Looking for Important Change/Differences in Studies of Responsiveness

DORCAS E. BEATON, CLAIRE BOMBARDIER, JEFFREY N. KATZ, JAMES G. WRIGHT, GEORGE WELLS, MAARTEN BOERS, VIBEKE STRAND, and BEVERLY SHEA, for the OMERACT MCID Working Group

ABSTRACT. The purpose of this paper is to describe a classification system for studies of responsiveness that was designed to help organize these studies, and identify those with the potential to provide information on minimal clinically important difference (MCID). We developed a 3 dimensional cube into which studies of responsiveness can be categorized based on their evaluation of 3 attributes: 1. individual or group setting; 2. which scores are contrasted; and 3. the type of change or difference being assessed. We present and discuss examples of studies that fit into categories in the classification cube. This classification system helps to focus attention on whether the literature is able to provide information on the specific type of change a person is interested in. It reinforces that the ability of an instrument to detect a certain category of discrimination within the cube does not mean it will necessarily be responsive to another category. The cube has been shown here as a means to separate out studies that address important change. These studies can then be examined as the source of information on MCID. (J Rheumatol 2001;28:400–5)

INTRODUCTION

Most clinical studies aim at discriminating (showing differences) between groups that are of interest. In evaluative studies, the difference of interest is usually in change over time, e.g., response to therapy. Studies of responsiveness evaluate the ability of an outcome measure to accurately detect change when it has occurred[1]. The number of studies addressing responsiveness has increased rapidly in recent years. Depending on the design of the study, however, it may or may not be able to provide information on the ability to detect *important* changes or differences, or that elusive minimal clinically important difference (MCID).

From the Institute for Work & Health, Toronto, Canada.

D.E. Beaton, BScOT, MSc, PhD, Institute for Work & Health, Department of Occupational Therapy, University of Toronto; C. Bombardier, MD, FRCP, Institute for Work & Health, Clinical Epidemiology and Health Care Research Program, Department of Health Administration, University of Toronto; J.N. Katz, MD, MS, Harvard Medical School, and the Robert Brigham Multipurpose Arthritis and Musculoskeletal Diseases Center, Boston, MA, USA; J.G. Wright, MD, FRCSC, MPH, Clinical Epidemiology and Health Care Research Program, Department of Surgery, University of Toronto; G. Wells, PhD, Faculty of Epidemiology and Community Medicine, University of Ottawa, Clinical Epidemiology Unit, Loeb Health Research Institute, Ottawa, Canada; M. Boers, Department of Clinical Epidemiology, VU University Hospital, Amsterdam, The Netherlands; V. Strand, Department of Medicine, Stanford University, Stanford, CA, USA; B. Shea, Clinical Epidemiology Unit, Loeb Health Research Institute, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada.

Dr. Beaton was supported by a PhD fellowship (health research) from the Medical Research Council of Canada and by the Institute for Work & Health during the time this research was done. Dr. Wright is the R.B. Salter Chair of Surgical Research and a Medical Research Council of Canada Scientist. Dr. Katz is supported through a grant from the NIH #AR36308 and the US National Arthritis Foundation.

Address reprint requests to Dr. D.E. Beaton, Institute for Work & Health, 250 Bloor St. East, Ste. 702, Toronto, Ontario M4W 1E6.
E-mail: dbeaton@iwh.on.ca

MEASURING RESPONSIVENESS

Each evaluation of responsiveness is built around a construct that suggests that the target attribute (e.g., disability, pain, joint count) has shifted or is different in some way. In some situations this takes the form of an expected pattern of recovery (such as one week before and 6 months following total hip replacement)[2]. In other situations a specific external marker of change is used, often in the form of a global index: Is your pain better? The ability of a measurement instrument to detect that variation is then described with summary statistics such as effect sizes, or other responsiveness statistics[3]. The construct within the study therefore plays an important role in defining that change has occurred.

Any given study of responsiveness can only provide information about the ability of an outcome measure to detect the specific construct of change designed into that study. Readers must appraise if information from that study will be useful to them. For instance, will it help them understand when an individual patient has had an important improvement. The growing volume of literature on responsiveness does not make this an easy task. The reader must critically appraise the studies and tease out exactly what kind of change a given study does in fact address. The purpose of this paper is to describe a classification system for studies of responsiveness that will help organize these studies, and identify those with the potential to provide information on MCID.

CLASSIFICATION OF STUDIES OF RESPONSIVENESS BASED ON THE CONSTRUCT OF CHANGE/DIFFERENCE USED

A review of the literature of responsiveness revealed 3 key features that help to define the attributes of the change/

difference designed into a given study[4]. The features are defined, either explicitly or implicitly, by the researchers involved within each study of responsiveness and are reflected in the design and analysis of that study. First is what we have called the "setting." The researchers decide whether they will target the analysis and the presentation of the results at a group level (average change in pain in patients getting total hip replacements), or at an individual level (smallest detectable change in joint width of one individual's radiograph). The second decision made by the researchers is which scores will be contrasted in that study, an axis we have labeled "which." Most responsiveness studies look at repeated scores in the same patients over time, but other possibilities also exist in the literature and will be described later. Finally, the researchers have also defined the type of change or difference that they are targeting, which spans 5 possible categories and also defines whose perspective is being sought (described in detail below).

These 3 key features are defined, intentionally or not, in each study. They are mutually independent, and because of that can be fit together into a "cube" (see Figure 1). Each cell within the cube is defined by its place along the 3 key features. And each cell becomes a description of the construct of change built into a study of responsiveness. The cube describes these cells as different, not more or less valid than each other. The cube becomes a classification system, classifying the nature of discrimination (either differences or changes) built into studies of responsiveness. Important change will be shown to be retrievable only from studies focusing on certain cells within the cube.

Each of the axes and its categories will now be described. It should be reinforced that the cube tries to reflect what is found in the literature, rather than what the authors feel makes up the "best" ways to approach responsiveness.

DESCRIPTION OF THE AXES OF THE CUBE OF DISCRIMINATION

Axis I: Setting. The first axis refers to the study setting. Specifically, whether the study results are targeting descrip-

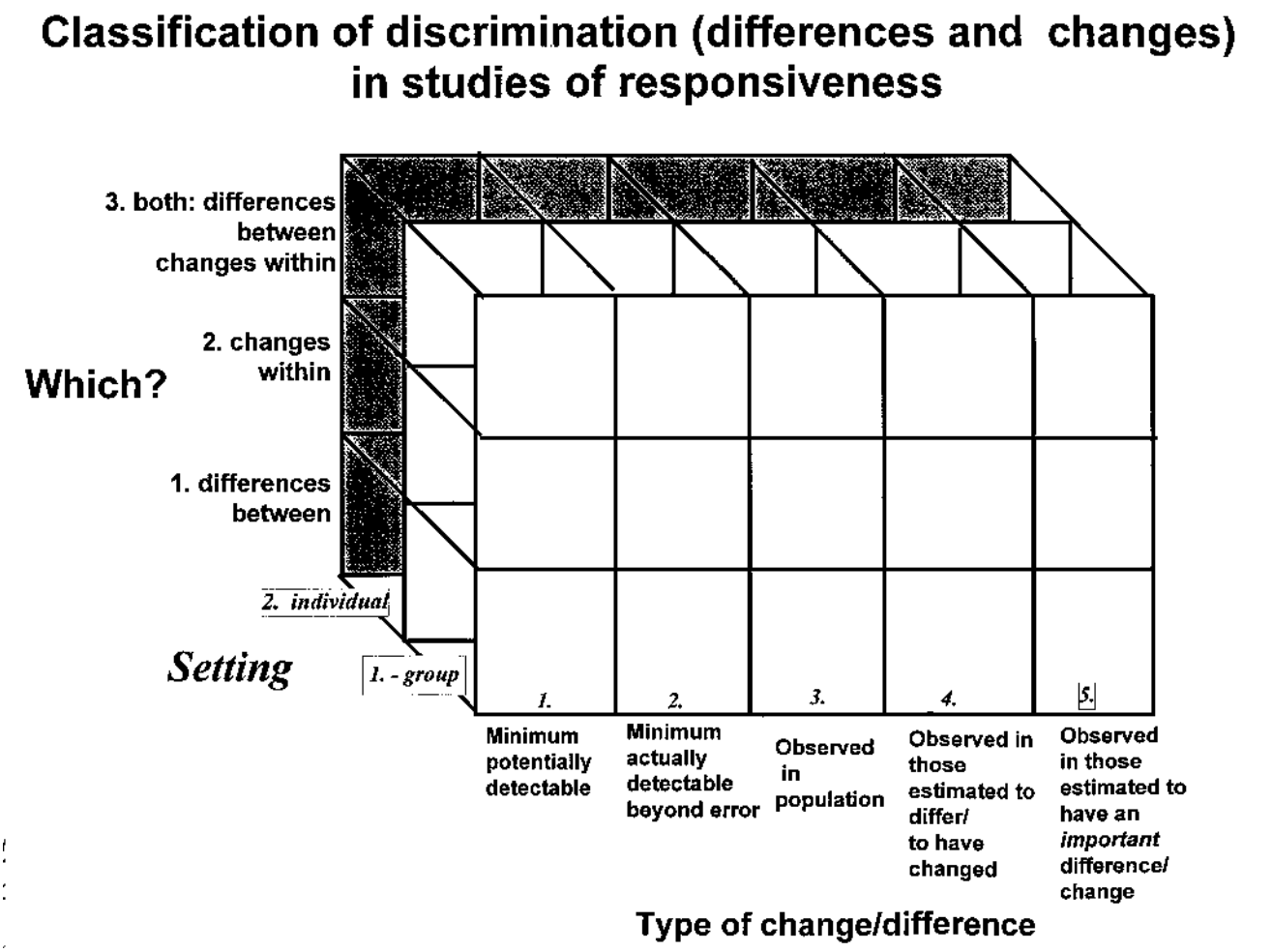

Figure 1. The taxonomy or "cube" of change and discrimination in studies of responsiveness/discrimination. Each cell in the cube describes the construct of change/difference that could be examined in a study under the rubric of responsiveness.

tion at a group level or at an individual patient level. Research has shown that the interpretation of change may vary depending on whether we are thinking at a group level (where smaller changes may be interpreted as important) or at an individual level, where larger changes are required before they are confidently accepted as indicating a meaningful change[5-10]. The results of studies usually are presented for one or the other level; however, sometimes both are presented in the results. For example, work by Redelmeier[9] discussed differences at a group level, but returned to the full set of data and reported that the average change in the group was not good at discriminating between improved and unimproved patients at an individual level.

Individual-level reporting requires analysis similar to diagnostic testing: with a description of a specific numeric score (reflecting either a change or difference in score) along with the sensitivity and specificity of that change in the study sample. A good example of this would be the recent work of Stratford[11] in low back pain patients with the Roland scale. A change score of 5 Roland Morris scale points was found to have sensitivity of 0.72 and specificity of 0.82 for discriminating between those achieving and not achieving important levels of improvement. This was done in a sample of patients undergoing physiotherapy for low back pain of less than 6 weeks' duration. Data were gathered on admission, and following 3 to 6 weeks of treatment, and the cut-point determined by receiver-operator characteristic curve analysis.

Few studies provide the individual-level analysis. In its absence, caution should be used when bringing things such as the average of a sample of people who experienced important change to the level of interpretation for an individual patient[10].

Axis II: Which? The Which axis defines which scores are being contrasted in the study. The base level, the bottom row, represents the discrimination at one point in time, as the scores between persons are contrasted[9,12]. Although this type of contrast may not typically be considered as a form of "responsiveness," work by Redelmeier has suggested that the differences in the scores of pairs of people, one of whom said they were "healthier" than the other, can be used to determine minimally clinically important differences[9,12-15]. Some would suggest that this row be excluded from the taxonomy because it is less informative about how responsiveness is traditionally quantified. Responsiveness is linked with longitudinal change in outcome measures' scores within patients. However, studies like Redelmeier's[14] will be found in the literature on responsiveness, and leaving this row in the taxonomy would allow people to place that in the taxonomy, and then correctly identify it as different than longitudinal studies that examine within-person changes over time[16]. The reader can then make the decision as to whether it informs what the construct of change/difference is in which they are interested. We believe, in the majority of the cases, it will not. Further empirical work may show they are interchangeable, but this does not yet appear to be the case.

As described above, studies of responsiveness are most commonly performed at the next level of the "which" axis of the cube: contrasting change within individuals or a group of individuals over time[2,17,18-20].

Finally, studies were found that combined these concepts of within and between and really reflect what we have labeled "both." This would be the randomized controlled trial where the focus is on between-group differences of within-person change[5,21], or the change in the treatment group relative to the change in the control group.

These 3 categories, between, within, and both, define the "which" axis, which scores are being contrasted in published studies of responsiveness.

Axis III: Types of change/difference? The third axis defines the type of change or difference being quantified in a study of responsiveness. The various categories are summarized in Table 1. For simplicity, we will refer only to the term "change" (within-person change over time) in the following descriptions; the same categories would apply to differences between persons and the hybrid (between-group differences of within-person change) contrasts as well.

The minimum potentially detectable change is the smallest increment of change possible on that instrument given the number of items and response options. It really provides an anchor for describing change because it would be difficult to interpret a change smaller than the smallest possible increment on an instrument.

The second type is the minimum detectable change (MDC) that needs to be observed before it is considered above the bounds of measurement error for that instrument and application[22-26]. This approach uses the standard error of measurement (SEM) and establishes the MDC that would be found to be statistically significantly different from zero

Table 1. Axis III: Type of change/difference being quantified in the study.

1. Minimum potentially detectable
2. Minimum actually detectable given the measurement error of the instrument
3. Observed change/difference measured by the instrument in a given population
4. Observed change/difference in a population deemed to have differed/changed by
   - the patient
   - the clinician/researcher
   - the payer
   - society
5. Observed change/difference in those deemed to have had an *important* difference/change by
   - the patient
   - the clinician/researcher
   - the payer
   - society

(no change) at a given level of confidence (usually 90% or 95%). This is calculated using the formula:

MDC (95% confidence level) = $1.96 * \sqrt{2} * \text{SEM}$

The SEM can be estimated by the standard deviation of baseline scores multiplied by the square root of one minus the reliability coefficient[22,23,27-29]. In a sense this is the upper limit of the 95% confidence interval (CI) around a change of zero, or no change. Similar thresholds could be determined at a 90% CI, or at any other level. The CI should therefore be specified.

When making estimates around change over time, the corresponding reliability coefficient to use in the formula is the test-retest reliability[30], although some have used the alpha coefficient[25,26]. The alpha coefficient is considered appropriate when estimating the precision around a single observation, that is, how close a given score is to the "true" value[30].

The MDC provides another threshold for interpretation. When a change score exceeds this level, there is reasonable certainty (95% CI, for instance) that it is true signal, and not just noise or error. The estimate has limitations. There is an assumption that the amount of change reflected in the MDC will be the same across the range of possible scores. However, this is being shown not to be the case and we must consider the degree to which the meaning of a change in score depends on where the person is on the scale (very disabled vs only mildly disabled)[24,31,32]. The MDC should be considered a guideline, not an absolute.

The third category on this axis of the cube is that of observed change/difference. This is quantified when scores are contrasted in situations where variation in the attribute is expected, but not specifically verified as having occurred. The clearest example might be the change observed before and after a treatment (usually of "known efficacy")[2,17,19]. Often, studies have used the changes expected before and after a reliable procedure such as total joint arthroplasty, or the usual course of a disorder such as the early stages of simple, acute low back pain[17,33]. In these cases the expected course becomes the construct of change.

The fourth and fifth categories use some sort of indicator to verify that change has occurred, or that a difference exists between people. This external indicator is the key point differentiating these from the observed type of change[16,34]. The sample is then stratified according to whether this has been experienced or not, and responsiveness is calculated on those who have changed/who are different. Examples of this would be the global indicators of change (compared to "Before your surgery are you: much worse...same....much better?"). Those reporting "better" would be considered for analysis of responsiveness to improvement, and those reporting "worse" would have their change scores examined for responsiveness to deterioration (not commonly done in traditional studies of responsiveness).

For important change (type 5), analysis is done on the change observed in those who have had an improvement/deterioration that was also an ***important*** improvement/deterioration. Someone assigns the value of importance to the experience of change. Similarly, it could also be a difference that was deemed to occur between persons, and was also considered to be an important difference. Determination of the *importance* of the change is critical, and should be examined carefully as a component of the validity of the results.

Both estimated and important change have 4 subdivisions, each defining the perspective used in determining the occurrence or importance of the change. The patient can determine the occurrence of change[17,35] or their difference from another person[13]. However, in other studies the clinician determines the change[36]. Other less frequently used, but nonetheless valid perspectives would be the payer, and finally society[37,38]. These 4 perspectives could lead to very different definitions of who has experienced "important improvement" and indeed in determining what is considered the minimal clinically important change/difference. Both estimated and important change should make the perspective taken explicit, and therefore define "from whose perspective?"[38-40].

APPLYING THE CUBE WHEN LOOKING FOR INFORMATION ON IMPORTANT CHANGE, OR MINIMAL CLINICALLY IMPORTANT DIFFERENCES

Studies of the responsiveness of outcome measures in a given clinical field can be classified according to their place in the cube (Figure 1). For example, Deyo's work on low back pain (onset to 3 weeks) followed all patients with an acute episode of low back pain from onset to 3 weeks, anticipating that a large proportion would be better, given the natural history of this disorder. In this particular analysis, no external indicator of whether or not they improved was used. This would fit into group-level, within-person observed change[17]. Bombardier's auranofin trials contrasted the responsiveness of different outcome measures using data gathered in a randomized clinical trial. Change in the treatment group over and above the control group was the focus of the analysis. This would be an example of group-level — both within-person and between-person — observed change[21]. Buchbinder's review of outcome measures used a similar approach, gathering data from multiple controlled trials, and would fall into the same cell in the taxonomy cube[41]. Stratford's work (described above) tried to determine the change score that could most accurately identify individuals who had experienced an important improvement (from a combined clinician and patient perspective) from those who did not[11]. This study looked at important change within persons over time, and the results were presented in a manner that could be used for individual patients.

Once studies have been placed into the appropriate cells

in the cube, the furthest right column, those studies that address "important change/difference," can be separated from the rest of the cube. These 6 cells in the column labeled "important change/difference" then become the focus of attention, as they will be the only source of information that can be used to determine the MCID. Wells, *et al*[40] provide a review of the different approaches that have been used to move from these cells in the cube, to the establishment of a MCID.

The cube therefore helps to sift and sort through the studies of responsiveness to define the constructs of change that have been studied on a given outcome measure, or across measures in a given clinical area. Many studies of responsiveness are not designed to address important change/differences or MCID because they have focused on an equally valid but different construct of change. These studies can be set aside in order to direct attention at those most likely to provide the information needed — those studies addressing important change.

The cube is also helpful if no studies are found that address a specific kind of change. In that situation the cube can be used to help design a new study by defining what variables should be considered: what setting (do you want analysis that will give information at an individual or group level?); which contrast or setting (between-person, within-person, or both); and type of change/difference. In this way, researchers can be sure that they will obtain the information that they ultimately need.

CONCLUSION

This paper has described responsiveness as the ability of an instrument to accurately detect changes or differences when they have occurred, and describes a classification system that helps unravel determining when change/difference has occurred, and what type of change/difference that was. The cube of discrimination reflects our efforts to clarify the types of studies that will be found in the literature under the rubric of responsiveness.

This classification system helps to focus attention on whether the literature is able to provide information on the specific type of change a person is interested in. It reinforces that the ability of an instrument to detect a certain category of discrimination within the cube does not mean it will necessarily be responsive to another category.

Minimal clinically important differences are meaningful thresholds in the distribution of *important* change scores. The cube has been shown here as a means to separate out studies that address important change from others. These studies can then be examined as the source of information on MCID. A companion article moves from this sorting exercise, to describe methods that have been used to determine the MCID from these studies[40].

The cube addresses just one component of the applicability of a study of responsiveness, and indeed the magnitude of the change quantified. Other factors such as the patient group, the intervention (some will produce much larger, more dramatic effects than others), and the timing between assessments (4 weeks or 12 months?) must also be considered[42].

REFERENCES

1. De Bruin AF, Diederiks JPM, De Witte LP, Stevens FCJ, Philipsen H. Assessing the responsiveness of a functional status measure: the Sickness Impact Profile versus the SIP68. J Clin Epidemiol 1997;50:529-40.
2. Laupacis A, Bourne R, Rorabeck C, et al. The effect of elective total hip replacement on health-related quality of life. J Bone Joint Surg 1993;75A:1619-26.
3. Wright JG, Young NL. A comparison of different indices of responsiveness. J Clin Epidemiol 1997;50:239-46.
4. Beaton DE. A taxonomy of responsiveness. In: Are you better? Describing and explaining changes in health status in persons with upper extremity musculoskeletal disorders [thesis]. Toronto: University of Toronto Press; 2000:15-47.
5. Goldsmith CH, Boers M, Bombardier C, Tugwell P. Criteria for clinically important changes in outcomes: development, scoring and evaluation of rheumatoid arthritis patient and trial profiles. J Rheumatol 1993;20:561-5.
6. Nesselroade JR. Interindividual differences in intraindividual change. In: Collins LM, Horn JL, editors. Best methods for the analysis of change. Washington, DC: American Psychological Association; 1991:92-105.
7. Peterson MGE, Williams PG. Composite index methodology [reply to letter]. Paulus H, Egger MJ, Ward JR, Williams HJ. Arthritis Rheum 1991;34:502-4.
8. Redelmeier DA, Tversky A. Discrepancy between medical decisions for individual patients and for groups. N Engl J Med 1990;322:1162-4.
9. Redelmeier DA, Lorig K. Assessing the clinical importance of symptomatic improvements: an illustration in rheumatology. Arch Intern Med 1993;153:1337-42.
10. Testa MA. Interpreting quality-of-life clinical trial data for use in the clinical practice of antihypertensive therapy. J Hypertension 1987;5 Suppl 1:S9-S13.
11. Stratford PW, Binkley JM, Riddle DL, Guyatt GH. Sensitivity to change of the Roland-Morris back pain questionnaire: 1. Phys Ther 1998;78:1186-96.
12. Redelmeier DA, Guyatt GH, Goldstein RS. Assessing the minimal important difference in symptoms: a comparison of two techniques. J Clin Epidemiol 1996;49:1215-9.
13. Wells GA, Tugwell P, Kraag GR, Baker PRA, Groh J, Redelmeier DA. Minimum important difference between patients with rheumatoid arthritis: the patient's perspective. J Rheumatol 1993;20:557-60.
14. Redelmeier DA, Goldstein R, Min ST, Hyland RH. Spirometry and dyspnea in patients with COPD: when small differences mean little. Chest 1996;109:1163-8.
15. Redelmeier DA, Guyatt GH, Goldstein RS. On the debate over methods for estimating the clinically important difference. J Clin Epidemiol 1996;49:1223-4.
16. Deyo RA, Patrick DL. The significance of treatment effects: The clinical perspective. Medical Care 1995;33:AS286-AS291.
17. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. J Chron Dis 1986;39:897-906.
18. Liang MH, Larson MG, Cullen KE, Schwartz JA. Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. Arthritis Rheum 1985;28:542-7.

19. Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. Medical Care 1990;28:632-42.
20. Liang MH. Evaluating measurement responsiveness. J Rheumatol 1995;22:1191-2.
21. Bombardier C, Raboud J. A comparison of health-related quality-of-life measures for rheumatoid arthritis research. Control Clin Trials 1991;12:243S-256S.
22. Christensen L, Mendoza JL. A method of assessing change in a single subject: an alteration of the RC index. Behav Therapy 1986;17:305-8.
23. Stratford PW, Binkley J, Soloman P, Finch E, Gill C, Moreland J. Defining the minimum level of detectable change for the Roland-Morris questionnaire. Phys Ther 1996;76:359-68.
24. Stratford PW, Riddle DL, Binkley JM, Spadoni G, Westaway MD, Padfield B. Using the neck disability index to make decisions concerning individual patients. Physiotherapy Can 1999; Spring:107-12.
25. Wyrwich KW, Nienaber MA, Tierney WM, Wolinsky FD. Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. Medical Care 1999;37:469-78.
26. Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting standard error of measurement based criterion for identifying meaningful intra-individual change in health-related quality of life. J Clin Epidemiol 1999;52:861-73.
27. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J Consulting Clin Psychol 1991;59:12-9.
28. Jacobson NS, Roberts LJ, Berns SB, McGlinchey JB. Methods for defining and determining the clinical significance of treatment effects: description, application, and alternatives. J Consulting Clin Psychol 1999;67:300-7.
29. Stratford PW, Finch E, Solomon P, Binkley J, Gill C, Moreland J. Using the Roland-Morris questionnaire to make decisions about individual patients. Physiotherapy Can 1996;48:107-10.
30. McHorney CA, Tarlov AR. Individual patient monitoring in clinical practice: Are available health status surveys adequate? Qual Life Res 1995;4:293-307.
31. Stucki G, Daltroy L, Katz JN, Johannesson M, Liang MH. Interpretation of change scores in ordinal clinical scales and health status measures: the whole may not be equal to the sum of the parts. J Clin Epidemiol 1996;49:711-7.
32. Francis DJ, Fletcher JM, Stuebing KK, Davidson KC, Thompson NM. Analysis of change: Modeling individual growth. J Consult Clin Psychol 1991;59:27-37.
33. Beaton DE, Hogg-Johnson S, Bombardier C. Evaluating changes in health status: reliability and responsiveness of five generic health status measures in workers with musculoskeletal disorders. J Clin Epidemiol 1997;50:79-93.
34. Lydick E, Epstein RS. Interpretation of quality of life changes. Qual Life Res 1993;2:221-6.
35. Beurskens AJ, de Vet HC, Koke AJ. Responsiveness of functional status in low back pain: a comparison of different instruments. Pain 1996;65:71-6.
36. Riddle DL, Stratford PW, Binkley JM. Sensitivity to change of the Roland-Morris back pain questionnaire: Part 2. Phys Ther 1998;78:1197-207.
37. Drummond M, O'Brien B. Clinical importance, statistical significance and the assessment of economic and quality-of-life outcomes. Health Economics 1993;2:205-12.
38. Naylor CD, Llewellyn-Thomas HA. Can there be a more patient-centred approach to determining clinically important effect sizes for randomized treatment trials? J Clin Epidemiol 1994;47:787-95.
39. Wright JG. The minimal important difference: who's to say what is important? J Clin Epidemiol 1996;49:1221-2.
40. Wells GA, Beaton DE, Shea B, et al. Minimal clinically important differences: Review of methods. J Rheumatol 2001;28:406-12.
41. Buchbinder R, Bombardier C, Yeung M, Tugwell P. Which outcome measures should be used in rheumatoid arthritis clinical trials? Arthritis Rheum 1995;38:1568-80.
42. Greenfield S, Nelson EC. Recent developments and future issues in the use of health status assessment measures in clinical settings. Medical Care 1992;30 Suppl:MS23-MS41.