Assessment of change in clinical evaluation

Middel, Lambertus Johannes

4

How to interpret the magnitude of change in health-related quality of life? A study on the use of Cohen's thresholds for effect size estimates.

Berrie Middel, MSc., Eric van Sonderen, Ph.D.

Submitted

ABSTRACT

This paper aims to identify problems in the evaluation of the magnitude of treatment-related change, or responsiveness of health status or health-related quality of life instruments, which are induced by standardizing change over time with the standard deviation of the difference score. This effect size is widely used and is represented as the Standardised Response Mean (SRM), and interpretation is problematic when it is used to estimate the magnitude of change over time with Cohen's rule of thumb for effect size (ES) which is based on standardisation with the pooled standard deviation. In the case of standardizing mean change with the SD of that change, application of the well-known cut-off points for pooled standard deviation units ('trivial' (ES < .20), 'small' (ES ≥ .20 <. 50), 'moderate' (ES ≥ .50 < .80), or large (ES ≥. .80) may lead to over- or underestimation of the magnitude of change over time due to the correlation between assessments.

.

Keywords: Responsiveness, Health Status, Sensitivity to change, Methodology, Effect size, Standardised Response Mean

4.1 INTRODUCTION

In the practice of health-related quality of life research, most researchers remain primarily interested in the statistical significance of the change in health-related functional status or quality of life in pre post designs. In combination with, e.g., the T-test approach, substantial effects can be detected [1-3] with an estimate of effect size. If a p-value is annotated as statistically significant, rejecting the null hypothesis does not imply an effect of important magnitude; likewise, a non-significant **p**-value does not indicate a trivial result, [4-7] although some researchers implicitly deem more important those results with smaller p-values.

In the last decade, however, a growing number of longitudinal intervention studies are focussed on questions like "If the change between baseline and outcome is statistically significant, what can we say about the magnitude (or amount) of change over time that has been detected? Can we interpret this difference in terms of an important difference or as a relevant (substantial) change?" To answer these questions, the responsiveness, i.e. the ability of quality of life outcome measures to detect change over time, has become crucial in the past decade. However, the responsiveness estimation is neglected in many clinical studies in which it could give information on the importance of change due to treatment effects supplementary to the statistical significance of change over time (e.g. before and after intervention) [8,9] Reporting effect sizes without appropriate statistical tests and associated p-values is misleading and potentially dangerous when the number of observations that is required to detect a difference has not been estimated with a power analysis. Effect size statistic should be provided to supplement (not as a substitute for) statistical testing, and only then, when the outcome is sufficiently extreme from what would have been expected on the basis of chance ($p < \alpha$).

Noteworthy in this respect is that in the field of psychological research, editorial policy indicates that "until there is a real impediment to doing so, authors should routinely present an effect size estimate along with the outcome of a significance test". [10,11]

Several quantitative indices have been developed [10-20] that belong to this family of effect sizes or standardized differences, each calculated with a different denominator in the

($\bar{X}1 - \bar{X}2$ /SD) formula, namely the SD of stable subjects, the SD of the baseline assessment, the SD of the observed difference score and the pooled standard deviation (SDp). Obviously, there is no consensus on how to declare a difference in terms of standard deviation units. Only in a small number of publications is this lack of consensus on the most appropriate effect size indicator signalled. [21-25] Despite the fact that different opinions exist on the method to estimate magnitude of difference

between groups or the magnitude of change within groups, researchers use the straitjacket of thresholds Cohen provided us with some 30 years ago. [26] However, these thresholds are taken for granted by many researchers for every version of effect size index. With regard to the correct use and interpretation of effect size indices as estimates of treatment related magnitude of change, we must revisit some basic assumptions:

1. the ES is developed and elaborated by Cohen to estimate power or the necessary sample size to detect relevant change with the basic principle of independent, equal size samples with common within-population standard deviation σ ;
2. in the case that this ES is used in paired samples or in a repeated measurement-design it must be adjusted for correct use of power tables and sample size tables;

4.1.1. Independent samples

Cohen represented the effect size (ES) on some dependent or outcome measure used in an experiment in terms of the difference (using the symbol d' to denote this ES) between the treatment and control group expressed in units of common within-population standard deviation (in samples this standard deviation is estimated with the pooled standard deviation) as follows:

[A]

$$ES = d' \quad \frac{\left(\bar{X}_{treatment} - \bar{X}_{control}\right)}{\sigma}$$

With this estimate of effect size, after analysing a wide sampling of behavioural research, Cohen developed his rules of thumb and reported that effect of .8σ being on the large end of the range, .5 σ was the medium, and .2 σ was at the small end of the range. [27]

4.1.2. Dependent samples or paired observations

The difference or change in matched observations within subjects is standardized by the common within-population σ, according to Cohen's [1977,p.13], but due to the removal of the variation in many extraneous characteristics of the subjects, the index must be adjusted (see appendix), dividing d' by $\sqrt{(1-r)}$. Cohen used the symbol d to denote this adjusted ES. [1]

1 As we will demonstrate, the effect size d is equivalent to the Stadardised Response Mean (SRM), i.e. mean change or difference divided by the standard deviation of that change of difference (see appendix)

[B]

$$d = \frac{d'}{\sqrt{(1-r)}}$$

d' = effect size for independent samples
d = adjusted effect size
r =correlation between baseline and outcome

This √(1-r) – correction of the denominator of formula A is necessary for a proper use of power and sample size tables since these assume 2(n-1) degrees of freedom where, in the case of paired observations, only n-1 are actually available. [26] This consequence for power and sample size estimation is something different from the use of the effect size d in evaluating efficacy of a new treatment in terms of amount of change in health status, which was not the aim of Cohen's work. Therefore, we did not abstract data from effect size estimates of health-related quality of life scales when they were used for the sole purpose of power analysis to draw conclusions from the results of the statistical analysis, or to answer the question whether the investigators had sufficient sample size to allow the detection of a relevant difference. [10,28,29]

Effect Size as an evaluative indicator of magnitude of difference inhealth-related functional status: Independent samples versus repeated measures

When effect sizes are calculated as the standardized difference in mean score to evaluate the efficacy of a new treatment with the use of Cohen's thresholds, for example between a treatment group and a control group, formula [A] should be used. The effect size can be calculated by pooling the estimates (pooled standard deviation) derived from sample data. In contrast to this independent sample case, effect sizes are also used in evaluation studies (pre- post study designs) as estimates of the responsiveness of (for example) a new outcome measure. Effects are often used to give meaning to change over time in terms of 'trivial' (ES < .20), 'small' (ES ≥ .20 < .50) ,'moderate' (ES ≥ .50 < .80) or 'large' (ES ≥ .80) change. Cohen [26] introduced this 'matched pairs' effect size (see appendix equation A2), which was later renamed the standardised response mean (SRM) by Liang et al. [28] to avoid confusion concerning other effect size indices. However, several researchers seem to have adopted the idea that **every** standardized difference is subject to Cohen's definitions of trivial, small, moderate and large effect. Such a belief could lead to

misinterpretations in studies focussing on treatment-related outcome in paired samples since these cut-off points of the magnitude of the difference were not established as a rule of thumb with the effect size d (dependent samples) but with the index d' (independent samples). Thus we argue that Cohen's thresholds are based on the assumption of common within-standard deviation (with matched pairs sample data we use the raw within-group pooled SD), resulting in an effect size we annotate as ES_P. Consequently, in matched pairs studies these thresholds cannot be used interchangeably for the SRM due to the role of the correlation between repeated measures or paired samples. In this article the attention is focussed on the standardized change in mean score between two points in time **within** a single group, estimated with the within-group effect size. In relation to the use of Cohen's rule of thumb for effect size interpretation, we evaluate the consequences of the calibration of the SRM with the ES_P and the role of the correlation between pre and post test scores.

To investigate how serious discrepancies can appear in effect size interpretation we first elaborate a theoretical example and used a sample of studies to evaluate the seriousness of these differences in practice. To evaluate the seriousness of the discrepancies between SRM and ES_P, the correlation of the subject's repeated measurements was needed. Empirical data were collected for the purpose of secondary analysis to draw conclusions in terms of the relative size of the SRM to the ES_P in relation to the size of the correlation. Applying Cohen's thresholds, which are based on the pooled estimate of effect, to interpret the SRM on the one hand may lead to similar results or subtle and trivial differences, but on the other hand also to meaningful shifts in classification of the amount of estimated change. In this article we analysed 148 SRMs interpreted using Cohen's rule of thumb and compared these SRMs with Cohen's ES_P from which these thresholds were derived. Furthermore, we calculated for the range of the correlation coefficient 0.01 to 0.99 the SRM adjusted for Cohen's cut-off points 0.20, 0.50 and 0.80 of the pooled effect size.

4.2 MATERIALS AND METHODS

To study the consequences of the impact of the association or correlation between repeated measures, we restrict the analysis to two effect size indices suitable for the evaluation and interpretation of magnitude of change over time (or responsiveness) within one group, namely the SRM and the ES_P. In this study we use the pooled SD

(SDP) as the standardizing unit (denominator) of mean change score over time (nominator) to calculate the effect size (ES_P)[2].

[C]

$$ES_p = \frac{\overline{X}_{change}}{SD_{(pooled)}}$$

The ES_P introduced by Cohen was made comparable to the SRM where the $SD_{(X\text{-}change)}$ is used as the denominator in which, as we will demonstrate below, the correlation between baseline and outcome scores is involved.

The SRM is the ratio between the mean change score and the variability (the standard deviation) of that change score within the same group.

[D]

$$SRM = \frac{\overline{X}_{change}}{SD_{(Xchange)}}$$

The relationship between ES_P (d') and SRM (d) and the correlationbetween baseline and outcome scores

One of our purposes was to get an indication of how the SRM varies in accordance with the size of the correlation between pre and post test scores when the correct pooled effect size estimate is used. An example may illustrate the role of r, the correlation of a person's health status measurements over time: In a study in which the outcome of a medical intervention was evaluated with a health-related quality of life measure, and in the case of improvement, a lower mean score after intervention was hypothesized. The investigator finds at baseline a mean score of 11.12 with a standard deviation of 4.43 and a mean score of 9.16 (SD: 4.88) at follow up. The estimate of the common within-standard deviation, which is the square root of $(SD_{baseline})^2 + (SD_{outcome})^2 / 2$), thus 4.66, and the pooled effect size (ES_P) is then 0.420 (11.12 – 9.16 /4.66). Before we compare the ES_P and SRM in relation to the correlation between repeated measurements, we must solve the problem of the

[2] $\text{effect size} = \frac{\left(\overline{X}_{baseline} - \overline{X}_{outcome}\right)}{\text{pooled SD}}$ where $\text{pooled SD} = \sqrt{\frac{(SD_{baseline})^2 + (SD_{outcome})^2}{2}}$ for $N_{baseline} = N_{outcome}$

equation of both formulas C and D. According to Cohen, the difference between means for **dependent** samples is standardised by a value "which is √2 (1-r) as large as would be the case were they independent". Cohen 1977, p.49
From equation A4 in the appendix, (d'/√2) / √(1-r) is equivalent to the SRM and alternatively SRM * √ 2 * √(1-r) is equivalent to d' and both indices will vary with the size of r . In table 4.1 we have elaborated the hypothetical example in which the effect size ES_P (d') = 0.42, is transformed into the SRM for a series of values of r. Both effect sizes are equal in the case that r = 0.50): ES_P = (0.42/√2) / √(1-0.50) = SRM, and the SRM for r .50 is then (0.42/1.41) / 0.71 = 0.42. In table 7.1 it is shown that the SRM gets larger for larger values of r. For example, an effect size of 0.42 indicating ' small effect' corresponds with a 'medium effect' (SRM = 0.50) if the correlation between the repeated measurements is approximately .64. This small effect estimated with the ES_P corresponds with a 'large effect' (SRM ≥ .80) if this correlation is approximately .86.

Table 4.1 *The conversion of an effect size calculated with the pooled SD (ES_P) of 0.42 into a SRM with correlation coefficients ranging from .00 - .90*

corr.	.00	0.10	.20	.30	.40	.50	.60	.65	.70	.80	.86	.90
$\frac{(.42/\sqrt{2})}{\sqrt{(1-r)}}$	.297	.313	.332	.355	.384	.420	.470	.502	.543	.664	.794	.940

4.2.2. The sample of studies

In examining the role of the correlation in the estimation of a within-group effect size index, we searched Medline and Psyclit for the years 1984-1999 and Current Contents for 1996 and 2000. We searched for studies with key words 'quality of life', 'health status', and 'questionnaire' and articles were scanned for the terms 'responsiveness', or 'sensitivity to change'.
The primary selection consisted of 151 publications and showed the existence of differing opinions about the appropriateness of the effect size as originally proposed by Cohen, which has led to the introduction of new methods of estimating the magnitude of change assessed over time. Due to the variation in the definition of the mean change scores in the nominator and in the definition of the standard deviation in the denominator of the effect size index formula, not all the studies were appropriate for this study. Therefore, 29 studies suitable for our analysis were selected using the following inclusion criteria:

1. the research should encompass repeated (self reported) health outcome assessments evaluating change within one group (paired observations);

2. a SRM must be represented accompanied with Cohen's thresholds;
3. health outcome must be assessed by (self reported) questionnaires with a disease specific or general mode.

Some other causes of the large reduction to 29 studies appropriate for the purpose of this article are:

- standardised Response Means were used without referring to Cohen's thresholds;
- missing information inhibited us from calculating the effect sizes needed;
- the topics of responsiveness and sensitivity to change were discussed purely from methodological or statistical perspective.

4.3 RESULTS

The original sample comprised 142 or scales belonging to scattered dimensions of health-related quality of life or health status measures. The current selection of 29 papers was determined by the condition that, a SRM had to be represented with referral to Cohen's thresholds for interpretation of change magnitude. [1,21,22,28,30-35,35-53] These 29 publications comprised 411 Standardised Response Mean indices which sizes were interpreted with referring to Cohen's thresholds. From this sample of SRM indices, 148 were published together with sufficient information to estimate the ES_P calculated with an estimated correlation coefficient (r) (see Appendix equation A5). The correlation coefficient is needed to compare the SRM that was shown with reference to Cohen's thresholds, with the ES_P as the correct yardstick for SRM interpretation. Additionally, 263 Standardised Response Means were detected with the investigator's reference to Cohen's rule of thumb (which is derived from pooled estimates of standard deviation) but unfortunately, not sufficient information was given to estimate the correlation between assessments.

4.3.1. The classification of treatment effect with the SRM with Cohen's thresholds for ES_P

In the interpretation of these two effect size indices, the thresholds proposed by Cohen [26-54] as operational definitions of magnitude of change cannot be used interchangeably as is generally assumed. Table 4.2 summarizes the results of the SRM's substituted into ES_P 's (using equation A4 in the appendix). It shows clearly that mistakes can be made in the classification of the magnitude of detected change in health-related quality of life if Cohen's rule of thumb of the ES_P is assumed in the interpretation of the SRM. In 148 SRM's that were adjusted, the magnitude of the

correlation coefficient caused no change in classification in 77%. Approximately, 34 of the 148 estimated effect sizes (23 percent) did not fall in the category indicating the same magnitude of change. Underestimation of effect size according to Cohen's thresholds occurred in 5 SRM's (3.4%), whereas 29 SRM's (19.6%) were overestimates of effect size.

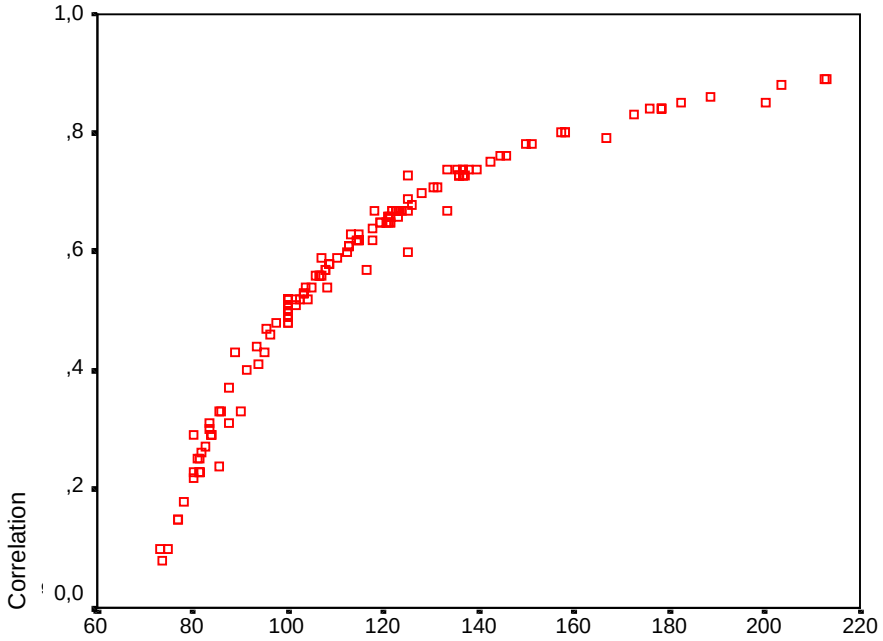
Table 4.2 *Similarities and differences between the Standardised Response Mean (SRM) and pooled effect size (ES_P) interpreted using Cohen's thresholds (N=411: 148 + 263)*

Calculated by equation A4[1]						Equation A4[2] not applicable
ES_{pooled}	ES < .20 Trivial effect	ES ≥ .20< .50 Small effect	ES ≥50< .80 Medium effect	ES ≥ .80 Large effect	Total	
SRM						
< .20	43	2			45	33
≥.20 <. 50	6	35	2		43	92
≥50 <. 80		11	13	1	25	69
≥ .80			12	23	35	69
total	49	48	27	24	148	263

[1] See appendix.
[2] SRM used with interpretation according to Cohen's thresholds for ES_{pooled}

To get a better understanding of the role of the calculated correlation between baseline and follow-up score in the relationship between these two effect size indices, we have, for these 148 estimates, expressed the SRM as the percentage of the value of the ES_P . In figure 4.1 it is shown that the SRM covers the ES_P 100% at the x-axis with the calculated r = .50 at the y-axis for each of the instrument scales of which the pre-post test correlation r was recalculated. The depicted curve shows, irrespective of the values of the effect sizes estimated in our sample of health-related quality of life scales, that the relative distance from the SRM to the ES_P varies with the size of the baseline-follow-up correlation.

Figure 4.1 *The relationship between the correlation and the relative ratio of the SRM and ES_P (N=148)*

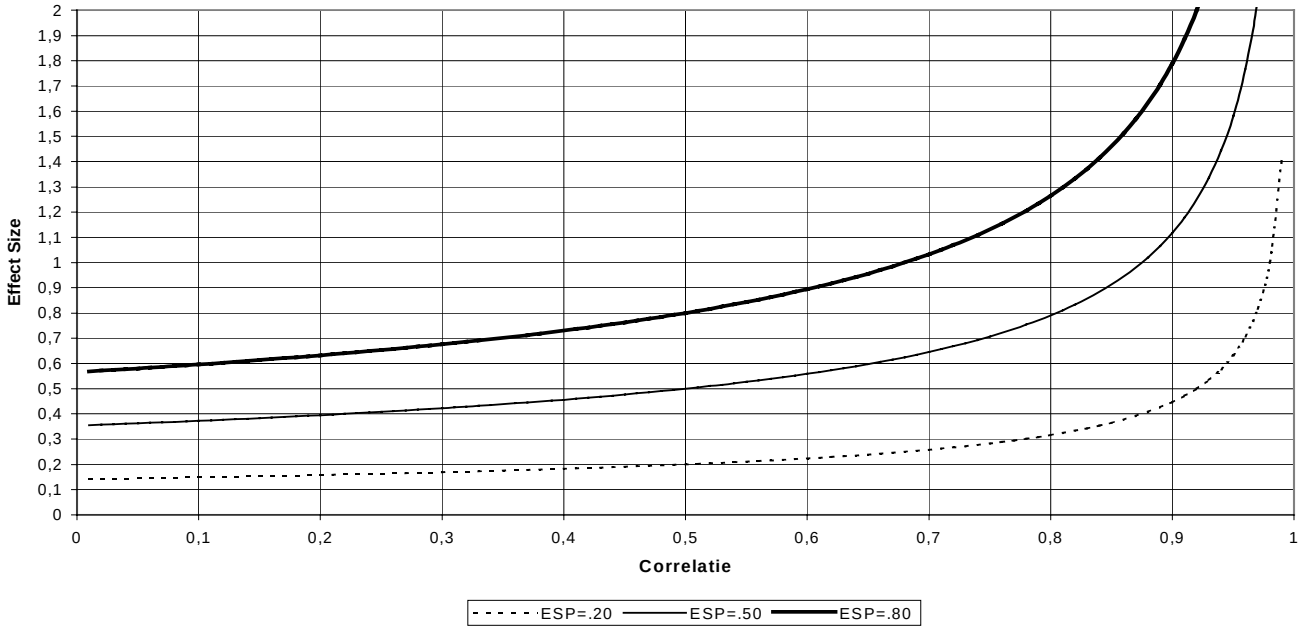


(SRM / ES_P * 100)

To avoid invalid interpretations in the evaluation of responsiveness with SRM index we have, for every value of the correlation between baseline and follow-up score, calculated the corresponding ES_P 's for Cohen's thresholds of .20 = small, .50 = medium, and .80 = large. Indices that lie within the interval that corresponds with these thresholds are not depicted. To classify the magnitude of change estimated with the SRM more precisely, this effect size index is adjusted for every value of the correlation coefficient (r) between baseline and follow-up assessments and brought into line with Cohen's thresholds for effect size. Figure 4.2 shows that SRM' s of 0.20, 0.50 and 0.80, don't deviate after calibration with Cohen's ES_P taken as the original standard, when r =.50. A SRM of 0.20 must be tagged as trivial effect as long as the correlation coefficient ranges from r=.01 to r=.49. With large corresponding correlation coefficients a small SRM of 0.20 must be tagged as moderate (.20/√2 / √1- .92 = .50) or large (.20/√2 / √1-.97 = .80) The class midpoint 0.35 of the 'small

effect' range of effect (not depicted) has to be classified as moderate or large effect with correlation coefficients of 0.76 (.35/√2 / √1-.76 = .50) and 0.91 (.35/√2 / √1-.91 = .80) respectively.
SRM's of 0.80 has to be tagged as 'moderate' effect if the correlation ranges from r = 0.01 to 0.49. The SRM ≥ 0.80 cannot drop below the cut-off points of small and trivial due to the correlation magnitude between baseline and outcome measurements. 'Moderate' effect (SRM = 0.50) must be tagged as 'small' if the correlation between repeated measures is below 0.49 and has to be classified as 'large' in case of r = .81. The class midpoint 0.65 (not depicted) of the 'moderate effect' range of effect must be valued as 'small' with a r = 0.14 (.65/√2/√1-.14 = .49).
In contrast with the fixed threshold values .20, .50 and .80 in figure 4.2, in the analysis of 148 effect size estimates from which the correlation of a person's health status measurements over time was calculated, we found SRM values ranging from 0.04 to 2.42. Correlation coefficients ranged from .08 to 0.89 and 70% of the 148 coefficients were larger than 0.50. Overestimates of effect size (see table 4.2) are not depicted in figure 2, but are easily estimated. For example A SRM of 0.85 interpreted by the researcher as large effect, changed into a moderate effect according to Cohen's thresholds, due to a correlation of 0.12 between repeated measurements

Figure 4.2 *Cohen's threshold's for effect size SRM corrected for the size of the correllation coefficent between repeated measurements*



(

4.4 DISCUSSION

The values used in effect size classification for difference between means as small, medium, and large was arbitrary but seemed reasonable, Cohen stated some 30 years ago. In the debate over which standardizing unit of the difference one should take in a within- group situation, we propose that estimating the magnitude of change by using either the SD of the change score or the pooled SD is preferable to the use of the SD at baseline as proposed by Kazis et al., [12] although the SRM must be adjusted to make correct use of Cohen's thresholds when magnitude of change over time is estimated in evaluation research. These thresholds of Cohen are now being cited without distinguishing between the units by which the assessed change over time is standardized. This is surprising since there is unequivocally no doubt that his rule of thumb was derived from the pooled SD as the estimate of the common within variance. Moreover, routine action in calculating effect sizes may have led to a reduced awareness of factors originally considered only in the calculation of power and sample size. For instance, the calculation of power of the detected change or difference without using the information of r can lead to the wrong inferences. Cohen, p. 50

In evaluation research on treatment-related quality of life, researchers seem to overlook the fact that, in assessing change over time within one subject, the experimental technique of 'self-matching' reduces the proportion of the total variance due to extraneous variables not related to the treatment or intervention per se. [55]

We may conclude that the rule of thumb proposed by Cohen can induce differences in the interpretation of the size of estimated effects. At present it does not appear to us that a single set of rules that are unequivocal or normative at some level is available. We have begun to explore alternative methods in effect size estimation and have assessed the interrelation between two effect sizes as estimates of magnitude of change over time within groups. As we have demonstrated, errors can easily be made and different interpretations of the magnitude of detected change may occur. In analysing the data from our sample of published studies on change over time in health-related quality of life, we saw meaningful shifts in magnitude of detected change in relation to the size of the correlation between pre- and post-test scores. In this article we have attempted to draw the attention to the problem of over- or underestimation of effect sizes when the Standardized Response Mean is used. Studies in which the mean change over time is standardized with the SDbaseline according to Kazis et al. [12] should report the ES_P to show that the results were not dependent on the choice of deominator in the d-index formula.

Due to their increasing appearance, it is important that all aspects of estimating the

magnitude of change be inspected. One of these aspects is the consequence of the hidden role of the correlation coefficient between repeated measurements, which increases the risk of incorrect conclusions. This initial effort may provide a moderate step toward the development of a precise and useful index in quality of life assessment in clinical trials.

Acknowledgements

Appreciation is expressed to Drs. Roy Stewart for providing valuable assistance with several aspects of the analysis, and a critical review of the manuscript. Prof.dr. Wim van den Heuvel and dr. Mike de Jongste provide helpful reviews of the manuscript.

APPENDIX

Given Cohen's formula 1 for the Effect Size index for means from matched samples:

$$(A1) \quad d_z' = \frac{m_z}{\sigma_z} = SRM$$

where:

$$\sigma z = \sigma(X_{baseline} - X_{outcome}) = \sqrt{(\sigma_{xbaseline})^2 + (\sigma_{xoutcome})^2 - 2r\sigma_{xbaseline}\sigma_{xoutcome}}$$

and assumed equal variance, i.e.:

$$\sigma x^2_{baseline} = \sigma x^2_{outcome} = \sigma x^2$$

(A2) gives:

$$\sigma_z = \sigma(x_{baseline} - x_{outcome}) = \sqrt{2\sigma^2 - 2r\sigma^2} = \sigma\sqrt{2(1-r)}$$

and for the Effect Size index for means of independent samples the standardizing unit is:

$$(A3) \quad d4' = \frac{m_{xbaseline} - m_{xoutcome}}{\sigma_P} = ES_P$$

where:

$$SD_P = \sqrt{\frac{(\sigma x_{baseline})^2 + (\sigma x_{outcome})^2}{2}} \qquad \text{for : } N_{baseline} = N_{outcome}$$

Now from equation A2 and A3 we use the difference between the standardizing unit for difference in means for matched samples (SRM) being $\sigma\sqrt{2(1-r)}/\sigma = \sqrt{2(1-r)}$ as large as would be in the case of independent samples (ES_P) [26, p.48-52]. Now we can subsitute SRM into ES_P by:

(A4)

$$SRM = d'_z = \frac{mz}{\sigma} \qquad ES_P = d_4' \frac{m_{baseline} - m_{outcome}}{\sigma}$$

$$d = d'_z \times \sqrt{2} \qquad d = \frac{d_4'}{\sqrt{(1-r)}}$$

$$d = SRM \times \sqrt{2} \qquad d = \frac{ES_P}{\sqrt{(1-r)}}$$

$$SRM \times \sqrt{2} = \frac{ES_P}{\sqrt{(1-r)}} \left\{ \text{with } r = 0 : SRM \times \sqrt{2} = ES_P \right\}$$

$$SRM \times \sqrt{2} \times \sqrt{(1-r)} = ES_P \quad \text{and} \quad \left(ES_P/\sqrt{2}\right)/\sqrt{(1-r)} = SRM$$

we note that r is estimated in cases in which the standard deviation at baseline, outcome, as well as the standard deviation of the difference or change score were published:

(A5)

$$\sigma_{change} = \sqrt{(\sigma_{baseline})^2 + (\sigma_{outcome})^2 - 2r\sigma x_{baseline}\sigma x_{outcome}}$$

$$\sigma_{change}{}^2 = (\sigma_{baseline})^2 + (\sigma_{outcome})^2 - 2r\sigma x_{baseline}\sigma x_{outcome}$$

$$\sigma_{change}{}^2 - (\sigma_{baseline})^2 - (\sigma_{outcome})^2 = -2r\sigma x_{baseline}\sigma x_{outcome}$$

$$2r = \frac{(\sigma_{baseline})^2 + (\sigma_{outcome})^2 - (\sigma_{change})^2}{(\sigma x_{baseline})(\sigma x_{outcome})} \approx r = 1/2 \frac{(SD_1)^2 + (SD_2)^2 - (SDchange)^2}{(SD_1)(SD_2)}$$

REFERENCES

1. Leon AC, Shear K, Portera L, Klerman GL. Effect Size as a Measure of Symptom-Specific Drug Change in Clinical Trials. Psychopharmacology Bulletin 1993;29(2):163-7.
2. Pulver AE, Bartko JJ, McGrath JA. The Power of Analysis: Statistical Perspectives. Part 1. Psychiatry Research 1988;23:295-9.
3. Brewer JK. Effect Size: The most troublesome of the hypothesis testing considerations. CEDR Quarterly 1978;11(4):7-10.
4. Rosnow RL, Rosenthal R. Statistical procedures and the justification of knowledge in psychological science. American Psychologist 1989;44:1276-84.
5. Rosenthal R. Progress in clinical psychology: Is there any? Clinical Psychology: Science and Practice 1995;2:133-50.
6. Rosenthal R, Rubin DB. The Counternull value of an effect size: a new statistic. Psychological Science 1994;5(6):329-34.
7. Bartko JJ, Pulver AE, Carpenter WT. The Power of Analysis: Statistical Perspectives. Part 2. Psychiatry Research 1988;23:301-9.
8. Borenstein M. A Note on the use of confidence intervals in psychiatric research. Psychopharmacology Bulletin 1994;30(2):235-8.
9. Cooper HM. On the significance of effects and the effects of significance. Journal of Personality and Social Psychology 1981;41(5):1013-8.
10. Tuley MR, Mulrow CD, McMahan CA. Estimating and testing an index of responsiveness and the relationship of the index to power. J.Clinical Epidemiology 1991;44(4/5):417-21.
11. Thompson B. Editorial policies regarding statistical significance tests: Futher comments. Educ.Res. 1997;26(5):29-32.
12. Kazis LE, Anderson JJ, Meenan RF. Effect Sizes for Interpreting Changes in Health Status. Medical Care 1989;27(3,Supplement):S178-S189
13. Guyatt GH, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. Journal Chron.Dis. 1987;40(2):171-8.
14. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Control.Clin.Trials. 1991;12(4 Suppl):142S-58S.
15. Pfennings LEMA, van der Ploeg HM, Cohen L, Polman CH. A comparison of responsiveness indices in multiple sclerosis patients. Qual.Life Res. 1999;8:481-9.
16. Wright JG and Young NL. A comparison of different indices of responsiveness. J Clin Epidemiol 1997;50(3):239-46.
17. Hillers ThK, Guyatt GH, Oldridge N, Crowe J, Willan A, Griffith L, Feeny D. Quality of life after myocardial infarction. Journal of Clinical Epidemiology 1994;47(11):1287-96.

18. de Beurs E, van Balkom AJLM, Lange A, Koele P, van Dyck R. Treatment of Panic Disorder With Agoraphobia: Comparison of Fluvoxamine, Placebo, and Psychological Panic Management Combined With Exposure and of Exposure in Vivo Alone. American Journal of Psychiatry 1995;152(5):683-91.
19. Taylor S, Woody S, McLean PD, Koch WJ. Sensitivity of outcome measures for treatments of generalized social phobia. Assessment 1997;4(2):181-91.
20. Wiebe S, Rose K, Derry P, McLachlan R. Outcome assessment in epilepsy: comparative responsiveness of quality of life and psychosocial instruments. Epilepsia 1997;38(4):430-8.
21. Beurskens AJHM, de Vet HCW, Koke AJA. Responsiveness of functional status in low back pain: a comparison of different instruments. Pain 1996;65:71-6.
22. van Bennekom CAM, Jelles F, Lankhorst GJ, Bouter LM. Responsiveness of the Rehabilitation Activities Profile and the Barthel Index. Journal of Cliical Epidemiology 1996;49(1):39-44.
23. Lachs MS. The more things change... Journal of Clinical Epidemiology 1993;46(10):1091-2.
24. Kempen GIJM, Miedema I, van den Bos GAM, Ormel J. Relationship of domain-specific measures of health to perceived overall health among older subjects. J Clin Epidemiol 1998;51(1):11-8.
25. Kraemer HC. Reporting the size of effects in research studies to facilitate assessment of practical or clinical significance. Psychoneuroendocrinology 1992;17(6):527-36.
26. Cohen J. Statistical power analysis for the behavioural sciences. revised edition ed. New York: Academic Press; 1977.
27. Lipsey MW. Design sensitivity. Statistical power for experimental research. SAGE Publications,London.; 1990.
28. Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. Medical Care 1990;28(7):632-42.
29. Diehr P, Psaty BM, Patrick DL. Effect size and power for clinical trials that measure years of healthy life. Stat.Med. 1997;16(11):1211-23.
30. Katz JN, Larson MG, Phillips CB, Fossel AH, Liang MH. Comparative measurement sensitivity of short and longer health status instruments. Medical Care 1992;30(10):917-25.
31. Garratt AM, Ruta DA, Abdalla MI, Russell T. Responsiveness of the SF-36 and a condition-specific measure of health for patients with varicose veins. Quality of Life Research 1996;(5):223-34.
32. Jacobs HM, Touw-Otten FWMM, de Melker RA. The Evaluation of Changes in Functional Health Status in Patients with Abdominal Complaints. Journal of Clinical Epidemiology 1996;49(2):163-71.
33. Doeglas D, Krol B, Guillemin F, Suurmeijer Th, Sanderman R, Smedstad LM, van den Heuvel WJA. The Asessment of Functional Status in Rheumatoid Arthritis: A Cross Cultural, Longitudinal Comparison of the Health Assessment Questionnaire and the Groningen Activity Restriction Scale. The Journal of Rheumatology 1995;22(10):1834-43.

34. Bouchet C, Guillemin F, Briancon S. [Comparison of 3 quality of life instruments in the longitudinal study of rheumatoid arthritis] Comparaison de trois instruments de qualité de vie pour l'etude longitudinale de la polyarthrite rheumatoide. Rev.Epidemiol.Sante.Publique. 1995;43(3):250-8.
35. Koes BW. Efficacy of manual therapy and physiotherapy for back and neck complaints. (dissertation). Maastricht: University of Limburg; 1992.
36. Vliet-Vlieland ThPM, Zwinderman AH, Breedveld FC, Hazes JMW. Measurement of morning stiffness in rheumatoid arthritis clinical trials. J Clin Epidemiol 1997;50(7):757-63.
37. Husted JA, Cook RJ, Farewell VTGDD. Methods for assessing responsiveness: a critical review and recommendations. J Clin Epidemiol 2000;53:459-68.
38. Middel B, Kuipers-Upmeijer H, Bouma J, Staal MJ, Oenema D, Postma Th, Terpstra S, Stewart R. Effect of intrathecal baclofen delivered by an implanted programmable pump on health related quality of life in patients with severe spasticity. J Neurol Neurosurg Psychiatry 1997;63:204-9.
39. Gordon JE, Powell C, Rockwood K. Goal attainment scaling as a measure of clinically important change in nursing-home patients. Age and Ageing 1999;28:275-81.
40. Wells G, Boers M, Shea B, Tugwell P, Westhovens R, Saurez-Almazor M, Buchbinder R. Sensitivity to change of generic quality of life instruments in patients with rheumatoid arthritis: preliminary findings in the generic health OMERACT Study. The Journal of Rheumatology 1999;26(1):217-21.
41. O'Carrol RE, Cossar JA, Couston MC, Hayes PC. Sensitivity to change following liver transplantation. A comparison of three instruments that measure quality of life. Journal of Health Psychology 2000;5(1):69-74.
42. Macduff C, Russell E. The problem of measuring change in individual health-related quality of life by postal questionnaire: use of the patient-generated index in a disabled population. Qual.Life Res. 1998;7:761-9.
43. Brunner HI, Feldman BM, Bombardier C, Silverman ED. Sensitivity of the systemic lupus erythematosus disease activity index, british isles lupus assessment group index,and systemic lupus activity measure in the evaluation of clinical change in childhood-onset systemic lupus erythematosus. Arthritis and Rheumatism 1999;42(7):1354-60.
44. Sneeuw KCA, Aaronson NK, Sprangers MAG, Detmar SB, Wever LDV, Schornagel JH. Comparison of patient and proxy EORTC QLQ-C30 ratings in assessing the quality of life of cancer patients. J Clin Epidemiol 1998;51(7):617-31.
45. Vulink NCC, Overgaauw DM, Jessurun GA, Ten Vaarwerk IAM, Kropman TJB, Van der Schans CP, Middel B, Staal MJ, De Jongste MJL. The effects of spinal cord stimulation on quality of life in patients with therapeutically chronic refractory angina pectoris. Neuromodulation 1999;2:33-40.
46. Gliklich RE, Hilinsky JM. Longitudinal sensitivity of generic and specific health measures in chronic sinusitis. Quality of Life Research 1995;4:27-32.
47. Sneeuw KCA, Aaronson NK, Osoba D, Muller MJ, Hsu M-A, Yung AWK, Brada M, Newlands ES. The Use of Significant Others as Proxy Raters of the Quality of Life of Patients with Brain Cancer. Medical Care 1997;35(5):490-506.

48. Vaile JH, Mathers M, Ramos-Remus C, Russel AS. Generic health instruments do not comprehensively capture patient perceived improvements in patients with carpal tunnel syndrome. The Journal of Rheumatology 1999;26(5):1163-6.
49. Bruin AFd, 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(5):529-40.
50. De Witte LP. After the rehabilitation centre; a study into the course of functioning after discharge from rehabilitation. (dissertation). Amsterdam/Lisse: Zwets en Zeitlinger; 1992.
51. Janssen M. Personal Networks of chronic patients. (dissertation). Maastricht: University of Limburg; 1997.
52. Hidding A, Van der Linden Sj, Boers M, Gielen X, Kester A, De Witte LP, Dijkmans B, Moonenburgh J. Is group physical therapy superior to individual therapy in ankylosing spondylitis, A randomized controlled trial. Arthritis Care and Research 1993;6(3):117-25.
53. Courtens AM. Kenmerken van zorg en kwaliteit van leven bij patienten met kanker (Characteristics of Care and Quality of Life in cancer patients) (dissertation). Maastricht: University of Limburg; 1993.
54. Cohen J. A Power Primer. Psychological Bulletin 1992;112(1):155-9.
55. Winer BJ. Statistical principles in experimental design. second ed. Tokyo: McGraw-Hill Kogakusha; 1962.