DOI 10.1007/s11136-013-0443-4

Don't middle your MIDs: regression to the mean shrinks estimates of minimally important differences

Peter M. Fayers · Ron D. Hays

Accepted: 21 May 2013 / Published online: 31 May 2013

Abstract Minimal important differences (MIDs) for patient-reported outcomes (PROs) are often estimated by selecting a clinical variable to serve as an anchor. Then, differences in the clinical anchor regarded as clinically meaningful or important can be used to estimate the corresponding value of the PRO. Although these MID values are sometimes estimated by regression techniques, we show that this is a biased procedure and should not be used; alternative methods are proposed.

Keywords Minimally important difference · Clinical significance · Quality of life · Patient-reported outcomes · Regression to the mean

Abbreviations

MCID	Minimal clinically important difference
MID	Minimal important difference
r	Correlation coefficient
NEI VFQ-25	Eye Institute Visual Function Questionnaire-25
PRO	Patient-reported outcome
SD	Standard deviation

P. M. Fayers (✉)
Institute of Applied Health Sciences, University of Aberdeen, Aberdeen, UK
e-mail: P.Fayers@abdn.ac.uk

P. M. Fayers
Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

R. D. Hays
Department of Medicine, UCLA, 911 Broxton Avenue, Los Angeles, CA 90024, USA
e-mail: drhays@ucla.edu

Introduction

To help interpret changes in patient-reported outcomes (PROs), it is useful to establish the amount that is large enough to be discernible and regarded as important. The minimal important difference or *MID*, also in the past sometimes called the minimal clinically important difference or *MCID*, has been defined as "the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management" [1]. The *MID* is the average change among the subgroup of people deemed to change a minimal (but important) amount according to an anchor or anchors. Hence, the *MID* estimate is the expected group mean change for people who have improved enough (but not too much) according to the external standard. The *MID* is different from "*responder*" [2]. If one uses the expected mean change among those who were deemed to have minimally important change to define responders, then about half the people in the group who changed a minimally important amount on the anchor will be classified as responders, assuming a normal distribution. The problem with this is that the group average is not an appropriate threshold for individual change. Group change and individual change have different standard errors, and thus group-level estimates should not be used to define responders. [3].

One method for estimating the *MID* is to take an initial or baseline assessment followed by a later assessment of the PRO, with respondents being asked after the second assessment whether and how much their condition has changed. This "global rating of change" (also known as a transition rating) can be compared with the observed change in the PRO to estimate the *MID*. Such patientfocused anchors are widely used despite criticisms of potential bias due to, for example, response shift [4], recall bias [5] and implicit theories of change [6].

Because of these criticisms, it has been suggested that alternative anchors be considered, instead of or in addition to global ratings of change [7, 8]. These anchors may be clinical endpoints, change in other PRO measures, or some combination of clinical and patient-based measures, and should be anchors for which observers or preferably patients are able to specify what constitutes an important difference. They should measure a similar construct to that of the target, and value of these anchors depends on how well they reflect underlying change. Cohen's [9] rules of thumb suggest 0.8 is a large effect size for comparing two independent groups, and since the equation relating correlations (r) to group mean differences (d) is $d = 2r/\sqrt{1 - r^2}$, derived $r > 0.371$ as a large effect size for a correlation coefficient; thus Hays et al. [7] recommend 0.371 as a correlation threshold to define a noteworthy (large effect) association between anchors and observed change on the PRO. Hays et al. also recommend that there should be multiple anchors and that the correlations should be reported. Then, patients can be classified into change groups according to the anchor and the observed changes on the PRO used to estimate the *MID*. However, Hays et al. only hinted at the method of analysis. The purpose of this statistical note is to explore the approach that has been used in published reports, and to make recommendations about the correct way to link changes in the anchor to the corresponding changes in the target outcome.

Example of an anchor-based *MID*

We take as an example the paper by Suňer et al. [10], who followed conventional methods of analysis. They used changes in visual acuity as an anchor, to propose a *MID* for the National Eye Institute Visual Function Questionnaire-25 (NEI VFQ-25). The authors note that a 15-letter change in best-corrected visual acuity is frequently used as a primary endpoint in clinical trials and is generally accepted as clinically significant. Thus, they used 15-character changes in visual acuity as the anchor for determining the *MID* for the corresponding changes in the overall composite score of the VFQ-25, formed by the mean of 24 items (excluding the single item for general health). The mean visual acuity (letter count) was 53.5, with standard deviation (SD) of 13.2, and the mean score for the VFQ was 69.3, with SD of 19.2. Patients were assessed again at 12 months. The investigators grouped patients into those who gained at least 15 letters, had less than 15 letter change, or lost at least 15 letters. Fitting linear regression models, they estimated a *MID* of 4.34 based on mean change on the VFQ-25 composite score for those with at least a 15-letter change in visual acuity.

One problem with the approach used in this example is that all change that was equal to or greater than the designated important change on the anchor was treated equally in the analysis. Hence, the authors lumped smaller and larger amounts of change together. This will inflate the estimated *MID*. The *MID* is best estimated by honing in on the change group that has improved by a non-trivial important amount but not by a medium or large amount. Although this or similar approaches are frequently used, here we focus on another issue.

The *MID* reported corresponds to an effect size of 4.34/19.2 = 0.22 SDs which, applying Cohen's guide, is a small effect. In contrast, the anchoring outcome has an effect size of 15/13.2 = 1.1 SDs, which according to Cohen's rules is a large effect (although it should be noted that 15 is simply the threshold and some patients will have had scores substantially larger than this) [9]. How can we explain this apparent discrepancy?

Impact of correlation

In the above example, the authors, following conventional practice, analysed group mean scores and used linear regression models. However, the correlation, r, between the anchoring variable and the target score affects the slope of the regression line. Here, the authors reported the correlation to be less than 0.3 (exact value is not specified), and since r^2 is a measure of the proportion of variance explained, this indicates that the anchor variable can account for less than 10 % of the variation in the target variable. The value $r = 0.3$ also implies that the slope of the regression line will be shallow, because for regression analysis the slope, b, is given by $b = r \times \left(\mathrm{SD_{Target}}/\mathrm{SD_{Anchor}}\right)$, and the smaller the value of r, the smaller the corresponding value of b. Thus, if there were perfect correlation with $r = 1.0$, the regression would have slope $= b = \left(\mathrm{SD_{Target}}/\mathrm{SD_{Anchor}}\right)$. Then, a standardised effect size of N standard deviations in the anchor would equate to a *MID* of N standard deviations in the target. Similarly, a correlation of zero leads to a slope of $b = 0.0$, and then the best estimate of the *MID* for the target would be zero, because the anchor is totally uninformative. Thus, with r of any intermediate value between 0 and 1, any specified effect size in the anchor will result in an attenuated estimate of the target-variable *MID*, and the degree of attenuation relates to the correlation coefficient. In the case of Suňer et al., we know $r < 0.3$, and so the *MID* is less than a third of the effect size specified for the anchor.

According to the suggestions of Cohen, an effect size of 0.5 SD is a medium effect. Let us therefore consider an anchor in which 0.5 SD does represent a medium effect. Now suppose this anchor is correlated $r = 0.8$ with the target, we will obtain 0.4 SD as the estimated *MID*. However, if we select another anchor, this time with correlation $r = 0.6$, the same 0.5 effect size reduces the estimated *MID* to 0.3 SD; or, for an anchor with correlation $r = 0.4$, the *MID* becomes 0.2 SD. Basically, we can obtain a *MID* as small as we want by choosing an anchor with as weak a correlation as we dare try to justify. This seems illogical.

Regression and prediction

Why does this attenuation occur? Regression aims to provide the optimal prediction of an outcome, *Y*, for individual subjects, where optimal is defined in terms of a criterion such as least squares or maximum likelihood. For simplicity, we consider linear regression, which can be written as $Y = a + bX$ where *Y* is the value of the outcome, or “dependent” variable, *X* is the value of the predictive factor, or “independent” variable, and *a*, *b* are numbers representing a constant offset and the slope of the line, respectively. Correlation assesses the predictive power of the factor *X*: the smaller the correlation, the weaker the predictive power. At one extreme, if there is perfect correlation of 1.0, the predictive factor *X* would suffice to provide a perfect estimate of the outcome (or *Y*-variable) for a future patient. At the other extreme, if there is zero correlation, factor *X* is of no value for predictive purposes and then the best estimate of outcome *Y* for a future patient is simply the mean value of the previously observed *Y*-scores. That is why, as noted above, the correlations affect the slope when using regression models because these are intended for prediction.

Thus, as the predictive power of *X* becomes weaker (*r* tends towards zero), so the best estimate for *Y* will shrink towards the mid-value, or mean. This phenomenon is well known and, for PROs, also affects the mapping from profile to preference-based measures [11]. It was first recognised by Francis Galton who observed that children of taller than average parents tend to be shorter and closer to the mean than their parents, and similarly shorter than average parents tend to have children who are taller than they are—and he termed this “regression to the mean” [12]. Nowadays, we simply say “regression” and far too often forget about the implications that Galton noted. Two other characteristics of the prediction-model approach may be noted. First, prediction is sometimes optimised by including additional factors or covariates, such as visual acuity, age and gender. Second, the equation is not symmetric, in that predicting *Y* from *X* is very different from predicting *X* from *Y*. That is, there are two regression lines, and the angle between these lines increases as *r* becomes smaller.

Perhaps most important of all, when using regression, the best prediction for an individual who is observed to be in (for example) the top 10 % of the distribution is that their *Y* outcome will be nearer to the mean.

Regression versus linking

Regression aims to *predict* the expected scores for *individual patients*, which is not the same as cross-calibration or *linking* of scales [11]. Instead, for linking, we are concerned with estimating the *equivalent* value on the target scale that corresponds to an observed change on the anchoring outcome. In contrast to the regression/prediction model, when linking from one scale to another, it is commonly agreed that an individual who lies *N* standard deviations above the mean on one scale might be expected to be similarly *N* standard deviations above the mean on the other scale. Similarly, if *P* % of individuals experience a clinically significant change in the anchor variable, we might expect roughly *P* % also to experience a change in the related target variable. For two scales, *X* and *Y*, this implies equality of the standardised values (difference from the mean divided by SD), resulting in the simple linear *linking function* [11, 13]:

$$\frac{X - \mathrm{Mean}_X}{\mathrm{SD}_X} = \frac{Y - \mathrm{Mean}_Y}{\mathrm{SD}_Y}$$

For the anchoring of *MID*s, this equation can be rearranged and written in terms of standardised anchor and target scales. It may be noted that, unlike the regression equation in which $b = r \times \left(\mathrm{SD}_{\mathrm{Target}}/\mathrm{SD}_{\mathrm{Anchor}}\right)$, the linking equation is unaffected by *r*:

$$MID = \text{Anchor Change} \times \left(\mathrm{SD}_{\mathrm{Target}}/\mathrm{SD}_{\mathrm{Anchor}}\right).$$

Thus, the *MID* equals the specified clinically important change in the anchor, scaled by the ratio of the standard deviations. It is equivalent to applying the effect-size ratio of the anchor to the target. An advantage of this approach is that it can be readily applied either to change-scores or to cross-sectional data, as it simply scales the anchor change by the respective SDs and we do not require regression analysis. For the above example, applying this approach would have resulted in $MID = 15 \times 19.2/13.2 = 21.8$, which is very different from the attenuated value proposed by Suňer et al. (It is important to bear in mind, however, that the problem with lumping together all those who changed equal to *or more than* the threshold for a minimally important change on the clinical measure counteracts the problem with regression to the mean in an unknown way for this example.)

Recommendations

Because we are in effect mapping the anchor and target scales against each other, the linking function which does not involve *r* should be used when calculating a *MID* from an anchor; that avoids the attenuation of the estimated values. It also carries the implication that if the widely used 0.5 × $\mathrm{SD}_{\mathrm{Anchor}}$ is regarded as a medium and clinically significant value that represents a medium effect size for the anchor [9, 14], then 0.5 × $\mathrm{SD}_{\mathrm{Target}}$ will always be the corresponding *MID* for the target scale—making the use of an anchor irrelevant. However, we also agree with authors who suggest that for a *minimal* important difference, it might be more appropriate to consider 0.2 × SD, which Cohen [9] described as a *small* effect size, rather than the medium effect of 0.5 [15].

Although the linking equation does not involve *r*, it is clearly desirable that the anchoring scale should be highly correlated with the target scores. If the correlation is too low, we cannot obtain a valid estimate for the *MID*. For establishing a *MID*, which is a group-based estimate, we recommend that *r* should be at least 0.371 [7, 8]. It is also strongly recommended that multiple anchors be considered, including self-rated anchors (global ratings of change), and if these converge towards a single number, the results will be more convincing [8].

Conclusions

Some studies using anchor variables to determine *MID*s have applied regression models. However, because of regression to the mean, these models are inappropriate for determining *MID*s; they result in estimates that are shrunk towards the middle of the distribution and are smaller than the correct values. When, for example, correlations between anchor and target approach 0.33, the estimate of the *MID* would be about a third of the true value. Linking of scales provides a simple way to rectify the problem.

Acknowledgments Ron D. Hays was supported in part by grants from the NIA (P30-AG021684) and the NIMHD (P20MD000182).

References

1. Jaeschke, R., Singer, J., & Guyatt, G. H. (1989). Measurement of health status: ascertaining the minimal clinically important difference. *Controlled Clinical Trials, 10*, 407–415.
2. US Food and Drug Administration. (2009). Patient-reported outcome measures: Use in medical product development to support labeling claims. Guidance for industry. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM071975.pdf. Accessed March 20, 2013.
3. McLeod, L. D., Coon, C. D., Martin, S., Fehnel, S. E., & Hays, R. D. (2011). Interpreting patient-reported outcome results: FDA guidance and emerging methods. *Expert Review of Pharmacoeconomics and Outcomes Research, 11*, 163–169.
4. Kvam, A. K., Wisløff, F., & Fayers, P. M. (2010). Minimal important differences and response shift in health-related quality of life; A longitudinal study in patients with multiple myeloma. *Health and Quality of Life Outcomes, 8*, 79.
5. Schwartz, N., & Sudman, S. (1994). *Autobiographical memory and the validity of retrospective reports*. New York: Springer.
6. Norman, G. (2003). Hi! How are you? Response shift, implicit theories and differing epistemologies. *Quality of Life Research, 12*, 239–249.
7. Hays, R. D., Farivar, S. S., & Liu, H. (2005). Approaches and recommendations for estimating minimally important differences for health-related quality of life measures. *Journal of Chronic Obstructive Pulmonary Disease, 2*, 63–67.
8. Revicki, D., Hays, R. D., Cella, D., & Sloan, J. (2008). Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of Clinical Epidemiology, 61*, 102–109.
9. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
10. Suňer, I. J., Kokame, G. T., Yu, E., Ward, J., Dolan, C., & Bressler, N. M. (2009). Responsiveness of NEI VFQ-25 to changes in visual acuity in neovascular AMD: Validation studies from two phase 3 clinical trials. *Investigative Ophthalmology & Visual Science, 50*, 3629–3635.
11. Fayers, P. M., & Hays, R. D. (2013). Linking should replace regression when mapping from profile to preference-based measures. *Value in Health* (submitted).
12. Galton, F. (1889). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain, 15*, 246–263.
13. Dorans, N. J. (2007). Linking scores from multiple health outcome instruments. *Quality of Life Research, 16*(Suppl 1), 85–94.
14. Norman, G. R., Sloan, J. A., & Wyrwich, K. W. (2003). Interpretation of changes in health-related quality of life: The remarkable universality of a half a standard deviation. *Medical Care, 41*, 582–592.
15. Farivar, S. S., Liu, H., & Hays, R. D. (2004). Half standard deviation estimate of the minimally important difference in HRQOL scores? *Expert Review of Pharmacoeconomics and Outcomes Research, 4*, 515–523.