methods vary in how they estimate error–for example, the time 1 standard deviation (SD), standard error of measurement (SEM; $\left(SD_1\sqrt{1-\text{reliability}}\right)$, standard error of estimation $\left(SD_1\sqrt{\text{reliability}\,(1-\text{reliability})}\right)$ and the standard error of prediction $\left(SD_1\sqrt{1-\text{reliability}^2}\right)$. The $\sqrt{2}$ SEM is used for the reliable change index (RCI). Note that these formulae

Table 1 Formula for Evaluating Individual Change

Statistic	Formula
Standard deviation index	$(X_2 - X_1)/SD_1$
Standard error of measurement (SEM)	$SD_1\sqrt{1-\text{reliability}}$
95% Confidence interval around SEM	$X_1 \pm 1.96$ SEM
Standard error of estimation	$SD_1\sqrt{\text{reliability}(1-\text{reliability})}$
Standard error of prediction	$SD_1\sqrt{1-\text{reliability}^2}$
Reliable change index	$(X_2 - X_1)/\sqrt{2}$ SEM
Coefficient of repeatability[a]	$1.96\sqrt{2}$ SEM
Reliable change index (practice effects)	$(X_2 - X_1$ - practice effects$)/\sqrt{2}$ SEM
Reliable change index (item response theory)	$(X_2 - X_1)/\sqrt{SE_1^2 + SE_2^2}$
Regression-based	$\left(X_2 - X_{2p}^b\right)/\sqrt{S_1^2 + S_2^2}\sqrt{1-\text{reliability}}$

[a]Also known as the smallest detectable change, minimally detectable change, or smallest real difference. A group-level version of the coefficient of repeatability has been proposed by dividing the formula by the $\sqrt{n}$

[b]Prediction of time 2 from time 1 variables

Note: X_2 = score at time 2; X_1 = score at time 1; SD_1 = standard deviation at time 1; SE_1 = standard error at time 1; SE_2 = standard error at time 2; S_1^2 = variance at time 1; S_2^2 = variance at time 2 used instead of SD_1. The SD of change can be used instead of SD_1. The estimated true score (mean + reliability (X_1-mean)) may be used instead of X_1 to account for regression to the mean

have between group variance in the denominators that may not be representative of variance in individual change.

Following the conventional $p < 0.05$ threshold for group-level research, responders are usually defined by an RCI of 1.96 or larger. A variant of the RCI used for cognitive measures corrects for practice effects [12], though caution has been raised about its use [13]. The denominator of the RCI for item response theory (IRT) calibrated measures uses IRT standard errors at time 1 and time 2 [14, 15]. The coefficient of repeatability indicates the amount of change necessary to be significant on the RCI and is, therefore, equivalent to it. This coefficient is also known as the minimally detectable change, smallest real difference, and the smallest detectable change [16].

Variations to these methods have been proposed to account for regression to the mean (see Table 1). Regression-based approaches compare observed scores at time 2 with regression predicted scores based on time 1 score and other time 1 variables. This can be useful clinically because time 2 status is compared to what would be expected based on time 1 characteristics.

MIC thresholds should not be used to identify responders to treatment

There are two major problems with applying group-based MIC methods to categorize individual patients as having changed or not: one conceptual and one statistical. The conceptual issue regards using averages derived from groups that may not be relevant to any one patient. MIC estimates are averages from distributions of individual MICs; small changes may be meaningful for some and large changes for others [17, 18]. Even if such MIC estimates are derived from patient-reported anchors representing the construct of interest, these averages may not represent change that is meaningful to individuals. For example, an individual patient may consider only a large magnitude improvement in physical function to be meaningful and be uninterested in achieving the average improvement, since the average value falls below that individual's perception of meaningful change. The statistical issue is that group-based MIC methods drastically underestimate the amount of change needed to be significant at the individual level due to the large measurement error around individual change scores [19]. "Any inspection of measured data reveals an order of magnitude difference between the variability in group versus individual changes" [20]. Thus, group-based MIC estimates will often be indistinguishable from individual measurement error [21].

Abu et al. [22] is a recent example of using MIC thresholds to identify whether patients improved or declined on the Atrial Fibrillation Effect QualiTy-of-Life (AFEQT) Questionnaire. A five-point change threshold was used as the threshold for "clinically meaningful change". This threshold was based on group-level MIC estimates from a prior study of the AFEQT MIC that used physician assessment of functional status [23]. The authors concluded that 22% declined and 40% improved from baseline to 1 year later in a sample of 1097 older adults with atrial fibrillation. Table 2 shows the standard deviations, internal consistency reliabilities, and coefficients of repeatability for the four AFEQT scores. The coefficients of repeatability are two-to-three times larger than the 5-point change threshold the authors used. Ironically, Abu et al. could have adopted the more appropriate SDC estimates (equal to the coefficient of repeatability) reported by Spertus et al. [23].

The Abu et al. [22] paper is one where the MIC, derived from group-based estimates, falls well below the coefficient of repeatability. When this is the case, Kemmler et al. [21] suggest increasing the MIC thresholds to the coefficients of repeatability. Terwee et al. [16] recommend looking to see how measurement error might be reduced by: (1) increasing homogeneity of the study sample's scores at the first measurement timepoint and thereby reducing the SD; and/or (2) increasing the reliability of the measure. Both options are made difficult if the amount of SD reduction or reliability increase is not trivial.

Using the Abu et al. [22] example, we calculated and plotted the SD's needed at 0.90, 0.95 and 0.99 reliability on the AFEQT. Figure 1 uses the approximate SD (~ 17.5) and coefficient of repeatability (~ 15) observed for the AFEQT overall scale at 90% as a starting point so that scenarios under which the reliability is increased or the SD is decreased can be examined. As seen in the plot, at 0.90 reliability the SD must drop to about 5 for the coefficient of repeatability to equal the MIC. If the reliability were 0.99,

Table 2 Amount of change in atrial fibrillation effect on QualiTy-of-Life (AFEQT) scores needed for significant individual change (coefficient of repeatability)

	Overall score	Symptoms	Daily activities	Treatment concerns
Standard deviation	17.8	17.5	24.5	19.3
Internal consistency reliability	0.90*	0.95	0.94	0.90
Coefficient of repeatability	15.6	10.8	16.6	16.9

*Exact reliability not reported in Abu et al. [22] so we estimated this from prior work[23]

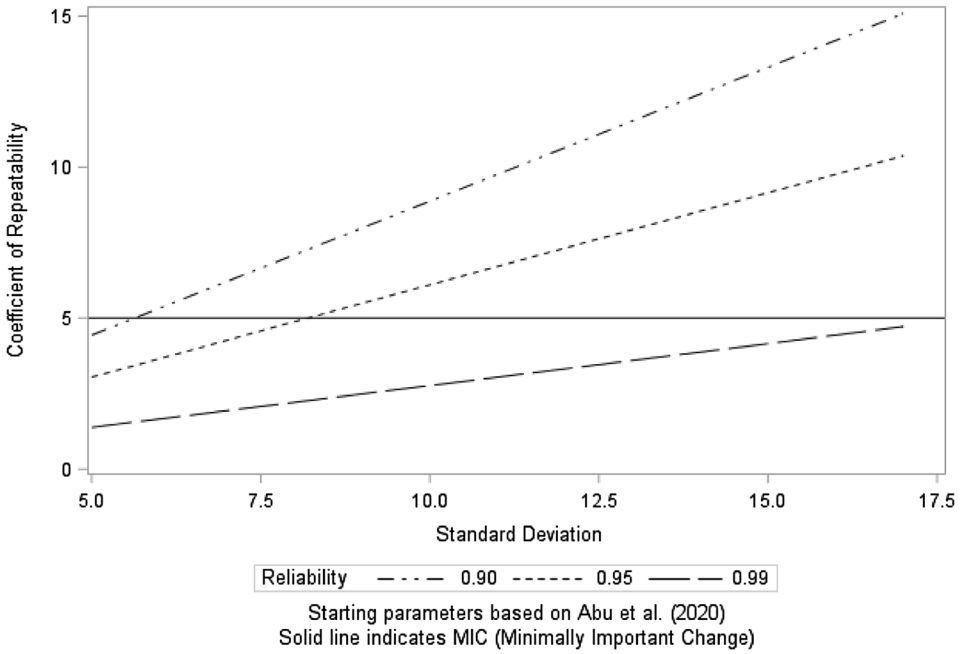

Fig. 1 Needed change in measurement error parameters for coefficient of repeatability equal to MIC on Atrail fibrillation effect on QuailTy-of-Life (AFEQT)

SDs under 17.5 would result in a coefficient of repeatability at or less than the MIC. This example demonstrates the types of conditions required for an instrument's coefficient of repeatability to equal its MIC. Many instruments will not achieve such low SD's or high reliabilities under any circumstances. Note that the coefficient of repeatability does not provide information about the amount of change that is meaningful to an individual.

Statistically significant and meaningful individual change

A clinician or researcher might also regard relative standing on the measure at the follow-up time point to be important. In some areas of medicine, change in clinical status alone is enough to be important. For example, COVID-19 patients who changed to a more positive level on a six-point ordinal scale (not hospitalized; hospitalized but not requiring supplemental oxygen; hospitalized, requiring supplemental oxygen; hospitalized, requiring nasal high-flow oxygen therapy, non-invasive ventilation, or both; hospitalized, requiring invasive mechanical ventilation, extracorporeal membrane oxygenation, or both; dead) were regarded as improved in one study [24]. Or, a primary care physician might be interested in whether a patient ends up within the normal blood pressure range following initiation of high blood pressure medicine. Similarly, a rehabilitation clinician might want to know if a patient with impaired physical functioning at the beginning of treatment ends up functioning as well as other people with a similar condition. The FDA has suggested that meaningful change needs to be assessed in addition to significant individual change [1]. Some contend that any individual change that is significant at $p < 0.05$ is substantial and likely to be meaningful to patients [10, 25].

Several years ago, Jacobson and Truax [26] made suggestions for how both significant and clinically meaningful change might be used together. They classified change as (1) "recovered" (statistically significant and clinically significant); (2) improved (statistically significant but not clinically significant); (3) unchanged (not statistically significant), and (4) deteriorated (statistically significant decrement). In one study, responders were those with significant individual improvement on the Functional Disability Inventory (FDI) and improvement in the FDI severity level (no/minimal disability, moderate disability, severe disability) [27]. These change categories offered by Jacobson and Truax may be more appealing than use of either statistically significant change (coefficient of repeatability) or the MIC alone.

Secondary analysis considering significant and meaningful individual change

To illustrate how significant individual change and meaningful individual change can be presented together, we conduct a secondary analysis of the Impact Stratification Score (ISS) administered in a prospective comparative effectiveness clinical trial of 750 active-duty U.S. military personnel [28]. The average age of the sample was 31; 76% were males and 67% white. Most of the participants reported low back pain for more than 3 months.

The ISS was proposed for use with chronic low back pain patients by a National Institutes of Health Pain Consortium research task force. The ISS is the sum of the PROMIS-29 v2.1 physical function, pain interference and pain intensity scores [29]. The ISS has a possible range of 8 (least impact) to 50 (greatest impact). Physical function (4 items with response options ranging from *without any difficulty* = 1 to *unable to do* = 5) and pain interference (4 items with response options ranging from *not at all* = 1 to *very much* = 5) each contribute from 4 to 20 points, and the pain intensity item contributes from 0 to 10 points. The task force proposed three categories of ISS severity: 8–27 (mild), 28–34 (moderate), and 35–50 (severe).

Following guidelines by de Vet et al. [30], Dutmer et al. [7] estimated a SEM of 5.2 for the ISS based on test–retest reliability. But test–retest reliability estimates can be problematic. Test–retest reliability can underestimate reliability when there is true underlying change. Reeve et al. [31] noted that:

> ISOQOL respondents agreed that as a minimum standard a multi-item PRO measure should be assessed for internal consistency reliability…
> However, they did not support as a minimum standard that a multi-item PRO measure should be required to have evidence of test–retest reliability. They noted practical concerns regarding test–retest reliability; primarily that some populations studied in PCOR are not stable and that their HRQOL can fluctuate. This phenomenon would reduce estimates of test–retest reliability, making the PRO measure look unreliable when it may be accurately detecting changes over time. In addition, memory effects will positively influence the test–retest reliability when the two survey points are scheduled close to each other (p. 1895).

We estimated a much smaller SEM of 2.4 using an internal consistency reliability estimate from another study [28]. In this dataset, we examine significance of individual change on the ISS between baseline and 6 weeks later using the coefficient of repeatability (= 6.6). In addition, we compare the significance of change with self-reports on a retrospective rating of change item administered at 6 months: "Compared to your first visit, your low back pain is: *much worse*, *a little worse*, *about the same*, *a little better*, *moderately better*, *much better* or *completely gone*"?

Thirty-seven percent of the sample improved significantly on the ISS over these 6 weeks and 59% reported on the retrospective change item that they were better (16% a little better, 14% moderately better, 23% much better, and 6% completely gone). Among those who improved significantly on the ISS, 89% reported they were better on the retrospective rating item. Thirty-three percent of the sample improved significantly and reported improvement on the retrospective change item (statistically and clinically meaningful), 4% improved significantly but did not report that they were better on the retrospective change item (statistically but not clinically meaningful), 26% did not improve significantly but reported improvement on the change item, and 37% did not improve significantly or report improvement on the change item.

Extending this application to further illustrate how group-based methods of estimating MICs can underestimate significant individual change, we compared two alternative ways of defining improvement on a retrospective rating of change item to identify optimal cut points on the ISS. The first way is more inclusive in that improvement from baseline to 6 weeks later included those who reported on the retrospective change item at 6 weeks that their back pain was either *a little better*, *moderately better*, *much better* or *completely gone*. The second way is more restrictive as improvement was limited to those who reported their back pain was *moderately better*, *much better* or *completely gone* on the retrospective change item.

The Youden index [32], (sensitivity + specificity) −1, suggested an optimal cut point of 5 points for change on ISS from baseline to 6 weeks later for the first definition of improvement: sensitivity of 65%, specificity of 82%, negative predictive value of 62%, and positive predictive value of 84%. For the second definition of improvement, the Youden index indicated an optimal cut point of 7 points for ISS change: sensitivity of 66%, specificity of 85%, negative predictive value of 77%, and positive predictive value of 76%. The group-level thresholds estimated for the second definition that excluded those who said they were *a little better* from the improvement group were closer to the coefficient of repeatability.

Discussion

In contrast to significant group-level change that can be trivial in magnitude if the sample size is large, significant individual change is substantial and worth noting regardless of whether the patient reports that they have improved. As suggested by Jacobson and Truax [26], researchers and clinicians may also be interested in whether those who have significantly improved on a HRQOL measure perceive that they have done so. One can separate people who improved significantly and report at time 2 that they have improved since time 1 from those who do not perceive they have improved. One could also note who reaches a desirable status such as becoming symptom free or ending up within the "normal" range at time 2.

Using group-level estimates of meaningful change (group means) to classify individuals as responders to treatment is inappropriate. Doing so results in overoptimistic estimates

of the number of people who improve (i.e., too many will be classified as improved). Ironically, a MIC estimate might yield similar numbers of responders as individual-level significance tests if the estimate erroneously includes those who changed by more than a minimally important amount [33]. In our secondary analysis we observed that the optimal cut-point on the ISS using one way of classifying improvement (i.e., those who reported that they were *moderately better*, *much better* or their back pain was *completely gone*) over 6 weeks was similar to the coefficient of repeatability for individual change. But including people who felt they were a little better as improvers resulted in an overoptimistic number of responders. Future work is needed to investigate whether group-level threshold estimates based on retrospective ratings of more than a little improvement converge with appropriate individual-level significance tests.

A fundamental criterion for a responder is that the individual improves significantly (i.e., individual change is greater than estimated measurement error). Individual-level statistical indices such as the RCI or the equivalent coefficient of repeatability have been available for decades. These or parallel item response theory approaches [15, 34] that allow reliability to vary across the true score continuum need to be used to determine if patients have stayed the same, deteriorated, or improved.

Some may argue for using a significance level other than $p < 0.05$ to identify individual change that doesn't meet the conventional cutoff used for group-level comparisons. Individual differences may be important even if they do not equal or exceed the conventional 0.05 significance level. One possible strategy is to use a combination of one-tailed and two-tailed tests of significance and report five levels of change: *definitely worse* (two-tailed), *probably worse* (one-tailed), *same* (one-tailed), *probably better* (one-tailed), and *definitely better* (two-tailed). This classification preserves more information and, therefore, helps to address to some extent concerns about missing noteworthy individual change Others might favor even more liberal significance levels to capture more potential responders. Indeed, Donaldson [20] entertained focusing on likely instead of unlikely values and classifying individuals into categories such as: *almost certainly changed*, *quite likely changed*, and *probably stayed the same*. Others [35] have suggested using a Bayes factor to indicate the evidence in the data for or against true change but note that it requires at least three data points. More work is needed about what cutoffs should be used to identify important individual change.

Variance between individuals was used in previous research and in the examples in this paper. As noted earlier, it would be ideal if individual-level variation was available, but the number of observations required to do this is prohibitive for most research studies [8]. The SD of change is more consistent epistemologically with the assessment of individual change and is analogous to the denominator used for the standardized response mean and the responsiveness statistic when evaluating responsiveness of measures [36]. Future research is needed that compares SD between individuals with the SD of change. In addition, different ways of identifying meaningful change need to be examined and group-level threshold cutoff approaches (e.g., area under the curve) compared with use of individual ratings of change.

We recommend that investigators conducting clinical trials and observational studies routinely report responders to treatment using the significance of individual change. Presenting individual statistical significance and whether the individual feels that they have improved together could be especially useful. We also suggest that clinicians evaluate statistically significance and meaningful change in their patients.

Author contributions RDH wrote the first draft and JDP provided edits to it.

Funding Hays received partial funding support from the University of California, Los Angeles (UCLA), Resource Centers for Minority Aging Research Center for Health Improvement of Minority Elderly (RCMAR/CHIME) under NIH/NIA Grant P30-AG021684. Dr. Peipert received partial funding support from the National Institute on Aging (P30-AG059988).

Declarations

Conflicts of interest The author declare that they have no conflict of interest.

References

1. FDA. (2018). Patient-focused drug development guidance public workshop. Methods to identify what is important to patients and select, develop or modify fit-for-purpose clinical outcomes assessments. https://www.fda.gov/media/116277/download. Accessed 4 Nov 2020
2. Coon, C. D., & Cook, K. F. (2018). Moving from significance to real-world meaning: methods for interpreting change in clinical outcome assessment scores. *Quality of Life Research, 27*(1), 33–40. https://doi.org/10.1007/s11136-017-1616-3
3. Schwartz, N., & Sudman, S. (1994). *Autobiographical memory and the validity of retrospective reports*. New York: Springer-Verlag.
4. Norman, G. R., Stratford, P., & Regehr, G. (1997). Methodological problems in the retrospective computation of responsiveness to change: The lesson of Cronbach. *Journal of Clinical Epidemiology, 50*(8), 869–879. https://doi.org/10.1016/S0895-4356(97)00097-8
5. Hays, R. D., & Reeve, B. B. (2010). Measurement and modeling of health-related quality of life. In J. Killewo, H. K. Heggenhougen, & S. R. Quah (Eds.), *Epidemiology and demography in public health* (pp. 195–205). Netherlands: Elsevier.
6. Yuan, L., Zeng, Y., Chen, Z., Li, W., Zhang, X., & Ni, J. (2020). Risk factors associated with failure to reach minimal clinically

important difference after correction surgery in patients with degenerative lumbar scoliosis. *Spine, 45*(24), E1669–E1676. https://doi.org/10.1097/BRS.0000000000003713
7. Dutmer, A. L., Reneman, M. F., Schiphorst Preuper, H. R., Wolff, A. P., Speijer, B. L., & Soer, R. (2019). The NIH minimal dataset for chronic low back pain: Responsiveness and minimal clinically important change. *Spine, 44*(20), E1211–E1218. https://doi.org/10.1097/BRS.0000000000003107
8. Borckardt, J. J., Nash, M. R., Murphy, M. D., Moore, M., Shaw, D., & O'Neil, P. (2008). Clinical practice as natural laboratory for psychotherapy research: A guide to case-based time series analysis. *American Psychologist, 63*(2), 77–95. https://doi.org/10.1037/0003-066X.63.2.77
9. Moinpour, C. M., Donaldson, G. W., et al. (2017). The challenge of measuring intra-individual change in fatigue during cancer treatment. *Quality of Life Research, 26*(2), 259–271. https://doi.org/10.1007/s11136-016-1372-9
10. Hays, R. D., Brodsky, M., Johnson, M. F., Spritzer, K. L., & Hui, K. K. (2005). Evaluating the statistical significance of health-related quality of life change in individual patients. *Evaluation and the Health Professions, 28*(2), 160–171. https://doi.org/10.1177/0163278705275339
11. Duff, K. (2012). Evidence-based indicators of neuropsychological change in the individual patient: Relevant concepts and methods. *Archives of Clinical Neuropsychology, 27*(3), 248–261. https://doi.org/10.1093/arclin/acr120
12. Bruggemans, E. F., Van de Vijver, F. J., & Huysmans, H. A. (1997). Assessment of cognitive deterioration in individual patients following cardiac surgery: Correcting for measurement error and practice effects. *Journal of Clinical and Experimental Neuropsychology, 19*(4), 543–559. https://doi.org/10.1080/01688639708403743
13. Maassen, G. H. (2000). Principles of defining reliable change indices. *Journal of Clinical and Experimental Neuropsychology, 22*(5), 622–632. https://doi.org/10.1076/1380-3395(200010)22:5;1-9;FT622
14. Jabrayilov, R., Emons, W. H. M., & Sijtsma, K. (2016). Comparison of classical test theory and item response theory in individual change assessment. *Applied Psychological Measurement, 40*(8), 559–572. https://doi.org/10.1177/0146621616664046
15. Brouwer, D., Meijer, R. R., & Zevalkink, J. (2013). Measuring individual significant change on the beck depression inventory-II through IRT-based statistics. *Psychotherapy Research, 23*(5), 489–501. https://doi.org/10.1002/mpr.1496
16. Terwee, C. B., Terluin, B., Knol, D. L., & de Vet, H. C. W. (2011). Combining clinical relevance and statistical significance for evaluating quality of life changes in the individual patient. *Journal of Clinical Epidemiology, 64*(12), 1465–1467. https://doi.org/10.1016/j.jclinepi.2011.06.015
17. Ingelsrud, L. H., Roos, E. M., Terluin, B., Gromov, K., Husted, H., & Troelsen, A. (2018). Minimal important change values for the Oxford knee score and the forgotten joint score at 1 year after total knee replacement. *Acta Orthopaedica*, *89*(5), 541–547. https://doi.org/10.1080/17453674.2018.1480739.
18. Terluin, B., Eekhout, I., & Terwee, C. B. (2017). 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. *Journal of Clinical Epidemiology, 83*, 90–100. https://doi.org/10.1016/j.jclinepi.2016.12.015
19. Lord, F. M. (1963). Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring change.*Madison: University of Wisconsin Press.
20. Donaldson, G. (2008). Patient-reported outcomes and the mandate of measurement. *Quality of Life Research, 17*(10), 1303–1313. https://doi.org/10.1007/s11136-008-9408-4
21. Kemmler, G., Zabernigg, A., Gattringer, K., Rumpold, G., Giesinger, J., Sperner-Unterweger, B., et al. (2010). A new approach to combining clinical relevance and statistical significance for evaluation of quality of life changes in the individual patient. *Journal of Clinical Epidemiology, 63*(2), 171–179. https://doi.org/10.1016/j.jclinepi.2009.03.016
22. Abu, H. O., Saczynski, J. S., Mehawej, J., Tisminetzky, M., Kiefe, C. I., Goldberg, R. J., et al. (2020). Clinically meaningful change in quality of life and associated factors among older patients with atrial fibrillation. *Journal of the American Heart Association, 9*(18), e016651. https://doi.org/10.1161/JAHA.120.016651
23. Spertus, J., Dorian, P., Bubien, R., Lewis, S., Godejohn, D., Reynolds, M. R., et al. (2011). Development and validation of the Atrial fibrillation effect on QualiTy-of-Life (AFEQT) questionnaire in patients with atrial fibrillation. *Circulation Arrhythmia and Electrophysiology, 4*(1), 15–25. https://doi.org/10.1161/CIRCEP.110.958033
24. McElvaney, O. J., Hobbs, B. D., Qiao, D., McElvaney, O. F., Moll, M., McEvoy, N. L., et al. (2020). A linear prognostic score based on the ratio of interleukin-6 to interleukin-10 predicts outcomes in COVID-19. *eBioMedicine, 61*, 103026. https://doi.org/10.1016/j.ebiom.2020.103026
25. King, M. T., Dueck, A. C., & Revicki, D. A. (2019). Can methods developed for interpreting group-level patient-reported outcome data be applied to individual patient management? *Medical Care, 57*(Suppl 51), S38–S45. https://doi.org/10.1097/MLR.0000000000001111
26. Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*(1), 12–19. https://doi.org/10.1037//0022-006x.59.1.12
27. Sil, S., Arnold, L. M., Lynch-Jordan, A., et al. (2014). Identifying treatment responders and predictors of improvement after cognitive-behavioral therapy for juvenile fibromyalgia. *Pain, 155*(7), 1206–1212. https://doi.org/10.1016/j.pain.2014.03.005
28. Goertz, C. M., Long, C. R., Vining, R. D., Pohlam, K. A., Kane, B., Corber, L., et al. (2016). Assessment of chiropractic treatment for active duty, U.S. military personnel with low back pain: Study protocol for a randomized controlled trial. *Trials, 17*, 70. https://doi.org/10.1186/s13063-016-1193-8
29. Deyo, R. A., Dworkin, S. F., Amtmann, D., et al. (2014). Report of the NIH Task Force on research standards for chronic low back pain. *Pain Medicine, 15*(6), 569–585. https://doi.org/10.1016/j.jpain.2014.03.005
30. de Vet, H. C. W., Terwee, C. B., Knol, D. L., & Bouter, L. M. (2006). When to use agreement versus reliability measures. *Journal of Clinical Epidemiology, 59*(10), 1033–1039. https://doi.org/10.1016/j.jclinepi.2005.10.015
31. Reeve, B. B., Wyrwich, K. W., Wu, A. W., Velikova, G., Terwee, C. B., Synder, C., et al. (2013). ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. *Quality of Life Research, 22*(8), 1889–1905. https://doi.org/10.1007/s11136-012-0344-y
32. Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer, 3*, 32–35.
33. Yuksel, S., Ayhan, S., Nabiyev, V., Domingo-Sabat, M., Vila-Casademunt, A., Obeid, I., et al. (2019). Minimum clinically important difference of the health-related quality of life scales in

adult's deformity calculated by latent class analysis: Is it appropriate to use the same values for surgical and nonsurgical patients? *The Spine Journal, 19*(1), 71–78. https://doi.org/10.1016/j.spinee.2018.07.005

34. Hays, R. D., Spritzer, K. L., & Reise, S. P. (in press). Using item response theory to identify responders to treatment: Examples with the patient-reported outcomes measurement information system (PROMIS®) physical functioning and emotional distress scales. *Psychometrika*
35. De Vries, R., Meijer, R. R., Van Bruggen, V., & Morey, R. D. (2016). Improving the analysis of routine outcome measurement data: What a Bayesian approach can do for you. *International Journal of Methods in Psychiatric Research, 25*(3), 155–167. https://doi.org/10.1002/mpr.1496
36. Guyatt, G., Walter, S., & Norman, G. (1987). Measuring change over time: Assessing the usefulness of evaluative instruments. *Journal of Chronic Diseases, 40*, 171–178.