Title

Should Linking Replace Regression When Mapping from Profile-Based Measures to Preference-Based Measures?

Journal

Value in Health, 17(2)

ISSN

1098-3015

Authors

Fayers, Peter M
Hays, Ron D

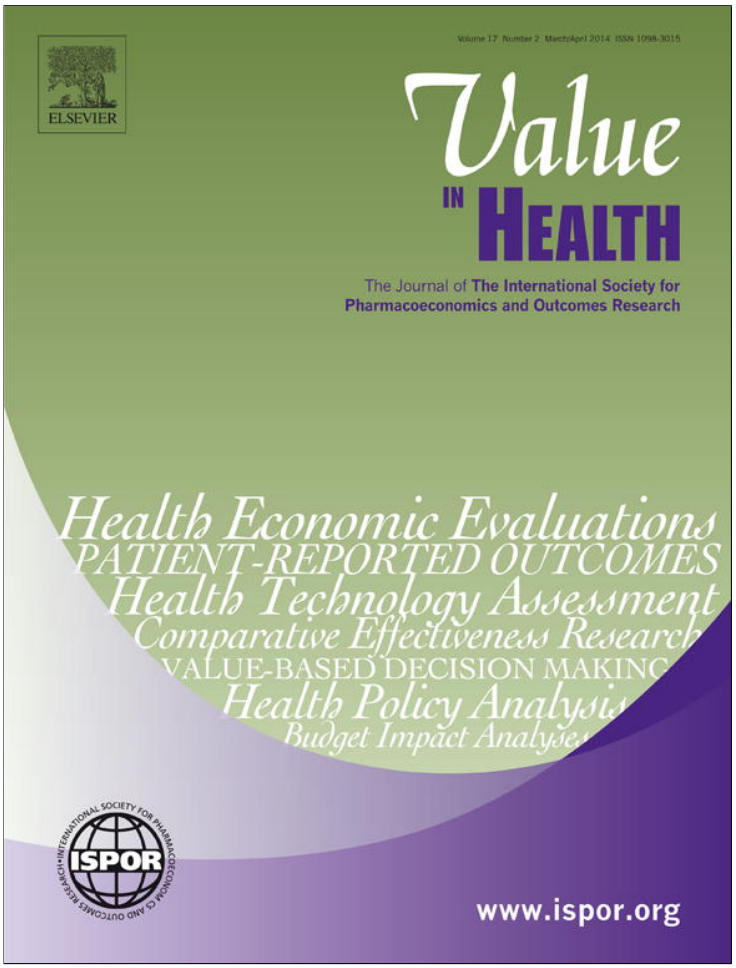
Publication Date

2014-03-01

DOI

10.1016/j.jval.2013.12.002

Peer reviewed



VALUE IN HEALTH 17 (2014) 261–265

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jval

METHODOLOGICAL ARTICLE

Should Linking Replace Regression When Mapping from Profile-Based Measures to Preference-Based Measures?

Peter M. Fayers, PhD[1,2,*], *Ron D. Hays, PhD*[3]

[1]*Institute of Applied Health Sciences, University of Aberdeen, Aberdeen, UK;* [2]*Department of Cancer Research and Molecular Medicine, Faculty of Medicine, Norwegian University of Science and Technology (NTNU), Trondheim, Norway;* [3]*UCLA Department of Medicine, Los Angeles, CA, USA*

ABSTRACT

Background: Profile instruments are frequently used to assess health-related quality of life and other patient-reported outcomes. However, preference-based measures are required for health-economic cost-utility evaluations. **Results:** Although regression-based approaches are commonly used to map from profile measures to preference measures, we show that this results in biased estimates because of regression to the mean. **Conclusions:** Linking (scale-aligning) is proposed as an alternative.

Keywords: linking values, mapping functions, patient-reported outcomes, preference-based measures, profile instruments, scale-aligning, test equating.

Introduction

Various health-related quality-of-life (HRQOL) measures are used in clinical trials and observational studies. Although the diversity of approaches is welcomed because measures need to be fit for purpose, it makes comparing and combining results challenging. A major advantage of preference-based measures is that they yield the single summary score needed to estimate quality-adjusted life-years (QALYs) for cost-utility evaluations. But it is frequently recommended that clinical trials should use disease-targeted instruments that are sensitive to clinically relevant changes, while at the same time minimizing patient burden and thereby maximizing survey response and item completion rates [1]. Thus, many studies include only generic [2] and disease-targeted [3] profile measures, and there has been great interest in mapping profile-based measures to preference-based measures to enable the calculation of QALYs [4–8]. "Mapping" is the equating (or "linking") of values from a source instrument to equivalent values on a target instrument. We emphasize that for health-economic evaluations the purpose of mapping is to obtain group-averaged estimates of QALYs with corresponding SDs to enable comparison of interventions (e.g., treatments or management policies) in clinical trials, observation studies, and meta-analyses.

Brazier et al. [6] reviewed 30 studies that reported 119 different models mapping profile-based measures to preference-based measures. The most common target measure was the EuroQol five-dimensional (EQ-5D) questionnaire, and the most widely used starting measures were the Short-Form 12 Health Survey (SF-12) or Short-Form 36 Health Survey (SF-36) profile measures. Brazier et al. comment that the performance of mapping functions in terms of goodness of fit and prediction is variable, and so it is impossible to generalize across instruments. Most importantly, the majority of mapping functions were estimated by using ordinary least squares. Some studies explored generalized linear models with random effects, adjusted least square regression models, weighted least squares, and other approaches, but, one way or another, they all used regression methods. We explain why these least-squares regression-based approaches are problematic for mapping.

Regression to the Mean

Predictions from regression models result in attenuated estimates. Indeed, the very term "regression" is short for "regression to the mean," and was defined by Francis Galton in 1886 in his seminal article, "Regression towards mediocrity in hereditary stature" [9]. (The Galton article was sullied by the pejorative term "mediocrity" and the eugenic beliefs that are now considered reprehensible.) Later, it was shown that the "mediocre" value toward which regression estimates tend is the central, or mean, value ("regression to the mean"). What triggered Galton's article was his observation that although tall fathers tend to have tall sons, these sons are usually less tall than their fathers; and short fathers similarly have sons who, while usually short, are also less extreme than their fathers. Superficially, this may appear to imply that over time everyone will have the same height, a

* *Address correspondence to:* Peter M. Fayers, 82 Tillydrone Avenue, Aberdeen AB24 2TN, UK.
E-mail: p.fayers@abdn.ac.uk.

1098-3015/$36.00 – see front matter Copyright © 2014, International Society for Pharmacoeconomics and Outcomes Research (ISPOR). Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.jval.2013.12.002

conclusion that is manifestly untrue. It is the presence of additional random variation that suffices to ensure that there are always new people at the two extreme ends of the distribution. What it does mean, however, is that for any individual, the best predicted ("true") score is less extreme—that is, regressed toward the mean.

Another way to think of regression to the mean is that if someone has a better than average score, his or her score is likely to be partly the consequence of an above-average underlying or "true" level and partly luck, so that the true score is likely to be closer to the mean value. If we assume that each person had two assessments, a test and a retest, then the retest score is likely to be closer to the mean. This same effect is observed in reverse, too: someone with a higher than average retest score is likely to have had a score nearer the mean on the original test. This effect is observed whenever there is less than perfect correlation between the two assessments. Prediction models shrink estimates toward the mean.

Mapping versus Prediction

The distinction between mapping (using scale-alignment) and prediction (using regression) has been recognized in the educational field for almost a century [10] as well as the "fallacy of using regression lines to show a true correspondence" [11]. In brief, when mapping educational examination scores, one is not interested in *predicting* the score a student might have obtained on another examination. Rather, one wants to know what score is *equivalent* for the second examination, such that students of a particular ranking on one examination are assigned the same ranking on the other examination. For example, suppose students are randomly assigned to take either examination *X* or examination *Y*. Because the assignment is random, students in the two groups should, on average, be of similar ability. Let us now assume that the two examinations assess the same underlying construct and that we wish to convert all scores to be on a single metric. One simple approach, known as equipercentile linking, is to ensure that a student in the top 5% for examination *X* will also be in the top 5% when the *X*-examination scores are converted to *Y*-examination scores. This is in contrast to predicting scores from regression analyses, when regression to the mean results in predicted scores that are closer to the mean value, with the best students who completed examination *X* therefore unfairly receiving a lower predicted score on examination *Y* than they deserve (and the less able students receiving the advantage of a higher than deserved score). When converting to *Y*-examination scores, using regression-based methods, the scores of students who completed examination *X* with low or high results become unfairly biased toward the mean *Y*-examination score.

The aim of prediction is typically to predict the most likely true score on the basis of information that is known about the respondent. Thus, other factors such as socioeconomic status, age, and sex might also be included if predictive. In contrast, mapping, or scale-aligning, does not predict scores for one instrument from another. Instead, it aims to align the scales so that the distributions are matched and an individual with a particular score on one scale can be compared with similar individuals assessed on the other scale [12]. Regression does not achieve this: as discussed below, the predicted *Y*-examination scores for individuals assessed using examination *X* will be less extreme than the observed scores of similar individuals who were assessed using examination *Y*, and therefore the overall ranking of individuals who took examination *X* will be biased relative to those who took examination *Y*.

Regression is a robust technique that is assuredly the most appropriate approach whenever the aims are to predict outcomes, evaluate explanatory factors, or explore potentially causal relationships. Scale-aligning has a very specific and different objective.

Shrinkage and Variance of the Predicted Scores

Brazier et al. [6, p. 221] comment that

"These papers also found that the predicted values from the mapping functions tend to have lower levels of variance than the original observed values."

Because of regression to the mean, this is hardly surprising and, in fact, is what is to be expected. In the simplest case of linear regression between two normally distributed variables, we have

$$\sigma^2_{\text{Predicted}} = r^2 \times \sigma^2_y, \quad (1)$$

where r is the correlation between *X* and *Y*, σ^2_y is the variance of the *Y* scores, and $\sigma^2_{\text{Predicted}}$ is the variance of the predicted *Y* scores based on the observed values of *X* scores. Thus, the amount of shrinkage of the variance is directly proportional to r^2. This tells us that as r tends toward 1.0, there is little or no loss of variability. Crucially, as the correlation between *X* and *Y* decreases (i.e., as r tends toward 0), the variance of the predicted values becomes smaller because regression-based predictions are increasingly shrunk toward the mean. As might be expected, in the extreme case of zero correlation, the predictor variable provides no information and then the best predicted value is simply the mean *Y* score and so the variance of the predicted values becomes zero.

Another effect of regression to the mean and the consequent reduced variance in predicted scores is that we should anticipate that the "misfit" will be most apparent for the highest-scoring individuals, whose scores will be underestimated relative to subjects who were assessed by using instrument *Y*, and the lowest-scoring persons, whose scores will be overestimated. This has been frequently noted in those mapping studies that use regression-based approaches, in which it has been observed that the cumulative distribution function of the predicted scores is shrunk at the tails in comparison with the observed values of the target distribution [6,13,14]. Finally, it is also worth noting that *X* and *Y* are interchangeable: if we instead use the values of *Y* to try to predict *X* scores, the variance of the predicted values of *X* would also have a variance shrunk by r^2.

Consequences for Mapping Profile-Based Measures to Preference-Based Measures

In the HRQOL setting, instrument *X* is usually a profile instrument such as the SF-12 and the short-form 36 health survey or the condition-specific European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire-Core 30; the target instrument, *Y*, is a preference-based measure such as the EQ-5D questionnaire or six-dimensional health state short form (derived from short-form 36 health survey). To estimate QALYs for a clinical trial that only used a profile instrument *X*, the *X* scores are first mapped to preference scores *Y* by using a published mapping equation that is typically based on linear regression [6], and then combined with patient survival times. One review found that many studies report the variance of *Y* that is explained by *X* to be generally above 50% for generic profile instruments, but lower for condition-specific instruments [6], while another review reported 33 comparisons with a median of 49% [7]. In such studies, the *Y* scores predicted by regression

will have a variance that is 50% or less than the variance that would have been observed if instrument Y had been used for direct measurements. The falsely low variances may result in optimistic claims of precision, with unduly narrow confidence intervals (CIs). For example, if r is 0.7 (corresponding to 49% variance explained), the CIs for the estimated Y scores will appear to be only 70% of the true width. When mapping profile-based to preference-based measures, few studies report $r = 0.8$ or greater, although when equating patient-reported outcomes for two instruments measuring the same domains it is reported that correlations are commonly 0.8 or greater [15]. We suggest that only when $r > 0.9$ can regression to the mean be ignored, which is rarely the case when mapping to preference-based measures, and even then variances are shrunk to 81% and CIs to 90% of the true value.

Within a single clinical trial, many statistical tests (e.g., t test) are unaffected by linear transformations and so if X is statistically significant then Y will be, too, despite regression to the mean. Also, in the simple single-study case, it may be possible to use compensatory adjustment to correct the shrunken CIs. When estimates for individual patients are used for calculating QALYs, however, the distribution of the QALY estimates will have a variance that is substantially shrunken in some unspecified manner that will depend on survival times, and it is unclear how this will affect the estimated QALYs and their CIs. In theory, it should be possible to adjust the regression-shrunk individual-patient estimates, but scale-aligning is a more direct and simpler approach. Care should also be taken with meta-analyses that combine results from several trials. If Y scores are derived from linear regression in some clinical trials while other trials directly used instrument Y, the regression-predicted Y scores and the directly observed Y scores will not be on scales with comparable frequency distributions, potentially invalidating significance tests and CIs unless compensatory adjustment is made. Nonlinear regression is even more complex. Scale-aligning preserves the mean and variance, thus avoiding these problems.

Methods for Mapping Scores

As mentioned above, test-linking and aligning have a long tradition in educational testing. Mapping and equating of examinations is "high stakes" because it determines the future prospects of students, and it is essential to be fair and unbiased when comparing those who have taken different examinations. Educational research has developed methods for comparison of students who have taken different examinations, and these techniques have been evaluated and applied to large samples covering students from a wide range of abilities [16–19]. Thus, we turn to education for details of suitable methodology for scale-aligning, while bearing in mind that unlike the educational setting, clinical trials and meta-analyses are group-based and we are more concerned with (1) estimating group effects than making precise estimates of scores for individuals and (2) preserving the properties of the estimated means and avoiding variance shrinkage.

Five requirements have been proposed for equating of scores to be valid [20], although these are intended for the equating of individuals rather than generating group-based statistics. Angoff [21] used the term "calibration" for linking scores that have differing reliability (relaxing requirement b) or different difficulty, and also described the scale-alignment of tests measuring different constructs as providing "comparable scores" [21]. Kolen [22] described the linking of different, but similar, constructs by using a common population of respondents as "battery scaling." We use the term "scale-aligning" and suggest that the same conditions are applicable when the focus is scale-aligning for group comparisons, with the exception of (b), which is not applicable for scale-alignment [12].

1. Equal constructs: The tests should measure the same constructs.
2. Equal reliability: The tests should have the same reliability.
3. Symmetry: The function for linking scores of Y to those of X should be the inverse of the function for linking scores of X to those of Y.
4. Equity: It should be a matter of indifference as to which of the two equated tests is used.
5. Population invariance: The choice of the subpopulation used to estimate the linking function between the scores of tests X and Y should not matter; the linking function should be population invariant.

Approaches that may be applied include the following:

1. Simple linear equating, based on equating the mean and SD of the two scales.
2. Equipercentile equating, which matches two cumulative distribution functions to each other either via smooth functions or in a nonparametric manner [12,17].
3. Item response theory–based methods that map onto logistic scales, possibly together with equipercentile equating [23].

We focus here on 1), the simple linear equating approach. Equipercentile equating is nontrivial because it requires presmoothing of the X and Y distributions and/or postsmoothing of the equipercentile relationship [12,17] because of discrete categories used in many HRQOL instruments. Equipercentile equating and item response theory have been used for mapping unidimensional patient-reported outcomes [15]. Linear equating, however, is the simplest method, and is the most analogous to linear regression. Most HRQOL mapping studies use a single group approach, in which all respondents complete both the profile instrument and the target preference-based instrument; this is rarely feasible in educational settings, where more complex designs are frequently used.

Equipercentile equating, mentioned above, provides a nonparametric approach that matches the entire cumulative distribution. To derive a linear scale-aligning function that is comparable to linear regression, the equipercentile requirement can be applied by ensuring that X scores and Y scores correspond to the same number of SDs above or below the mean. That leads to the following linking function that transforms the X scores to have the same mean and SD as the Y scores:

$$Y = \mu_Y + (\sigma_Y/\sigma_X)(X - \mu_X), \quad (2)$$

where μ_X and μ_Y are the mean values of X and Y, and σ_X and σ_Y are the SDs. Note that in contrast to the linear linking of Equation 2, the linear regression function (Equation 3) does involve the correlation, r, because the slope of the regression line is $\beta = r(\sigma_Y/\sigma_X)$.

$$Y = \mu_Y + \beta(X - \mu_X) = \mu_Y + r(\sigma_Y/\sigma_X)(X - \mu_X), \quad (3)$$

Equation 2 is symmetrical, and so it does not matter whether we convert from X to Y or from Y to X, and the linear relationship represents a single line. Equation 3, however, results in two different regression lines according to whether X or Y is regarded as the dependent (outcome) variable, and as the correlation tends toward 0 these lines increasingly diverge. Geometrically, for $r = 0.70$, the regression lines of Y on X and X on Y subtend an angle of approximately 20° [24], with the scale-aligning line roughly midway between the two; as r becomes smaller, the two regression lines diverge, regression to the mean increases, and shrinkage of variance in the predicted values becomes greater.

Although we used a simple scale-aligning argument to derive Equation 2, it can be shown that this equation also represents a form of regression in which it is assumed that random errors occur not only in Y but also in X; this "geometric mean regression" is explained in the Statistical Appendix in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2013.12.002 (although called "regression," the estimated values no longer "regress" to the mean). Lu et al. [25], using a different approach, also conclude that ordinary least squares regression is not coherent and geometric mean regression is preferable. In practice, many HRQOL mapping studies make use of multiple subscales from the generic measure, and both the regression equation (Equation 3) and the linking equation (Equation 2) can be extended to these more general models.

Validity and Goodness of Fit

Goodness of fit is frequently assessed in terms of root mean square error (RMSE), which is the square root of the summed squared differences between the observed and expected values. If the parameters for linear regression have been evaluated on the basis of ordinary least-squares estimation, linear regression is by definition optimal in terms of a linear relationship that yields the smallest RMSE. Thus, all other linear scale-aligning, linking, or mapping functions will inevitably show poorer fit in terms of RMSE statistics. For mapping, it is inappropriate to define goodness of fit in terms of predictive ability. The role of mapping or scale-alignment is to determine equivalent scores such that respondents taking either test will achieve the same overall rank score as if they had taken the other test, with, as a consequence, the equivalent cumulative distribution function, mean, and standard distribution for observed and estimated scores.

Other methods than goodness of fit are required for comparing different approaches. Longworth and Rowen [26] review methods of validating and evaluating mapping studies. They suggest assessing performance and validity of a linking function by predictive ability and elements that include 1) content validity and the extent to which the tests measure similar constructs; 2) strength of association between the scores; 3) the quality of the linking data and the mapping study (e.g., qualitative and descriptive review); 4) comparison of the distributions and cumulative distributions of the variables; and 5) studies to evaluate the population invariance [20] of the linking functions.

Content Validity and Similarity of Constructs

The validity of a mapping depends on the assumption that the two instruments assess the same or closely similar constructs. This may, for example, be assessed by using qualitative methods in which patients and experts formally compare the wording and meaning of items. Blome et al. [27] map a skin-targeted HRQOL instrument for psoriasis to the EQ-5D questionnaire. The authors found poor association between the constructs that "seem to be too different to be equivalent to each other, because the two instruments assess largely different aspects of patient impairment." They acknowledged that this mapping "has severe limitations in validity and clinical relevance" and postulate that "comparable results could be derived from studies on other skin diseases. Consequently, the EQ-5D questionnaire or comparable instruments should be implemented in studies aiming to measure utilities, because utilities cannot reliably be estimated from other study variables." This conclusion is likely to apply whenever condition-targeted profile scales are mapped to generic preference scales. Although similarity of constructs is essentially a qualitative judgment, high similarity may be expected to lead to strong correlation between the scales.

Strength of Association between the Scores

Association between X and Y scores can be measured by the correlation coefficient, r, or the multiple correlation coefficient, R, when more complex models with covariates are used. Although the correlation coefficient does not appear explicitly in the scale-aligning function (Equation 2), for both regression and scale-aligning the correlation should be high, and if the two instruments really are assessing similar constructs it will be. We have been impressed, however, by how low the correlations are in many studies mapping disease-specific profiles to preference measures. As observed above, reviews have reported that in as many as half of the mapping studies, R^2 (the variance explained) fails to reach 50%. Thus, Blome et al. [27] reported that R^2 was only 0.24. So, what value is acceptable for mapping and scale-aligning? In educational settings, it has been suggested that correlations must exceed 0.87 for adequate scale-aligning of individuals [28,29]. For group-based estimates, as when comparing the average number of QALYs from clinical trials in health economic assessments, the magnitude of the correlation need not be as high but it should still be reasonably large. By analogy with the widely used threshold of correlations for reliability of group-level comparisons, and on the basis of our experience, we suggest a threshold of 0.70 as the lowest acceptable. As many as half the published studies are rejected by this criterion. This level of correlation, however, still represents poor agreement between the observed scores X and Y, and implies that the proportion of variance explained will only be 49%; thus, many would argue that even 0.70 is too low a threshold.

Quality of the Linking Data and the Mapping Study

Longworth and Rowen [26] provide guidance about the requirements for a mapping study. Generally, the mapping study will either comprise data from one or more clinical trials, or will be a purpose-designed mapping study. To be confident about the generalizability of the mapping function to future samples, the characteristics of the estimation sample should be as similar as possible to the characteristics of the sample to which the mapping algorithm will be applied.

Similarity of Distribution Functions

Similarity of distribution functions is a prerequisite for linear alignment of scales; the linear linking functions merely adjust for different means and SDs. If the content of the scales is similar, this requirement is likely to follow. Problems arise, however, when linking profile measures to the EQ-5D questionnaire, because the EQ-5D questionnaire returns scores that follow a bimodal (or even trimodal) distribution, with respondents grouped into low and high values and very few in the middle [6,13,14,30]. The six-dimensional health state short form (derived from short-form 36 health survey), however, follows a smooth unimodal continuum—as is commonly observed with scales from health profile instruments such as the European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire-Core 30 [31]—and which seems more plausible for samples of patients in clinical trials. This disparity is also seen with the EQ-5D questionnaire cumulative distribution function. Clearly, these indexes measure different things. Possible causes are that their underlying constructs are fundamentally different, or their scoring algorithms are inconsistent. Arguably, no mapping is likely to compensate for one measure smoothly covering the continuum while the other is strangely bimodal. At the very least, nonlinear (possibly nonparametric) solutions must be considered.

Population Invariance

Dorans and Holland [20] show that an effective way of confirming the validity of linking scores is by assessing whether the linking function is invariant in diverse subpopulations, because differences in constructs or reliability of instruments are manifested by population invariance, as are nonlinearity and other departures from the model [20]. For HRQOL instruments, possible subgroups to explore might be disease type, disease severity, age, sex, and race/ethnicity. Dorans and Holland describe and illustrate suitable tests.

Conclusions

Regression, which is a method for predicting outcomes and is normally quite justifiably the method of choice, differs from scale-alignment, which is appropriate when mapping between instruments. The differences are largely attributable to regression to the mean, which is a frequently overlooked and misunderstood phenomenon. For simplicity of exposition, we have focused on the simple case of linear regression and linear linking functions in single group designs. In HRQOL research, the term "mapping" has usually been implemented by using regression-based prediction. The use of regression models, however, is inappropriate for that task and results in biased estimates. Approaches such as nonparametric equipercentile methods or parametric linking functions should instead be used for mapping profile-based measures to preference-based measures. Other options include item response theory [23], but care should still be taken to distinguish between prediction of individual scores (when uncertainty shrinks estimated values toward the mean) and mapping.

Source of financial support: Ron D. Hays was supported in part by grants from the National Institute on Aging (grant no. P30-AG021684) and the National Institute on Minority Health and Health Disparities (grant no. P20MD000182).

Supplemental Materials

Supplemental materials accompanying this article can be found in the online version as a hyperlink at http://dx.doi.org/10.1016/j.jval.2013.12.002 or, if a hard copy of article, at www.valueinhealthjournal.com/issues (select volume, issue, and article).

REFERENCES

[1] Fayers PM, Machin D. Quality of Life: The Assessment, Analysis and Interpretation of Patient-Reported Outcomes. (2nd ed). Chichester: Wiley & Sons Ltd, 2007.
[2] Hunt SM, McEwan J. The development of a subjective health indicator. Sociol Health Illness 1980;2:231–46.
[3] Hays RD, Mangione CM, Ellwein L, et al. Psychometric properties of the National Eye Institute-Refractive Error Quality of Life instrument. Ophthalmology 2003;110:2292–301.
[4] Revicki DA, Kawata AK, Harnam N, et al. Predicting EuroQol (EQ-5D) scores from the Patient-Reported Outcomes Measurement Information System (PROMIS) global items and domain item banks in a United States sample. Qual Life Res 2009;18:783–91.
[5] Kowalski JW, Rentz AM, Walt JG, et al. Rasch analysis in the development of a simplified version of the National Eye Institute Visual-Function Questionnaire-25 for utility estimation. Qual Life Res 2011;21:323–34.
[6] Brazier JE, Yang Y, Tsuchiya A, Rowen DL. A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. Eur J Health Econ 2010;11:215–25.
[7] Mortimer D, Segal L. Comparing the incomparable? A systematic review of competing techniques for converting descriptive measures of health status into QALY-weights. Med Decis Making 2008;28:66–89.
[8] Chuang L-H, Whitehead SJ. Mapping for economic evaluation. Br Med Bull 2012;101:1–15.
[9] Galton F. Regression towards mediocrity in hereditary stature. J Anthrop Inst Gr Br 1886;15:246–63.
[10] Thorndike EL. On finding equivalent schools in tests of intelligence. J Appl Psychol 1922;6:29–33.
[11] Otis AS. The method for finding the correspondence between scores in two tests. J Educ Psychol 1922;13:529–45.
[12] Dorans NJ. Linking scores from multiple health outcome instruments. Qual Life Res 2007;16:85–94.
[13] Versteegh MM, Rowen D, Brazier JE, Stolk EA. Mapping onto EQ-5 D for patients in poor health. Health Qual Life Outcomes 2010;8:141.
[14] Rowen D, Brazier J, Roberts J. Mapping SF-36 onto the EQ-5D index: how reliable is the relationship? Health Qual Life Outcomes 2009;7:27.
[15] Choi SW, Podrabsky T, McKinney N, et al. PROSetta Stone® Analysis Report: A Rosetta Stone for Patient Reported Outcomes (Vol. 1). Chicago, IL: Department of Medical Social Sciences, Feinberg School of Medicine, Northwestern University, 2012.
[16] Brennan RL, ed. Educational Measurement. (4th ed.). Westport, CT: Praeger Publishers, 2007.
[17] Kolen MJ, Brennan RL. Test Equating, Scaling, and Linking: Methods and Practices. (2nd ed.) New York: Springer, 2004.
[18] von Davier AA, ed. Statistical Models for Test Equating, Scaling, and Linking. New York: Springer, 2010.
[19] Dorans NJ, Pommerich M, Holland PW, eds. Linking and Aligning Scores and Scales. New York: Springer, 2007.
[20] Dorans NJ, Holland PW. Population invariance and the equatability of tests: basic theory and the linear case. J Educ Meas 2000;37:281–306.
[21] Angoff WH. Scales, norms, and equivalent scores. In: Thorndike RL, ed. Educational Measurement. (2nd ed.). Washington, DC: American Council on Education, 1971:508–600.
[22] Kolen MJ. Linking assessments: concept and history. Appl Psychol Meas 2004;28:219–26.
[23] Kolen MJ. Comparison of traditional and item response theory methods for equating tests. J Educ Meas 1981;18:1–11.
[24] Schmid J. The relationship between the coefficient of correlation and the angle included between regression lines. J Educ Res 1947;41:311–3.
[25] Lu G, Brazier JE, Ades AE. Mapping from disease-specific to generic health-related quality-of-life scales: a common factor model. Value Health 2013;16:177–84.
[26] Longworth L, Rowen D. Mapping to obtain EQ-5D utility values for use in NICE health technology assessments. Value Health 2013;16:202–10.
[27] Blome C, Beikert FC, Rustenbach SJ, Augustin M. Mapping DLQI on EQ-5D in psoriasis: transformation of skin-specific health-related quality of life into utilities. Arch Dermatol Res 2013;305:197–204.
[28] Dorans NJ. Equating, concordance, and expectation. Appl Psychol Meas 2004;28:227–46.
[29] Dorans NJ, Walker ME. Sizing up linkages. In: Dorans NJ, Pommerich M, Holland PW, eds. Linking and Aligning Scores and Scales. New York: Springer, 2007.
[30] Brazier J, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D and SF-6D across seven patient groups. Health Econ 2004;13:873–84.
[31] Scott NW, Fayers PM, Aaronson NK, et al. EORTC QLQ-C30 Reference Values. Brussels: EORTC Quality of Life Group, 2008.