Key Issues in Identifying Responders to Treatment

Ron D. Hays, PhD[1] and John Devin Peipert, PhD[2]

[1]UCLA Department of Medicine, Division of General Internal Medicine & Health Services Research
[2]Northwestern University Feinberg School of Medicine

Correspondence: and reprint requests:

Ron D. Hays, Ph.D.
UCLA Division of General Internal Medicine & Health Services Research
1100 Glendon Avenue, Suite 850
Los Angeles, CA 90024
310-794-2294; 310-794-0732 FAX
drhays@ucla.edu
ORCID: 0000-0001-6697-907X

Declarations

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.

Conflicts of interest/Competing interests: None

Ethics approval: N/A

Consent to participate: N/A

Availability of data and material: N/A

Code availability: N/A

Authors' contributions: RDH wrote the first draft and JDP provided edits to it.

Word count: 2,285

Hays, R. D., & Peipert, J. (2020). Key issues in identifying responders to treatment.

Preprint at https://labs.dgsom.ucla.edu/hays/pages/favorite_links

Abstract

Purpose. Estimates of the minimally important difference (MID) for between group comparisons or minimally important change (MIC) over time can be used to evaluate whether group-level differences are large enough to be important. Responders to treatment are too often based upon group-level MIC thresholds. Improper use of the MIC leads to inaccurate classification of change over time. This article reviews options for individual-level statistics to assess whether individuals have improved, stayed the same, or declined.

Methods. Review of MID and MIC concepts. Illustrative examples of misapplication of MIC group-level estimates to assess individual change. Secondary data analyses.

Results. MIC thresholds are shown to yield over-optimistic conclusions (i.e., classify those who have not changed as responders to treatment). Proper individual change statistics can be used along with individual retrospective ratings of change.

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.

Keywords. Meaningful change; minimally important difference; responder; reliable change index

Changes in health-related quality of life (HRQOL) vary in magnitude. Trivial group-level differences or change within groups can be statistically significant if the sample size is large enough. Estimates of the minimally important difference (MID) for between group comparisons or minimally important change (MIC) over time can be used to evaluate whether statistically significant differences are large enough to be important or meaningful to patients or clinicians. Because the MIC is a subset of responsiveness to change [1], it is informative to compare change for those hypothesized to have smaller or larger changes than the MIC. There should be a monotonic association whereby those who change the most on an anchor have the largest average change on the target measure and those that do not change according to the anchor have minimal or no change on the target measure.

Change in physical function, for example, would vary for people hit by a feather, rock, bicycle or car. The average change in physical function for those hit by a car would not be small enough to provide a good estimate of the MIC. Those hit by a feather would not be useful either because no change in physical function would be expected. Best suited for estimating the MIC might be those hit by a rock, assuming change in physical function is expected to be important but not substantial [2]. An example of failure to limit MIC estimates to those with minimal improvement was the use of average HRQOL change in 123 adult surgical patients with adult spinal deformity among those who reported any improvement, including large improvements, on a retrospectively rating of change item [3]. Similarly, Dutmer et al. [4] estimated the MIC by including all patients who improved rather than limiting it to those with minimal

change. Including all those who change rather than focusing on those with minimal (but important) change leads to MIC estimates that are too large.

Inappropriate Uses of MIC to Identify Individual Change

MIC thresholds have been used erroneously to identify “responders” to treatment. It is inappropriate to apply the group-level concept of important change to individual change. First, as noted above, at the group-level, additional information is needed to supplement interpretation of significant change because trivial changes can be statistically significant with large sample sizes. However, at the individual level, substantial (not trivial) changes are needed to obtain statistical significance. Therefore, any change that is statistically significant at the individual level is very likely to be important or meaningful clinically and to the patient [5]. Second, since the threshold for significant group-level change is smaller than significant individual-level change, using group-level MIC estimates to make inferences about individual change leads to misclassification of patients who have not changed as getting better or worse [6]. Examples of this problem are studies where investigators inappropriately used MIC thresholds to classify significant improvement after corrective surgery for 123 adults with degenerative lumbar scoliosis [3] and to estimate responder thresholds for a Patient-reported Outcomes Measurement Information System (PROMIS®) physical function measure using Phase 3 data from the ENLIVEN trial [7]. Responder is an individual change concept that requires use of appropriate individual-level statistics.

While individual-level variation can be estimated by single-case time-series approaches when HRQOL has been assessed at several time points [8], most longitudinal HRQOL studies are limited to two time points. Thus, we focus on individual

change for two time points, and, for simplicity, $p < .05$ (two-tailed) significance testing. Table 1 provides several formulae used for estimating the significance of individual change.

The simplest approach is to subtract the time 1 score from the time 2 score and divide by the time 1 standard deviation (standard deviation index) and define responders by a z-score of 1.96 or larger [9]. Another approach is to use the confidence interval around the time 1 score based on the standard error of measurement (SEM) as was done in the Medical Outcomes Study [10]. Participants were categorized as not changing (time 2 score fell within the 95% confidence interval of the time 1 score), improved (time 2 score exceeded the upper bound of the time 1 95% confidence interval), or declined (time 2 score was less than the lower bound of the time 1 95% confidence interval). But setting confidence intervals around observed scores is discouraged because of regression to the mean. Rather, the standard error of the estimation is recommended so that inferences can be drawn about “true” rather than observed scores [11]. The standard error of estimation has been referred to as the standard deviation of the observed scores when the true score is held constant. The standard error of prediction is designed to evaluate whether a second assessment is beyond measurement error based on whether it is outside of the confidence interval of what would be expected on a retest.

The reliable change index (RCI) is the most used way to identify individual change. Responders are defined by those with an RCI of 1.96 or larger (or improve at least as much as the equivalent coefficient of repeatability). A group-level version of the coefficient of repeatability has been proposed by dividing the formula in Table 1 by the

$\sqrt{n}$ [3]. A variant of the RCI used for cognitive measures corrects for practice effects [10]. The denominator of the RCI for measures calibrated using item response theory (IRT) has IRT standard errors at time 1 and time 2 [12].

Regression-based approaches compare observed scores at time 2 with regression predicted scores based on time 1 score and other time 1 variables. This approach can be useful clinically because it compares where someone is at time 2 to what would be expected based on time 1 characteristics. As with any derivation and application of regression equations, it is important to cross-validate and account for regression to the mean.

A study that used MIC thresholds to identify whether patients improved or declined on the Atrial Fibrillation Effect QualiTy-of-Life (AFEQT) Questionnaire provides an illustration of problems with this approach [13]. Group-level MIC estimates were used based on estimates from a prior study of the AFEQT MIC from physician assessment of functional status [14]. 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 that coefficients of repeatability are two-to-three times larger than the 5-point change threshold used. Ironically, the publication the authors relied on for the MIC threshold [14] also included appropriate minimally detectable change estimates (equivalent to the coefficient of repeatability).

In the study of degenerative lumbar scoliosis noted above, the SF-36 physical and mental health summary scores were extremely reliable (0.96 in [3] and 0.93-0.94 in [5]), so the minimum detectable change estimates for these HRQOL scores were both about a half-standard deviation.

Use of "Meaningful" Change When Identifying Responders

The FDA and leading HRQOL researchers suggest that meaningful change needs to be assessed in addition to significant individual change [15-16]. Any change that is significant at the individual-level is substantial, but a clinician or researcher might also regard relative standing on the measure at the follow-up time point to be important. For example, 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. Guidelines published for the RAND-36 Health Status Inventory segment significant positive change into 1) positive, but insufficient; (2) favorable; (3) very favorable; or (4) optimal [17]. 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, ECMO, or both; dead) were regarded as improved in one study [18]. But responders in another study were defined by having significant individual improvement on the Functional Disability Inventory (FDI) and improvement in the FDI severity level (no/minimal disability, moderate disability, severe disability) [19].

A National Institutes of Health Pain Consortium research task force proposed an Impact Stratification Score (ISS) for chronic low back pain that is the sum of the

PROMIS-29 physical function, pain interference and pain intensity scores [20]. 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 0-10 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. [21], Dutmer et al. [4] estimated a SEM of 5.2 based on test-retest reliability of the ISS. But test-retest reliability estimates can be problematic. Test-retest reliability can underestimate reliability when there is true underlying change. We estimated a much smaller SEM of 2.4 using internal consistency reliability from other data [22]. In the same dataset, we examined the significance of change between baseline and 6 months later on the ISS using the coefficient of repeatability (= 6.6). We also compared 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 months 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 improved on the retrospective rating item. Thirty-three percent of the sample improved significantly and reported improvement on the retrospective change item.

We also estimated the optimal cut-point on the ISS for identifying improvement from baseline to 6 weeks later. We first defined improvement as reporting on the retrospective change item at 6 weeks that one's back pain was either *a little better*, *moderately better, much better* or *completely gone*. Next, we defined improvement as reporting that back pain was *moderately better, much better* or *completely gone* on the retrospective change item. The Youden index suggested an optimal cut point of 5 points for change on ISS from baseline to 6 weeks later for the first definition: 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 change on ISS: sensitivity of 66%, specificity of 85%, negative predictive value of 77%, and positive predictive value of 76%. Thus, 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 same size is large, significant individual change is substantial and likely important in and of itself. Looking at individual perceptions of change and individual status at follow-up may be valuable in addition to significant individual change. 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 report improvement. In addition, one could report the number of individuals that

reach a desirable status such as no or mild symptoms or is within the “normal” range at time 2.

Using group-level estimates of meaningful change as a basis for determining individual change and identifying responders to treatment is inappropriate. Doing so will result in overoptimistic estimates of the number of people who improve (i.e., too many will be classified as improved). It is possible that apparent use of the MIC could yield similar numbers of responders as proper individual-level change statistics. This can happen when the MIC estimate erroneously includes those who changed by more than a minimally important amount [23]. In our analysis of the ISS we observed that the optimal cut-point for 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. Including people who felt they were a little better as improvers resulted in an overoptimistic number of responders. Future work is needed to investigate if there are conditions when group-level threshold estimates converge with appropriate individual-level significance tests.

Clinical trials and observational studies should routinely report responders to treatment using the significance of individual change. A fundamental criterion for a responder is that the individual improves significantly (i.e., individual change is greater than measurement error) [24]. Improper use of group-level estimates for individual-level decisions needs to cease. Individual-level statistical indices such as the reliable change index or the equivalent coefficient of repeatability have been available for decades. These or parallel item response theory approaches 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.

References

1. 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). Elsevier.
2. Hays, R. D., Farivar, S. S., & Liu, H. (2005). Approaches and recommendations for estimating minimally important differences for health-related quality of life measures. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, *2*, 63-67
3. Yuan, L., Zeng, Y., Chen, Z., Li, W., Zhang, X., & Ni, J. (in press). Risk factors associated with failure to reach minimal clinically important difference after correction surgery in patients with degenerative lumbar scoliosis. *Spine* in press.
4. 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.
5. 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. *Eval Health Prof, 28*(2), 160-171.
6. Hays, R.D., & Peipert, J.D. (2018). Minimally important differences do not identify responders to treatment. *JOJ scim*, *1*(1): JOJS.MS.ID.555552
7. Speck, R.M., Ye, X., Bernthal, N.M., & Gelhorn, H.L. (2020) Psychometric properties of a custom Patient-Reported Outcomes Measurement Information

System (PROMIS) physical function short form and worst stiffness numeric rating scale in tenosynovial giant cell tumors. *J Patient Rep Outcomes, 4*(1):61. PMID: 32676941; PMCID: PMC7366525.

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. *Am Psychol, 63*(2), 77-95.

9. Duff, K. (2012). Evidence-based indicators of neuropsychological change in the individual patient: Relevant concepts and methods. *Archives of Clinical Neuropsychology, 27,* 248-261.

10. Ware, J. E., Bayliss, M. S., Rogers,W. H., Kosinski, M., & Tarlov, A. R. (1996). Differences in 4-year health outcomes for elderly and poor, chronically ill patients treated in HMO and fee-for-service systems: Results from the Medical Outcomes Study. *Journal of the American Medical Association*, *276*, 1039-1047.

11. McManus, I. C. (2012). The misinterpretation of the standard error of measurement in medical education: A primer on the problems, pitfalls and peculiarities of the three different standard errors of measurement. *Medical Teacher, 34*, 569-576.

12. 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.

13. 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*, e016651.

14. 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. *Circ Arrhythm Electrophysiol, 4*(1):15-25.

15. FDA. (2018). Patient-Focused Drug Development Guidance Public Workshop. Methods to Identify What is Important to Patients & Select, Develop or Modify Fit-for-Purpose Clinical Outcomes Assessments. https://www.fda.gov/media/116277/download. Accessed 4 November 2020.

16. Coon, C.D., Cook, K.F. (2018). Moving from significance to real-world meaning: methods for interpreting change in clinical outcome assessment scores. *Qual Life Res* 27(1):33-40. doi: 10.1007/s11136-017-1616-3. Epub 2017 Jun 15. PMID: 28620874.

17. Hays, R. D., Prince-Embury, S., & Chen, H. (1998). *RAND-36 Health Status Inventory*. San Antonio: TX: The Psychological Corporation.

18. 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.

19. 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.

20. Deyo RA, Dworkin SF, Amtmann D, et al. Report of the NIH Task Force on research standards for chronic low back pain. *Pain Med.* 2014;15(8):1249-1267.

21. 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, 1033-1039.

22. 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.

23. Yuksel, S., AyhaA, 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*, 71-78.

24. McLeod, L. D., Coon, C. D., Martin, S. A., Fehnel, S. E., & Hays, R. D. (2011). Interpreting patient-reported outcome results: US FDA guidance and emerging methods. *Expert Rev Pharmacoecon Outcomes Res, 11*(2), 163-169.

Table 1. Formula for Evaluating Individual Change

Statistic	Formula
Standard Deviation Index	$(X_2 - X_1)/\ SD_1$
Standard Error of Measurement (SEM)[b]	$SD_1\ \sqrt{1 - reliability}$
95% Confidence Interval around SEM	X_1[a] +/- $1.96\ SEM$
Standard error of estimation	$SD_1 \sqrt{reliability\ (1 - reliability)}$
Standard error of prediction	$SD_1 \sqrt{1 - reliability^2}$
Reliable Change Index	$(X_2 - X_1)/\ \sqrt{2}\ SEM$
Coefficient of repeatability*	$1.96\ \sqrt{2}\ SEM$
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	$(X_2 - X_{2p})/\sqrt{S_1^2 + S_2^2}\sqrt{1 - reliability}$
X_{2p}	Prediction of time 2 from time 1 variables

[a]Some have suggested that the SD of change be used.

[b]One can use the estimated true score (mean + reliability(t_1-mean) to account for regression to the mean.

*Also known as the smallest detectable change, minimally detectable change or smallest real difference

Table 2. Amount of Change in Atrial fibrillation Effect on QualiTy-of-Life (AFEQT) Scores Needed for Significant Change

	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

*Note: Exact reliability not reported in [13] so we estimated this from prior work [14].