Methods for assessing responsiveness: a critical review and recommendations

Janice A. Husted\textsuperscript{a,\*}, Richard J. Cook\textsuperscript{b}, Vern T. Farewell\textsuperscript{c}, Dafna D. Gladman\textsuperscript{d}

\textsuperscript{a}Department of Health Studies and Gerontology, University of Waterloo, 200 University Avenue, Waterloo, Ontario, Canada
\textsuperscript{b}Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada
\textsuperscript{c}Department of Statistical Science, University College London, London, UK
\textsuperscript{d}University of Toronto, Centre for Prognosis Studies in The Rheumatic Diseases, Toronto Western Hospital, Toronto, Ontario, Canada

Received 31 March 1999; received in revised form 21 July 1999; accepted 6 October 1999

Abstract

A review of the literature suggests there are two major aspects of responsiveness. We define the first as “internal responsiveness,” which characterizes the ability of a measure to change over a prespecified time frame, and the second as “external responsiveness,” which reflects the extent to which change in a measure relates to corresponding change in a reference measure of clinical or health status. The properties and interpretation of commonly used internal and external responsiveness statistics are examined. It is from the interpretation point of view that external responsiveness statistics are considered particularly attractive. The usefulness of regression models for assessing external responsiveness is also highlighted. © 2000 Elsevier Science Inc. All rights reserved.

Keywords: Clinical trials; Outcome measures; Responsiveness; Responsiveness statistics

1. Introduction

It is widely argued that outcome measures in clinical trials should be reliable, valid, and responsive [1–3]. A reliable measure is one that tends to produce the same results when administered on two or more occasions under identical conditions [4–6]. Reliability is typically assessed in test–retest studies by analyses based on the kappa statistic or the intraclass correlation. A valid measure is one that measures what it was intended to measure [4–6] and is assessed by estimation of sensitivity and specificity, ROC curve analyses, correlation analyses, or regression models.

There does not, however, appear to be a consensus in the literature on what constitutes a responsive measure nor, correspondingly, how responsiveness should be quantified. A review of the literature suggests there are two major aspects of responsiveness, each having its own definition and strategies for assessment. We define the first as “internal responsiveness,” which characterizes the ability of a measure to change over a particular prespecified time frame. One widely used method of assessing internal responsiveness is to evaluate the change in a measure within the context of a randomized clinical trial involving a treatment that has previously been shown to be efficacious [7–13]. Any observed change in the measure is typically attributed to clinically relevant changes in health. Alternatively, change in a measure has been assessed using a single group repeated measures design, where patients are assessed before and after a known efficacious treatment (e.g., total hip arthroplasty, back surgery, physiotherapy). This strategy has frequently been employed to compare change in various health status measures [14–19]. The internal responsiveness of a measure, evaluated by either of these methods, will depend upon both the particular treatment and the particular outcomes used to determine treatment efficacy.

We use the term “external responsiveness” to define the second aspect of responsiveness. External responsiveness reflects the extent to which changes in a measure over a specified time frame relate to corresponding changes in a reference measure of health status. In this context, in contrast to internal responsiveness, the measure is not in and of itself of primary interest. Rather, it is the relationship between change in the measure and change in the external standard. One motivation for this is that if the relationship is strong (i.e., the measure is shown to adequately capture changes in the standard), the measure may be used instead of the reference measure as an outcome in future clinical trials. Another motivation is more general and is not based on the assumption that the measure under study should be a re-
placement for a standard measure. Rather, change in the standard is viewed as an accepted indication of a change in the condition of a patient. By accepted, we mean change that would be widely regarded by clinicians as meaningful and important change in clinical status. If the standard changes then it follows that some change in the measure under investigation would also be expected. Note that, unlike internal responsiveness, the external responsiveness of a measure will depend only on the choice of the external standard and not on the treatments under investigation. This implies that external responsiveness is a property of a measure and therefore it has meaning in a wider range of settings than the more context specific concept of internal responsiveness.

The lack of consensus on what “responsive” actually means, and how one should assess it, has led to a proliferation of responsiveness statistics, with investigators often reporting several within one study. This makes comparisons of measures across and within studies difficult or impossible [2,3,18–22]. Beaton et al. [20] write: “there is no gold ‘standard’ for summarizing responsiveness, although some consensus is needed . . . ; the literature demonstrates inconsistency in the methods used for calculating responsiveness statistics, and readers must be cautioned to examine the formulae amid adaptations made to the different statistics.” Thus, the most appropriate responsiveness statistic remains a matter of debate and, indeed, if there are different aspects of responsiveness that are of interest, more than one statistic may be reported. However, there seems to be several statistics that have been proposed and are used that purport to reflect the same thing. This has motivated our current investigation.

The purpose of this article is to examine the property of responsiveness from a foundational standpoint. Many of the issues that we discuss have been explicitly or implicitly raised by others [2,3,17,20,21]. Particularly relevant references in this regard are [7,18]. Our intentions in renewing discussion are to: 1) highlight the distinction between internal and external responsiveness; 2) clarify both the properties and interpretation of frequently used responsiveness statistics, and readers must be cautioned to examine the formulae amid adaptations made to the different statistics; 3) recommend the use of regression models to assess external responsiveness; and 4) provide directions for future research. Our illustrative example is drawn from the rheumatologcal literature, although the general principles we highlight apply to all disciplines in which responsiveness is important.

2. Notation

Here we define some notation that we will use subsequently to present the various responsiveness statistics. We assume research participants are assessed at two timepoints and let $X_1$ and $X_2$ denote their responses on the measure at the first and second assessments respectively. We let $D_1 = X_2 - X_1$ represent the change in the response on the measure over time, with positive (negative) values for $D_1$ representing increase (decrease) in the response over time. We let the expected mean change between baseline and follow-up assessments be denoted by $\delta = E(X_2 - X_1)$. Furthermore, we let $SD(X)$ denote the estimated standard deviation in the first (time 1) and second (time 2) measurements, assumed to be the same, and $SD(D_1)$ the estimated standard deviation in the change score for this measure. When another instrument is available we let $Y_1$, $Y_2$, $D_2$, $\delta$, $SD(Y)$ and $SD(D_2)$ be analogously defined. When it is important to indicate responses for individual patients we will introduce another subscript. Thus, for example, $X_{i1}(X_{i2})$ is the response from patient $i$ at time 1 (2) and $D_{i1} = X_{i2} - X_{i1}$ is the corresponding difference. The mean response on the measure at times 1 and 2 are $\bar{X}_1$ and $\bar{X}_2$, respectively, and the mean change score is $\bar{D}$. Again similar quantities may be defined for other measures. We let $n$ denote the total number of subjects.

3. Internal responsiveness

The most frequently used responsiveness statistics fall into this group.

3.1. Paired $t$-test

This test statistic is used to test the hypothesis that there was no change in the average response on the measure over the two time points, or more formally $H_0: \delta = 0$. The paired $t$-test statistic has been used to analyze data originating from a one-group repeated measures design [14,21]. The observed value for $T$ is:

$$t_0 = \frac{\bar{D}}{SD(D_1)/\sqrt{n}}.$$

With a reasonable sample size ($n > 30$), a value of $t_0$ greater than 1.96 indicates that the data provide evidence to reject the null hypothesis. It therefore can be concluded that a statistically significant change in the measure occurred over time. On this basis the measure is judged as responsive [21].

The paired $t$-test statistic has also been used to compare two or in more measures in the same patient group. For example, Liang et al. [14] use the relative efficiency index, calculated by squaring the ratio of paired $t$-tests for two measures, where one serves as a standard.

The paired $t$-statistic is focused exclusively on the statistical significance of the observed change in the measure. Statistical significance depends on the magnitude of the observed change, which is clearly relevant, but also depends on sample size and the variability of the measure. As a result, in spite of its past use, it is not a particularly appealing method for assessing responsiveness. As many users will recognize, sample size has nothing to do with responsiveness and, for valid comparisons, $t_0$ statistics must be based on samples of the same size. Both the magnitude of the change and its variability could be thought relevant to responsiveness, but it would be more useful to examine these...
3.2. Effective size I

In contrast to the paired t-test, effect size statistics, first proposed by Cohen [24], provide direct information on the magnitude of change in the measure, expressed in terms of some measure of variation. For this reason they have been widely recommended for use as indicators of responsiveness. The first effect size statistic we consider, frequently referred to as the standardized effect size [20,25], is given by:

$$ES_I = \frac{D_x}{SD(D_x)}.$$

This definition of effect size is the difference between the mean baseline scores and follow-up scores on the measure, divided by the standard deviation of baseline scores. Thus, a measure that has a high level of variability at baseline in relation to mean change scores will have a small effect size. Note that the standardization on the basis of the baseline scores has no link with the concept of responsiveness-Treatment (RT) coefficient [7] or an efficiency index. Alternatively, where there are only two observations of the measure in clinically stable subjects (e.g., multiple baseline measures prior to an intervention). Alternatively, where there are only two observations of the measure (e.g., patients’ global ratings of changes in health). Minimal clinically important change reflects the magnitude of change in the measure associated with an arbitrary definition of smallest important change on the external standard. For example, the minimal clinically important change may be estimated by the average change score among those patients rating some improvement in health minus the average change score among those patients rating no change in health [18].

Note that $ES_{II}$ may be preferred over the paired t-test because it removes the dependence on sample size. $ES_{II}$ provides an estimate of change in the measure, standardized relative to the between patient variability in change scores. Values of 0.20, 0.50, and 0.80 or greater have been proposed to represent small, moderate, and large responsiveness, respectively [16,20,22]. The use of the same benchmarks for $ES_I$ and $ES_{II}$, two quite different indices of responsiveness, is perhaps indicative of the confusion, or at least the lack of consensus, in the literature.

Some investigators have developed techniques to reflect uncertainty in $ES_{II}$ through construction of confidence intervals. Beaton et al. [20] base their confidence intervals on the assumption that distribution of $ES_{II}$ is approximately normal. Liang et al. [16] used the jack-knife procedure to provide an improved estimate of $ES_{II}$ as well as a robust estimate of the variability. Confidence intervals can then be constructed based on the Student’s $t$-distribution, thereby permitting qualitative comparisons between two or more instruments within a single group of patients.

3.3. Effect size II: standardized response mean

The standardized response mean (SRM) is another type of effect size and is widely used today (e.g., [16]). In the literature the SRM is also sometimes referred to as a Responsiveness-Treatment (RT) coefficient [7] or an efficiency index [9]. It is defined as,

$$ES_{II} = SRM = \frac{D_x}{SD(D_x)}.$$

As shown the $ES_{II}$ is a ratio of observed change and the standard deviation reflecting the variability of the change scores. Thus, a measure that has a high level of variability in change scores in relation to mean change will have a small SRM value. Note that the SRM can also be defined as a function of the paired $t$-test (or vice versa):

$$ES_{II} = t_0 / \sqrt{n}$$

or alternatively

$$t_0 = ES_{II} \times \sqrt{n}.$$
measure, which may arise from measurement error and learning effects. Guyatt et al. [2] argue that spurious change is reflected in the variability in score changes among clinically stable patients and that, to be responsive, a measure must be able to detect minimally clinically important change that exceeds any spurious change.

Similar to \( ES_I \) and \( ES_{III} \), values of 0.20, 0.50, and 0.80 or greater have been used to represent small, moderate, and large responsiveness, respectively [7,18]. Again the use of the same benchmarks for very different indices of responsiveness reflects the confusion in the area.

Despite its perceived superiority, the index is not a widely used statistic (at least in rheumatology). Beaton et al. [20] state this is due to the fact that minimally clinically important change is not yet known for a number of measures. Deyo et al. [21] explain that standardized methods for estimating this quantity have not yet been developed and in fact that minimal clinically important change for a measure may vary across different patient populations.

3.5. Interpretation of the paired \( t \)-test and effect size statistics

All the measures discussed in this section share the limitation that they do not relate changes in the measure to corresponding changes on an external clinical or health status measure at the individual patient level. With the exception of formula \( ES_{III} \) as above, the statistics simply examine the extent of change in the measure over two occasions. This is problematic as a statistically significant change in the measure may occur without a corresponding change in clinical or health status [10,17,29]. In other words, the observed change in the measure may not reflect important change in the condition of the patient. In spite of the common usage of the words, validity and reliability are not concepts that can be adequately described in a dichotomous way. In the context of health status measures and many clinical indices, perfect validity and reliability are unlikely [30]. Therefore, reliance solely on previous assessments of acceptable reliability and validity and internal measures of responsiveness, which are divorced from external clinical measures, may give an incomplete picture of the usefulness of the measures for the purposes to which the study of responsiveness is directed.

To circumvent this problem some investigators calculate a paired \( t \)-test or an effect size statistic only for patients who rate their health as changed (e.g., improved) [20,21,25]. The major weakness of this approach is that it does not involve, and indeed specifically precludes, a comparison of change in the measure between patients who report changes in health and those who do not. Hence, it does not directly examine the relationship between change in response on the measure and corresponding change in the external criterion. A useful measure must be known to reflect “no change” as well as it does “change.”

By its construction, \( ES_{III} \) seems to have some attributes of a statistic for external responsiveness (specifically, it is determined in part by patients’ assessment of their change in health status). Indeed, Norman et al. [7] state that \( ES_{III} \) or RR coefficient is based on the correlational approach. It is natural then to ask why it is classified here as a measure of internal responsiveness. While \( ES_{III} \) does use an external criterion of change to ascertain minimally clinically important change, it does not directly relate changes in the measure to corresponding changes in the external criterion at the individual patient level. Our measures of external responsiveness do relate to change at the individual patient level, as correlational measures must. If for \( ES_{III} \) we dichotomize patients as having changed or not changed on the external standard and use the entire sample of patients to relate change in the measure to corresponding change in the standard, this is closer to our measures of external responsiveness; for instance, receiver operator characteristic curves (ROCs) that are discussed in Section 4. Furthermore, some investigators use the difference between pre- and postscores on the measure itself, following a known efficacious treatment, to estimate minimal clinically important change [2,21]. When calculated in this manner, \( ES_{III} \) is similar in character to the paired \( t \)-test statistic and \( ES_I \) and \( ES_{II} \) statistics.

We have alluded to some other potential limitations. Others [9,11,20] have also mentioned the important difficulties associated with the interpretation of findings based on the paired \( t \)-test (e.g., does the observed value reflect the true change in the measure over time, or simply the size of the sample?) and \( ES_I \) (e.g., does the observed value reflect true change, or simply the variance of baseline scores?). Norman et al. [7] have also pointed out difficulties associated with \( ES_{III} \), which they refer to as the RR coefficient. They show that values of \( ES_{III} \) will typically exceed zero, even when there is no important mean change in the patient group, as long as there is a positive correlation between the measure and the external standard of clinical change. This occurs “because the RR measure is based on individual variation in change scores, it will always be possible to identify a subgroup of patients at the high end of the distribution of changes who have improved and another near the middle of the distribution who have stayed the same, regardless of the nature of the intervention (or non-intervention).” This limitation of \( ES_{III} \) reemphasizes the need for responsiveness statistics, which are based on the correlational approach to use all patients in determining the relationship between two measures. These limitations of the internal responsiveness statistics are secondary to our discussion so are not addressed further here.

4. External responsiveness

4.1. Receiver operating characteristic method

Deyo and Centro [17] were among the first to propose the assessment of responsiveness using receiver operating characteristic curves (ROCs) in rheumatology. In this con-
text responsiveness is described in terms of sensitivity (probability of the measure correctly classifying patients who demonstrate change on an external criterion of clinical change) and specificity (probability of the measure correctly classifying patients who do not demonstrate change on the external criterion) [17,18,21]. In other words, it assesses the ability of a measure to reflect both change and no change in the external standard. Sensitivity and specificity for each value of change in the measure are calculated, using standard formulas developed to assess the validity of screening and diagnostic tools [31]. This information is used to plot a ROC curve for the measure. Values for sensitivity and for false-positive rates (1 – specificity) are plotted on the y and x axis of the curve, respectively. The area under the ROC curve is then calculated (see [17] and [18] for details). It represents the probability that a measure correctly classifies patients as improved (worse) or unimproved (not worse). ROC curves are also used to rank the ability of competing measures to detect clinical change and to establish the change score in the measure that best classifies patients as improved (worse) or unimproved (not worse).

The ROC curve provides a very useful overview of the relationship between a measure and an external indicator of change. The concepts involved in its calculation are straightforward. Perhaps the major disadvantage of the ROC method is that the external clinical change score must be dichotomized (e.g., improved and unimproved or worse and not worse). This sacrifices information on the magnitude of change in the external criterion. It also means that a separate analysis is required to determine responsiveness to both improvement and deterioration [17].

4.2. Correlation

Correlation analyses are thought to be well suited to assess responsiveness and so appear frequently in the literature. The Pearson product moment correlation is typically computed based on change scores from two measures and takes the form:

\[ r_{xy} = \frac{\sum_{i=1}^{n} (D_{yi} - \bar{D}_y)(D_{xi} - \bar{D}_x)}{n}. \]

This quantity lies between \(-1\) and \(+1\) with positive values indicating a positive association and negative values a negative association. Correlation measures of this sort indicate how change on two measures vary together. Often X will be a new health-related or quality of life measure and Y a traditional clinical outcome (e.g., articular index, grip strength, patient or physician global rating of function) [17–19,25,32]. If \(r_{xy}\) approaches \(+1\) then X is thought to capture the information on Y (i.e., it responds to changes on Y).

As pointed out by Armitage [33], there are some problems with the use of correlation coefficients. Perhaps the most serious, relevant for its use as a responsiveness statistic, is that its value may be affected by selection of particu-

lar values of one variable. For example, if Y and Z are bivariate normal variables with correlation \(\rho\) then restricting the range of Z by removing extreme values will tend to decrease the correlation coefficient so that it is less than \(\rho\). Thus, two clinics, say, with a different patient mix may estimate different correlation coefficients even though the basic relationship between two variables is the same in both clinics.

A second problem is that correlation measures the closeness to a linear relationship and the relationship between two measures may be close but nonlinear.

The attractiveness of the correlation coefficient in assessing responsiveness is that it examines whether a measure is responsive relative to a specific alternative outcome. This reflects an interest in how well changes in one measure predict changes in another. If so, it is natural to consider regression modeling, which has the additional advantage of not being subject to the problems associated with the correlation coefficient.

4.3. Regression models

Suppose, as before, the X is a new health-related quality of life measure and Y is a traditional clinical outcome. Consider the simple linear regression model

\[ D_{yi} = a + b d_{xi} + e_i \]

where \(e_i, i = 1, ..., n\) are independent \(N(0, \sigma^2)\) variables. The quantity \(a\) represents the mean change for \(Y\) when no change is observed in \(X\), and \(b\) indicates the average increase in the change in \(Y\) associated with a one unit change in \(X\). Values for \(b\) near zero suggest large changes observed in \(X\) may not be accompanied by changes in \(Y\) and large values of \(b\) mean the associated changes in \(Y\) will also be large. Of course, \(a\) and \(b\) are scale dependent (in the sense that they depend on the units used for measuring \(X\) and \(Y\)). The statistical significance of \(b\) should be examined although this does, of course, depend on study design. For the purpose of comparing different measures within a single study, some authors suggest the use of standardized regression coefficients [34,35]. As mentioned earlier, correlational analysis link naturally with these regression models but regression models have the potential to be more informative.

Husted et al. [36] considered a generalization of the previous regression model, which takes the form

\[ D_{yi} = a + b_1X_{1i} + b_2X_{2i} + e_i \]

where \(e_i, i = 1, ..., n\) are independently and normally distributed \(N(0, \sigma^2)\) variables. Here if \(b_1 = -b_2\) we retrieve the earlier regression model. Allowing \(b_1\) and \(b_2\) to be estimated from the data provides a better tool for the study of change and, therefore, improved inferences about external responsiveness.
Note that the model can be generalized further, if appropriate, by the inclusion of \( Y_i \) as an additional predictor variable in the model for \( D_{ij} \) [36].

We believe viewing external responsiveness as a regression problem is both natural and illuminating. Unlike the various alternative procedures that have been proposed, technically it generates an easily interpreted index in the form of the regression coefficient \( b \) (or its standardized form). Furthermore, one can carry out a goodness-of-fit assessment to check the plausibility of the regression model. It is possible that the relationship between a new measure and a traditional clinical measure is not linear, and in this case the model may be generalized. For example, one could add different powers of \( X \) to the model or control for confounding variables, or add interaction terms if external responsiveness is thought to vary across subgroups of patients. Thus, a regression analysis provides a comprehensive examination of the relationship between changes in an external standard and a measure under study and does so in a manner that can be replicated by other investigators. Finally, we note that the above discussion is not limited to normal regression models and that for external standards that are binary or categorical the same approach can be used with other types of regression models.

5. Responsiveness in psoriatic arthritis

The data originated from the University of Toronto psoriatic arthritis out-patient clinic [37]. Between 1994 and 1996, 70 patients (27 women and 43 men) completed three health status measures—the HAQ [38], AIMS2 [39], and SF-36 [40]—on two occasions, approximately 12–18 months apart [41].

Here we compute responsiveness statistics for the physical functioning dimension of the HAQ, the AIMS2, and the SF-36 in this sample of 70 patients. For the external responsiveness statistics, a health transition index that asks about changes in health is used as the reference measure. Patients rated change in health between the two testings, using a 5 point ordinal scale (1 = much better than a year ago; 2 = somewhat better than a year ago; 3 = about the same; 4 = somewhat worse than a year ago; and 5 = much worse than a year ago). When we calculate \( ES_{III} \) for each of the health status measures, minimally clinically important change (the numerator) is defined as the mean change in physical functioning for patients who reported their health was somewhat better than a year ago minus the mean change in physical functioning for patients who reported their health was the same. The denominator is the standard deviation of change in physical functioning in the subgroup of patients who reported their health was the same. For the ROC analysis, patients’ assessment of change in health status is coded as a binary variable. Patients who rated their health as somewhat better or much better than a year ago are classified as improved and patients who rated their health the same, somewhat worse, or much worse are considered not improved.

The health transitional index indicated that 49% of the patients experienced health changes between the two testings. Of these patients, 23% reported improvement and 26% reported deterioration. Significant changes in disease activity and severity were also noted by the examining rheumatologist. Approximately 79% of the patients experienced a change in the number of active joints, with 51% showing clinical improvement in inflammatory disease (a 30% decrease in number of active joints) and 29% showing clinical deterioration (a 30% increase in number of active joints). Thirty-six percent of the patients experienced progression in deformity status, with 30% showing an increase of 1 to 4 damaged joints and 6% an increase of 5 to 9 joints.

The internal responsiveness statistics suggest that there was no corresponding change in average physical functioning, as measured by either the HAQ, the AIMS2, or the SF-36 (Table 1). The observed value for \( T \) (\( t_o \)) for each of the measures is less than 1.96, which indicates a lack of significant change in average physical functioning between the clinic visits. With one exception the various effect sizes for each of the measures are approximately 0.20 or below, also suggesting that any observed change in physical functioning was not clinically meaningful.

On the other hand, the external responsiveness statistics indicate something different. The ROC analysis indicates that physical functioning change scores for each of the measures discriminates moderately between improved and not improved patients. The shape of and the area under the ROCs also suggest that the SF-36 may be a better discriminator than the HAQ and AIMS2 (see Table 1 and Fig. 1). The correlation analyses also indicated that change scores in physical functioning for the HAQ, the AIMS2 and the SF-36 were related to change in perceived health. The value of the product-moment correlations and the lower limit of the 95% confidence intervals suggest that both the HAQ and AIMS2 were weakly correlated with perceived change in health, whereas the SF-36 was moderately correlated with change in health.

In the linear regression analysis the individual regression coefficients were statistically significant and showed the magnitude of change in health associated with one unit change in physical functioning for each of the measures. For instance, between the two testings a one unit change in the HAQ physical functioning change score, on average, corresponds to a change of approximately one category in the health transitional index.

The amount of variation in the change scores for health explained by the three measures was 14%, 11%, and 31%, respectively. This information, which is based on \( R^2 \) statistics, provides the additional information that while there is a demonstrable relationship between the change in health scores and the change in the health status instruments, this relationship would not necessarily be good enough for predictive purposes. This may or may not be relevant in considering the usefulness of the instruments for specific purposes. It may be thought, for example, that a high \( R^2 \) is only
required if the measure is to be used as a replacement for the external standard.

This example serves to illustrate that the use of responsiveness statistics, divorced from an external criterion, may lead users to inaccurate or, at least, incomplete conclusions. It also illustrates the usefulness of regression analysis for assessing responsiveness. Unlike both the Guyatt index and ROC methods, separate analysis are not required to deter-

<table>
<thead>
<tr>
<th>Physical function change scores</th>
<th>Internal responsiveness</th>
<th>External responsiveness$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Paired t-test</td>
<td>Effect size I</td>
</tr>
<tr>
<td>HAQ</td>
<td>-0.617</td>
<td>-0.055</td>
</tr>
<tr>
<td>AIMS 2</td>
<td>0.261</td>
<td>0.005</td>
</tr>
<tr>
<td>SF-36</td>
<td>1.056</td>
<td>0.086</td>
</tr>
</tbody>
</table>

$^a$External criterion is patient assessment of change in health (1 = much better than a year ago; 2 = somewhat better; 3 = about the same; 4 = somewhat worse than a year ago; and 5 = much worse).

$^b$Numerator equals the mean change in instrument response for patients who reported their health was somewhat better minus the mean change in response for patients who reported their health was the same. Denominator is the standard deviation of change in response for the latter group of patients.

$^c$External criterion is coded as a binary variable. Patients who rated their health as somewhat better or much better than a year ago are classified as improved and patients who reported their health as the same, somewhat worse, or much worse are classified as not improved.

$^*P < 0.01$, $^{**}P < 0.0001$.

Table 1
Responsiveness statistics for the physical functioning dimension of the HAQ, AIMS 2, and SF-36 in a sample of 70 patients with psoriatic arthritis

Fig. 1. Receiver operating characteristics curve for the HAQ, AIMS2, and SF-36.
mine the extent to which change in a measure is related to both improvement and deterioration on the external criterion. Unlike both the correlational and ROC methods, a regression analysis estimates the magnitude of change in the external criterion that is associated with one unit (or specific units) of change in the measure.

6. Discussion

Further discussion of responsiveness is warranted. We have attempted to provide a structured framework within which such discussion can take place. Here we offer some preliminary thoughts based on our review of the literature.

The distinction between internal and external responsiveness is important. Stucki et al. [18] make a distinction in their work by referring to internal responsiveness simply as responsiveness and external responsiveness as discriminative ability. Kirshner and Guyatt [42] make a similar distinction, referring to internal responsiveness as responsiveness and to aspects of external responsiveness as longitudinal construct validity. We, nevertheless, feel there is a difference between longitudinal construct validity and external responsiveness. In a study of longitudinal construct validity external measures are selected to establish both convergent and discriminative validity. By contrast, in a study of external responsiveness the external measure is selected to represent an accepted indication of change in the condition of the patient. For instance, Felson et al. [43] recently suggested the use of a single uniform definition of improvement for clinical studies in rheumatoid arthritis. On this basis we feel there is some advantage to maintaining the term responsiveness as long as a clear distinction between the two aspects of responsiveness is made.

Studies using internal responsiveness statistics can be used to qualitatively assess which measures are more “internally responsive” than others. The only potential difficulty in this regard relates to the somewhat subjective choice of the minimal clinically important change for each measure, which is necessary for the calculation of ESIII scores. The use of internal responsiveness statistics, other than ESIII, for comparisons across studies is difficult, however. The calculation of the statistics is sufficiently specific to each study that there is no well-defined interpretation, independent of study design, which can be given to particular values of the statistics. As well, it must be remembered that the comparison is made on a population level. Where a measure has imperfect reliability and validity (which is often the case), the use of the statistics may give an incomplete picture of a measure’s ability to reflect clinical change at the individual patient level.

The choice between internal responsiveness statistics for use in a study should be based on their structure. For example, a decision would be made between ESII and ESIII based on what is felt to be the appropriate “standardization” for the mean difference. Guyatt et al. [2] argue that the between-subject variability of the individual changes in score over time is the appropriate standardization. We tend to agree with this view and thus favor the selection of ESIII over ESII. While the use of the paired t-test statistic for comparative purposes in a single study is valid, its dependence on sample size suggests that it should be complemented by the use of some other measure as well.

It is from the interpretation point of view that external responsiveness statistics are particularly attractive. In general, they characterize a relationship between change in an external standard and change in an outcome measure under study at the individual patient level. This relationship is examined in the context of a specific study, but, in the usual scientific sense, is generalizable across studies. That is, in another study of similar patients, the same relationship should be observed. As well, the interpretation of a particular value of the statistic is defined independent of the study and thus the results of different studies can be compared.

The use of regression models provides a very general approach to studying the relationship between two variables and therefore should be part of a study of external responsiveness. Some individuals may not find the use of regression coefficients as a summary measure of external responsiveness as intuitive as they would wish. In this regard ROC curves and correlation coefficients may be preferred. The limitations of these methodologies would be lessened if they were complemented by a regression analysis, however. It is to be hoped that with use, the understanding of regression coefficients in the community will increase.

While the interpretation of regression-based external responsiveness statistics is well defined, the reliance on an external standard may be viewed as a weakness as well as a strength. To illustrate, we may have a group of patients with important change over time in the sense that their disease status does “truly” change. We have defined our external standard as an accepted indication of change in a patient’s condition and have assumed that this measure would change at least for some patients. A strong correlation between the standard measure and a new measure should then imply that some change in the new measure will also be seen. However, if the selected standard does not capture the observed important change in the patients, and neither does the new measure, then these measures may still be correlated, leading users to inaccurate conclusions about external responsiveness of the new measure. This example highlights how important the choice of the external standard is for studies of external responsiveness. It also points out that some individual variation in the external standard must be present for a study of external responsiveness to be sensible. If there is no variation in the standard (i.e., all patients “improve”) this makes it impossible to examine how change in a measure and change in the external standard vary together.

A further weakness of relying on an external standard to assess responsiveness is that a new outcome measure may be designed specifically because it reflects a different aspect of patient status than currently available measures. While its
relationship with measures known to be responsive would still be of interest, this may not answer all the relevant questions. In this case, the use of internal responsiveness statistics may be required. In this regard there is motivation for more research into the design of internal responsiveness statistics that reflect patient level changes.

Finally, we would like to address an important issue raised by Norman et al. [7]. These investigators argue, and we agree, that responsiveness measures based on variation in change among patients in a cohort (e.g. $E_{gm}$ and our external responsiveness statistics) have no direct relationship to overall treatment effect in the same cohort of patients. On this basis it is concluded that such measures of responsiveness “yield little information about the ability of an instrument to detect treatment effects, and should not be used as a basis for choice of an instrument for applications to clinical trials.” Here we tend to differ. We feel that evidence on external responsiveness can be useful for selecting outcome measures for clinical studies. External responsiveness statistics measure responsiveness to an external standard, defined here as an accepted indication of change in the condition of a patient. If there is no change in the external standard, then it is assumed that there is no treatment effect and therefore the new measure should not be responsive to that particular treatment.

Acknowledgments

Supported by the Medical Research Council of Canada. The authors would like to thank Dr. Gordon Guyatt and an anonymous referee for helpful referee.

References


