ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

European Journal of Pain 11 (2007) 635–643

www.EuropeanJournalPain.com

Long-term reliable change of pain scores in individual myogenous TMD patients

Robert J. van Grootel, Andries van der Bilt, Hilbert W. van der Glas *

Department of Oral-Maxillofacial Surgery, Prosthodontics and Special Dental Care, University Medical Center Utrecht, STR 4.115, P.O. Box 85060, 3508 AB Utrecht, The Netherlands

Received 26 October 2005; received in revised form 3 October 2006; accepted 4 October 2006
Available online 22 November 2006

Abstract

A within-patient change in pain score after treatment is statistically ‘reliable’ when it exceeds the smallest detectable difference (SDD). The aims of the present study were to: (i) determine SDDs for VAS-scores of pain intensity, for sufficiently long test–retest intervals to include most biological fluctuations, (ii) examine whether SDD is invariant to baseline score, and (iii) discuss the value of reliable change (RC) for detecting clinically important difference (CID) or as a possible indicator of successful treatment. SDDs were determined using duplicate data from 118 patients with myogenous Temporomandibular disorders: (1) VAS-scores of pain intensity from the masticatory system in a pre-treatment diary, and (2) VAS-scores of pain intensity from the hand (cold-pressor test). RC was determined in VAS-scores from a pre- and post-treatment questionnaire. The long-term SDD was 49 mm. A regression analysis on duplicate VAS-scores showed that SDD was largely invariant to the baseline level. Because RC (change > SDD) exceeded CID, it might serve as an indicator of successful treatment. However, only 17% of the patients showed RC after treatment, mainly because the baseline was smaller than SDD in 67% of the patients thus making detection of any treatment effect impossible. For patients with possible detection (33%), the frequency of RC was 51%. If the detection threshold would be avoided by provoking pain in patients with a low baseline, a long-term RC in VAS-scores might occur in about half of all myogenous TMD patients and might then serve as an indicator of cases of treatment success.

Keywords: Decision; Pain rating; Reliability; Visual analogue scales; Temporomandibular disorders

1. Introduction

Temporomandibular disorders (TMD) are characterized by pain and restricted jaw movements. Scoring of, for example, pain intensity on a visual analogue scale (VAS) has been used to evaluate treatment effects (Dao et al., 1994; Ekberg et al., 1998; for a review, see Jensen, 2003).

A group analysis of treatment efficacy is based on the difference in pre- and post-treatment score averaged across patients. However, a mean difference value only provides information on the average treatment effect and not about the effect on individuals. Although a small value of the mean difference might become statistically significant for a large sample size, this small value might not become clinically significant or relevant because the score change in many individuals might not exceed the smallest difference detected above a statistical level of random fluctuations (Jacobson et al., 1984).

A clinically relevant treatment effect within a patient is characterized by: (i) a change in score that is statisti-

* Corresponding author. Tel.: +31 30 2533097/3610; fax: +31 30 2535537.
E-mail address: h.w.vanderglas@med.uu.nl (H.W. van der Glas).

1090-3801/$32 © 2006 European Federation of Chapters of the International Association for the Study of Pain. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.ejpain.2006.10.003

cally ‘reliable’ and (ii) a change that a patient considers as being beneficial (Jacobson et al., 1984; Kropmans et al., 1999a). A reliable change between two measurements exceeds the extent of random error that is present within a measurement system (Streiner and Norman, 2004). This random error, denoted as the smallest detectable difference (SDD; Kropmans et al., 1999a), is estimated from a variance value to which various factors contribute that influence the test–retest reproducibility of the measurement procedure.

Variations in repeated scores of pain intensity that are not due to treatment will mainly depend on: (1) the patient’s accuracy of handling a scale, (2) time variations in nociceptive mechanisms or mood, and (3) a patient’s memory for scoring. The contribution of the last two factors will depend on the test–retest interval. This interval for examining VAS-scores of pain intensity varied between up to four hours (Kropmans et al., 1999a) to a week (Kropmans et al., 2002). However, the usual interval between successive treatment evaluations of TMD is several weeks, varying between 6 and 10 weeks in efficacy studies (Dao et al., 1994; Rudy et al., 1995). The duration of an entire TMD treatment includes even several months. It is therefore relevant to determine SDDs for long time-intervals to assess whether treatment causes improvement in a patient that exceeds the effect of random fluctuations or the natural course of a disorder.

From studies linked to a randomized clinical trial, duplicate data on VAS scores of pain intensity were obtained that will serve three aims of the present study, of which at least its methodology has the potential for application to pain patients in general. The first aim was to determine long-term SDDs. The concept of SDD implicitly assumes that the variation of a score would be invariant to the baseline score level. Therefore, the second aim was to examine whether the same SDD is valid regardless of the patient’s baseline. The third aim was to deal with the problem of SDD being a threshold for enabling detection of reliable change, and the relationship between SDD and clinically important difference.

2. Materials and methods

2.1. Patients

The studies from which the data originated were approved by the Ethics Committee of the University Medical Center Utrecht. The patients with myogenous TMD ($n = 118$) met the following inclusion and exclusion criteria (for details, see van Grootel et al., 2005): (i) pain and tenderness of the muscles of mastication and restricted mandibular opening of 3 month duration or longer, (ii) no clinical and/or radiographic evidence of organic TMJ changes, (iii) no previous TMD treatment or recent (<1 year) other pain treatment, (iv) no evidence of serious psychopathology (no psychotherapy and/or psychomedication, no recent dramatic life events), and (v) between 18 and 65 years of age. The mean age of the patients was 31.6 years (SD 10.0); 93% were female. The median duration of pre-treatment pain was 1.1 years (range 3 months to 20 years). Our sample is a representative for myogenous TMD patients (in particular females) without any or a recent previous treatment of pain (van Grootel et al., 2005).

2.2. Procedures and data analysis

The patients who gave informed consent started with treatment 2–3 weeks after intake. The first aim was to determine long-term SDDs using sufficiently long test–retest intervals to include most fluctuations of biological origin in pain sensation. To that end, two sets of duplicate data were available related to two types of pain, i.e. TMD pain of the masticatory system (with relatively low score levels) and experimentally induced pain from the hand (with relatively high score levels; Table 1). Regarding the first data set, the patients were handed a paper-and-pencil diary to score pain variables for a period of 14 days prior to the start of treatment. While details can be found elsewhere (van Grootel et al., 2005), only an outline is given here. For each day, four different intervals were considered for scoring pain variables at the end of each interval. Being representative for other daily intervals, the results of the analysis will be presented for the interval between dinner and bedtime for which scoring was most complete (95–106 out of 118 cases). Scores of intensity of pain from the masticatory system were studied using a VAS of 100 mm, with ‘no pain’ and ‘the most pain one can imagine’ as anchor points. In order to determine variations of the VAS-scores, duplicate scores of pain intensity were examined at different inter-day intervals in the diary. Chosen around the center of the diary, the diary interval concerned the 7th and 8th diary day, the 4th and 11th day, and the 1st and 14th day (1, 7, and 13 day interval, respectively). The 7 day interval was directly comparable with one from the literature (cf. Section 4).

The second data set regarding SDD of VAS-scores of pain intensity, originated from 109 out of 118 patients who participated in a cold-pressor test (Chen et al., 1989; Table 1). This test is characterized by an experimentally induced constant noxious stimulus. To that end, the patients held the non-writing hand up to the wrist in 2 °C water as long as they could bear it or for a limit of 4 min. This water was contained in the central chamber of a two-chamber bath, while the outer chamber contained a slurry of ice and water. While immersing the hand, the water was gently agitated using a magnetic stirrer in a bottom compartment of the central chamber. The patient scored the maximal intensity of pain from the hand on a VAS of 100 mm. These scores were

Table 1
Relationship between aims of the study and used measures

Data set	Aim of the study		
	SDD	Relationship between variation of score and baseline score level	RC and CID
Pre-treatment diary (masticatory system; duplicate VAS-scores of pain intensity;predominantly small values)	+	+	
Cold-pressor test (hand; duplicate VAS-scores of pain intensity; predominantly large values)	+	+	
Questionnaire (masticatory system; pre- and post-treatment VAS-scores of pain intensity for area with predominant pain)			+

SDD, smallest detectable difference; RC, reliable change; and CID, clinically important difference.

obtained before treatment and during the last visit, between 10 a.m. and 3 p.m. The last visit occurred immediately after treatment for patients whose treatment was unsuccessful in the short term, or after a follow-up of 6 or 12 months. In this way, duplicate VAS-scores were obtained with an inter-score interval between 2 and 18 months. As a control for excluding any influence of long-term variations of the mechanisms underlying myogenous TMD, the cold-pressor test was also applied to 24 healthy subjects, matched for age and gender and using an interval of 6–12 months for retesting.

In order to determine the smallest detectable difference (SDD) in VAS-scores of pain intensity, the difference between duplicate VAS-scores was calculated for each subject. Since these difference values were nearly normally distributed (cf. Section 3), the SDD was well approached by 1.96× SDd, in which SDd is the standard deviation of the difference values. SDD is thus related to the limits of a 95% confidence interval for difference values. As the mean of the difference values was close to zero (no systematic differences between duplicate values), this definition of SDD is mathematically equivalent to the one used previously (Cronbach et al., 1972; Kropmans et al., 1999a; see Appendix).

The second aim of the study was to examine whether the variation of VAS-scores of pain intensity was similar regardless of the patient's baseline level. To that end, regression functions were determined between duplicate VAS-scores from the various patients, using two data sets: diary and cold-pressor test (Table 1). Fig. 1A shows an example of such a regression function for diary data. Furthermore, the residuals (the deviation in the second score from the predicted value according to the regression function for various values of the first score) have been determined as a function of the predicted score. The residuals have been normalized with respect to their standard deviation (Norušis, 1999). This normalization included a correction for variations in local variance, which tends to be larger the more the predicted score values are located near an end of the range of the regression functions (studentized residuals, Fig. 1B). If the variation in the second VAS-score were invariant to the value of the first score, the studentized residuals would randomly scatter around the zero-line in a residual–predicted value plot, within a band of constant width. In contrast, if, for example, the variation of the second score increased proportionally with the value of the first score, the range of scatter of the studentized residuals around the zero-line would increase linearly with the value of the predicted score. The band of scatter would then have a funnel shape around the zero line.

The third aim was dealing with the problem of SDD being a threshold for enabling detection of reliable change (RC) in the score value after treatment, and the relationship between SDD and clinically important difference. Detection of treatment effect by the occurrence of RC will only be possible for patients whose baseline score exceeds SDD. Regarding this aim, the patients scored the pain intensity of the area in the masticatory system for which pain was predominant, using, apart from the diary, a 100 mm VAS in a questionnaire (Table 1). Questionnaire's scores were considered which were obtained just before treatment was started and at the last visit (after treatment or follow-up, inter-score interval between 2 and 18 months). The frequency of RC was determined using the long-term SDD-value from the present study as a criterion (49 mm, cf. Section 3, diary) in two groups of patients, i.e. the entire patient group and in the subgroup of patients whose baseline score exceeded SDD. The frequency of RC in the subgroup gives a clue of the frequency of RC that might occur in the entire group of patients when the baseline of all patients would be located above SDD by applying a procedure of provoking pain when necessary (cf. Section 4).

The average effect of the various types of treatment procedures from a randomized clinical trial occurred in the beforementioned determination of frequency of RC. Treatment types involved were three conventional dental therapies, i.e. occlusal splint, occlusal adjustment or a combination of both, and physiotherapy of the masticatory system (details unnecessary for the present

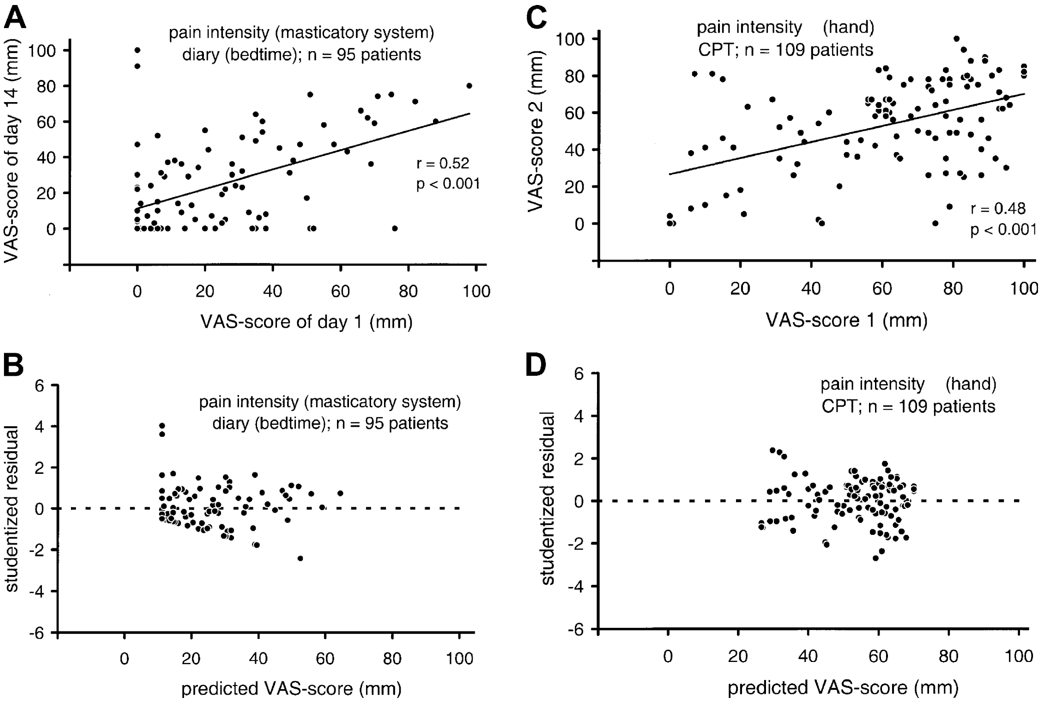

Fig. 1. Regression analysis of the relationship between duplicate VAS-scores of pain intensity. A and B, data from the diary on pain from the masticatory system, with diary days 1 and 14 as a representative example for two times of measurement. C and D, data from the cold-pressor test on pain from the hand with an interval of 2–18 months between VAS-scores 1 and 2. The equations of the regression lines in A and C (solid lines) are $Y = 0.543X + 11.2$ and $Y = 0.435X + 26.6$, respectively. Studentized residual, deviation (in the Y-direction) of the second score-value from the regression line, normalized with respect to the SD of the residuals and corrected for variations in local variance. Predicted VAS-score, score value according to the regression function for a given value of the VAS-score of time 1. Studentized residuals have been depicted as a function of predicted score-value (B and D) for assessing whether the amount of scatter is invariant to the predicted value (dashed line, zero level).

study). Pooling of treatments could be applied as no significant differences between treatment types occurred in success rate (unpublished observations) and the mean success rate (74% in the short-term) was similar to that (75–80%) reported in TMD textbooks (Clark, 1988; Greenwood, 1994; Okeson, 1996).

Clinically important difference (CID) was assessed using the criterion of a reduction of 30% in a score of pain intensity. This criterion has been derived from the relationship between percentage change in a score of pain intensity and the patient's assessesment of change, which is similar for various types of chronic pain patients regardless of their baseline level (Farrar et al., 2001). Furthermore, using data from the present study, the change in mean VAS-score from the questionnaire observed after interventions of similar and known efficacy (see above) was considered as an initial estimate of CID (Guyatt et al., 1987).

2.3. Statistical analysis

SPSS 9.0® was used to examine the normality of the distribution of VAS-scores and that of difference values between duplicate scores (Kolmogorov–Smirnov one-sample test), and for a regression analysis including residuals. Furthermore, the significance of differences between mean level of scores of pain intensity from the diary, the cold-pressor test, and the questionnaire was examined in the patients using a Student's t-test for paired observations. The significance of differences in SDD-values was determined by applying the F-test to values of the standard deviation of difference values of duplicate data (SDd) for two cases of unpaired observations, i.e. diary data from two patient subgroups ('short' and 'long' pre-treatment duration of pain), and cold-pressor test data from the patients and healthy subjects, respectively. Interclass correlation coefficients (ICCs, random, 2-way) between duplicate scores were determined as measure of test–retest reliability.

3. Results

3.1. Descriptive statistics, distribution of data samples, and SDD of VAS-scores of pain intensity

Table 2 shows data on VAS-scores of pain intensity from the masticatory system (diary data) and on scores of experimentally induced pain from the hand (cold-pressor test; CPT). The pain intensity of the masticatory system was on average similar for the various days (diary: columns 'VAS_1' and 'VAS_2'). The SD-values of

Table 2
Duplicate VAS-scores of pain intensity and their smallest detectable difference (SDD)

	n	VAS_1 (mm)	SD_1 (mm)	VAS_2 (mm)	SD_2 (mm)	ICC	Δ (mm)	SDd (mm)	SDD (1.96 × SDd) (mm)
Diary, patients									
Day 7–8	106	25.0	25.6	23.2	24.1	0.70	−1.8	19.2	37.6
Day 4–11	106	24.8	25.2	24.1	26.0	0.63	−0.7	22.0	43.1
Day 1–14	95	24.3	24.7	24.5	25.9	0.52	0.1	24.9	48.8
CPT patients	109	60.9	27.6	53.1	25.0	0.48	−7.8	26.9	52.7
CPT controls	24	53.3	23.8	54.8	25.2	0.60	1.5	21.9	42.9

Diary, day: day numbers of scoring of intensity of pain of the masticatory system between dinner and bedtime, in a pre-treatment pain diary of 14 consecutive days; CPT: cold-pressor test applied to the hand with scoring of maximal pain intensity of the hand; n: number of subjects; VAS_1 and VAS_2: VAS-score of the first and the second time, respectively; SD_1 and SD_2: standard deviation of these VAS-scores; ICC: interclass correlation-coefficient (random, two way); Δ: mean of the difference between duplicate data ($VAS_2 - VAS_1$); SDd: standard deviation of the difference values; and SDD: smallest detectable difference.

these scores were also similar (Table 2, diary: columns 'SD_1' and 'SD_2').

Before calculating SDD values, the normality of the distribution of the data samples was examined. The distributions of VAS-scores of pain intensity across patients were skewed and non-normal for all diary days ($p < 0.001$; Kolmogorov–Smirnov one-sample test; $n = 95$–106). The distributions of VAS-scores from the cold-pressor-test were also non-normal for the patients ($p < 0.05$–0.001; $n = 109$). Normality could just not be rejected for the distribution of the second VAS-scores from the cold-pressor test of the control subjects ($p = 0.051$; $n = 24$). Regardless of the origin of the duplicate VAS-scores (diary or cold-pressor test), the difference values of duplicate scores were non-skewed and nearly normally distributed (no significance in a Kolmogorov–Smirnov one-sample test for CPT values; $p > 0.06$). With a high degree of statistical power because of large subject samples ($n = 95$–106), normality was rejected for the distributions of difference values from the diary ($p < 0.001$). This rejection was mainly due to the incidence of a few outlying values in the tails of the sample distribution. However, these sample distributions were still nearly normal as a 95% confidence interval defined by the mean ±1.96 × SD including 93–95% of the values.

The mean difference in duplicate VAS-scores was close to zero (Table 2, column 'Δ'). The SD-values of the difference scores increased with the length of the interval between duplicate scores (Table 2, column SDd). Because of a nearly normal distribution of difference values (see above), the smallest detectable difference (SDD) was straightforwardly related to SDd (SDD = 1.96 × SDd). Therefore, the SDD also increased with the interval length and was 49 mm for the longest diary interval available (13 days). The SDD values were similar between two subgroups of patients (no significance in an F-test), i.e. for patients whose duration of pre-treatment pain was shorter than the median value (1.1 years) and for patients whose duration was equal to or larger than the median.

The patients' VAS-score of pain intensity was significantly larger ($p < 0.001$, Student's t-test for paired observations) for pain from the hand than for pain in the masticatory system. The mean pain intensity was (for example, for the first measurement in patients) 60.9 mm for the hand and (for example, for diary day 1–14) 24.3 mm for the masticatory system (Table 2). In contrast to the mean values, the patients' SD-values were similar between the cold-pressor test and diary, and between the patients and the healthy controls for the cold-pressor test. Furthermore, the SD-values for the difference between duplicate data were similar for these subject groups (Table 2, CPT, column 'SDd'; no significance in an F-test, $p > 0.1$). The long-term SDD from the cold-pressor-test (interval 2–18 months; 43–53 mm) was similar to the SDD from the diary for the longest interval (13 days; 49 mm).

3.2. Variation of a second VAS-score of pain intensity as a function of the level of the first score

Fig. 1A shows a representative example of a scatter plot including the regression function of duplicate VAS-scores of pain intensity of the masticatory system (diary data; predominantly small values). Fig. 1B shows the normalized residuals (the deviations in the second scores) as a function of the predicted VAS-score according to the regression line in Fig. 1A. Figs. 1C and D show the same for duplicate VAS-scores of pain intensity of the hand during the cold-pressor test (predominantly large values). The amount of scatter of the residuals did not increase proportionally with an increase in predicted VAS-score. Thus the band of scatter did not have a funnel shape around the zero line in Figs. 1B and D. In contrast, the residuals scattered largely in a random matter around this zero line, and the band of scatter had approximately a constant width regardless of the value of the predicted VAS-value. Only for extremely small values of the predicted VAS-value (about 15 mm, diary data, Fig. 1B), the values of the positive residuals tended to have a larger range than

Table 3
Incidence of reliable change (RC) in VAS-score of pain intensity after treatment and follow-up of myogenous TMD patients and incidence of relatively small and large initial score

SDD	Incidence of RC (criterion: score decreas > SDD) in entire patient sample	Incidence of pre-Tx VAS-score ⩽ SDD	Incidence of RC in patients whose pre-Tx VAS-score > SDD
49 mm	20/118 (16.9%)	79/118 (66.9%)	20/39 (51.3%)

SDD, smallest detectable difference between scores of pain intensity in the long-term (diary data) and incidence of pre-Tx VAS-score ⩽ SDD, number of patients whose pre-treatment VAS-score was smaller than or equal to SDD so that detection of treatment effect was inherently impossible in these cases, using SDD as criterion for reliable change.

those of the negative values. However, in all cases (diary as well as cold-pressor test) the degree of correlation was extremely low ($r < 0.0015$) between the values of the studentized residuals and the predicted VAS-scores (Figs. 1B and D), and the regression line describing this relationship nearly coincided with the zero line.

3.3. An initial estimate of CID, and the ability of detecting reliable change of VAS-scores of pain intensity for myogenous TMD

The questionnaire's VAS-score of pain intensity of the masticatory system was on average 40.0 mm (SD 22.3, $n = 118$) just before treatment was started and 15.8 mm (SD 22.0) at the last visit. The significant ($p < 0.001$; Student's *t*-test for paired observations) decrease in mean VAS-score reflects, at least in part, the mean effect of the various types of treatments used. This decrease (24.2 mm) is an initial estimate of clinically important difference (CID; cf. Section 2).

Using the long-term SDD-values of 49 mm (diary data), a reliable decrease had occurred in the questionnaire's VAS-score for 16.9% of all patients (Table 3). The proportion of patients in which reliable decrease can be detected by using a particular SDD-value depends on the number of patients in which detection of treatment effect is possible anyhow, i.e. only for patients whose baseline score exceeds SDD. Detection of treatment effect was inherently not possible in 66.9% of the patients using an SDD criterion of 49 mm. With respect to the number of patients in which detection of treatment effect could occur, the rate of detection of reliable decrease was 51.3% (Table 3).

4. Discussion

4.1. SDD of VAS-scores of pain intensity

Recently, a numerical rating scale (NRS) has been recommended for scoring pain intensity, for avoiding difficulties of completing VAS-scores by elderly patients or when opioid intake is involved (Dworkin et al., 2005). Such difficulties did not occur with myogenous TMD patients, and the consequences of our findings will likely also apply to NRS-scores, the more as both types of scores are greatly correlated in the same patients (for a review, see Jensen, 2003).

SDD observed in myogenous TMD patients for the greatest diary interval of 13 days (49 mm) is larger than the one reported for VAS-scores of actual, minimal or maximal pain in patients scheduled for general physiotherapy (28, 22 and 22 mm, respectively; Kropmans et al., 1999a), and for average, minimal and maximal pain in arthrogenous TMD patients (35, 25 and 43 mm, respectively; Kropmans et al., 2002). A larger SDD-value is most likely due to a larger test–retest interval. Patient type or the use of a diary rather than a questionnaire is a less likely explanatory factor because the 1-week SDD from the diary (42 mm) is similar to the 1-week SDD of questionnaire's scores for average or maximal pain, respectively, in arthrogenous TMD patients (35 and 43 mm, respectively; Kropmans et al., 2002). Furthermore, the long-term SDD-value from the diary is similar to the long-term SDD-value from the cold-pressor-test (see below). A few patients in our sample showed an extreme long-term variation of about 90 mm in the duplicate VAS-scores (Fig. 1). Such patients also occurred in a previous longitudinal study on TMD pain (see Fig. 2 in Raphael and Marbach (1992)). Because of a low incidence (2–3%), our results would not alter essentially after omitting such scores.

The duration of the pre-treatment diary has been limited to two weeks for ethical reasons and for avoiding influencing the patients' use of over-the-counter medication. For two reasons, a 2-week baseline in a diary will most likely be sufficient to reveal nearly the complete variation in pain scores. First, in a 4-week diary of patients suffering from tension headache, the degree of correlation between average pain scores from baselines of 1, 2, or 3 weeks with scores from the entire 4-week period closely approached the maximum when the baseline was at least 2 weeks (Blanchard et al., 1984). As pain is more sustained for myogenous TMD than for tension headache (van Grootel et al., 2005), a 2-week baseline will also be sufficient for TMD.

Second, the 2-week SDD for pain from the masticatory system (49 mm) is similar to that for experimentally induced pain from the hand with a test–retest interval of 2–18 months (53 mm for patients; 43 mm for controls). The repeated cold-pressor test with a constant noxious stimulus shows that large variations occur in VAS-scores of pain intensity that are directly due to fluctuations in nociceptive mechanisms (anyhow unaffected by TMD in the healthy control subjects), or in an indirect manner to variations in the subject's mood (Vendrig and Lousberg, 1997). One might argue that the magnitude of variations in pain scores might be influenced by differences in the quality of sensation between TMD pain and the pain induced during the cold-pressor test. However, such an influence is unlikely as the sensory and affective qualities of pain experience are separately mentally encoded and/or retrieved (Morley, 1993).

4.2. Invariance of SDD to the baseline level of VAS-scores of pain intensity

Similar SDD-values occurred while pain scores are larger for the hand (mean 61 mm) than for the masticatory system (22–25 mm in the diary). The conclusion of invariance of SDD to the score level is reinforced by the finding that in the relationship between studentized residual and predicted VAS-score, the residual values scatter almost randomly around the zero line within a band of constant width (Fig. 1). Only for extremely small values of the predicted VAS-value (about 15 mm, diary data), the values of the positive residuals tended to have a larger range than those of negative residuals. This asymmetry is due to a floor effect on VAS-scores of which the underbound is 0 mm. As the degree of correlation between the studentized residuals and the predicted VAS-scores is nearly zero, this extreme case does not distort the overall finding of random scattering.

4.3. Relationship between SDD and CID, and SDD being a threshold for detecting RC

In sight of within-patient reliable change that is also clinically relevant, three conditions are important, i.e. (i) a change in value should exceed the measurement error of the instrument (not related to biological factors of the patient), (ii) a long-term change should occur that exceeds fluctuations of biological origin within the interval of repeated clinical examinations, and (iii) the change should exceed one that a patient would consider as beneficial.

If the instrumental error as well as the biological variations were small, the frequency of RC might exceed that of cases with a clinically relevant change. Priority might then be given to a relevant change rather than one that is solely based on statistical criteria. However, the long-term SDD of 49 mm is so large that a change that exceeds this SDD will likely also be relevant. A reduction of 30% in a score of pain intensity represents a clinically important difference (CID), regardless of the baseline of various types of chronic pain patients (Farrar et al., 2001). Therefore, a decrease of 49 mm in VAS-score for a patient with a baseline of 49.1 mm (the minimal baseline level required for possibly detecting a reliable change of 49 mm) will exceed the CID of 15 mm belonging to this baseline (30% of 49 mm). Even if the baseline were maximally large (100 mm), a change of at least 49 mm would be larger than CID for this baseline (30 mm). The SDD of 49 mm is also large with respect to the difference in mean pre- and post-treatment VAS-scores (24 mm) as an initial estimate of CID according to questionnaire's data from the present study. Thus, both types of estimates of CID suggest that VAS-scores of pain intensity are not sufficiently accurate for detecting changes corresponding with CID, indicating a low responsiveness of the instrument (Guyatt et al., 1987). On the other hand, as reliable change in VAS-scores is related to clinically a very important difference anyhow it might be a simple indicator for cases of treatment that are considered as being successful according to various anamnestic and clinical criteria. However, RC in the questionnaire's scores has been detected in only 17% of the myogenous TMD patients. Apart from the large decrease required for attaining RC, this detection rate is low because any treatment effect was impossible to detect in those patients (67%) whose baseline score did not exceed SDD.

The ability of detecting RC will be increased by decreasing the variability of a score by averaging across repeated measurements (Kropmans et al., 2002), or by summation or averaging across scores from a multidimensional scale (Gracely and Kwilosz, 1988; Kropmans et al., 1999b). However, SDD still forms a considerable threshold for detecting RC when the multi-item sickness impact profile (SIP) is applied to stroke patients (Beckerman et al., 2001). Apart from an assessment burden on the patient, and considerations of cost-benefit in clinical practice, the gain of averaging repeated measurements is limited. For example, even if the long-term SDD were decreased by 37% (from 49 to 31 mm) by repeating scoring four times (Kropmans et al., 2002), there is still a detection threshold of 31 mm for the averaged score. Furthermore, the detection of RC cannot be improved by considering a relative (percentage) change rather than an absolute change in VAS-score because the variation does not increase proportionally with the baseline.

However, the threshold of detecting RC in scores of pain intensity at rest might be avoided by applying a procedure of provoking pain, only to those patients whose baseline score does not exceed SDD. For example, some quantified pressure might be applied on a tender jaw muscle. The amount of pressure should be sufficiently large to raise the patient's VAS-score above the SDD level, but also sufficiently small to avoid provoking pain in healthy subjects. When the measurement is repeated during another visit, the same patient-specific pressure and site should be used. Our findings of: (1) an invariance of SDD to baseline pain, and (2) a frequency of 51% of reliable decrease for those patients whose baseline score exceeds SDD, collectively suggest that if the level of baseline pain were larger than SDD in all patients, by provoking more pain when necessary, a long-term reliable decrease in VAS-scores might occur in about half of the patients. This proportion of the patients is only slightly smaller than the proportion that has a successful treatment in the long-term (59%, unpublished observations). In combination of provoking pain, RC in VAS-scores of pain intensity might therefore be of interest for predicting cases of successful treatment, in particular when there would be a good agreement between both cases.

Acknowledgements

This work was supported by the Netherlands Institute for Dental Sciences. We are grateful to Dr. M. Schipper (Department of Biostatistics, University of Utrecht) for statistical advice and to Dr. R. Buchner for comments on the manuscript.

Appendix

Because only two duplicate scores were available or used for each subject and systematic differences were absent between these scores, the smallest detectable difference (SDD) in VAS-scores of pain intensity could straightforwardly be calculated according to $\mathrm{SDD} = 1.96 \times \mathrm{SDd}$, in which SDd is the standard deviation of difference values between duplicate VAS-scores. SDD is related to the limits of a 95% confidence interval for the difference values between duplicate scores. In the literature (Cronbach et al., 1972; Kropmans et al., 1999a; Streiner and Norman, 2004), another approach has been described, i.e. an observed score is considered as the sum of the unknown true score and a random error value. The standard error of a measurement procedure (SEMp) is calculated from the standard deviation (SD) of the VAS-scores from a subject sample and a reliability coefficient (Portney and Watkins, 2000; Streiner and Norman, 2003). Therefore, $\mathrm{SEMp} = \mathrm{SD} \times \sqrt{}(1 - r)$. Assuming that SEMp of the observed score of the first and the second time are equal, $\mathrm{SDD} = 1.96 \times \sqrt{2} \times \mathrm{SEMp}$.

It is easy to show that both approaches are mathematically equivalent when Pearson's correlation coefficient between duplicate data can serve as reliability coefficient (true for the present study because the mean difference is close to zero, thus no systematic differences, and the variation of the second score is invariant to the level of the first score), and the same assumptions on variance of the VAS-scores hold:

$$\sigma^2(\mathrm{VAS}_2 - \mathrm{VAS}_1) = \sigma^2(\mathrm{VAS}_1) + \sigma^2(\mathrm{VAS}_2) - 2 \times c(\mathrm{VAS}_1, \mathrm{VAS}_2)$$

(in which 'σ^2' stands for variance, 'c' for covariance and VAS_1 and VAS_2 are the VAS-scores on times 1 and 2 of measurement, respectively, thus $\mathrm{VAS}_2 - \mathrm{VAS}_1$ refers to the difference between duplicate scores)

$$= \sigma^2(\mathrm{VAS}_1) + \sigma^2(\mathrm{VAS}_2) - 2 \times r \times \sqrt{}(\sigma^2(\mathrm{VAS}_1) \times \sigma^2(\mathrm{VAS}_2))$$

(in which 'r' is Pearson's correlation coefficient) If one assumes that the variance (or standard deviation) is the same at time 1 and time 2

$$(\sigma^2(\mathrm{VAS}_1) = \sigma^2(\mathrm{VAS}_2) = \sigma^2(\mathrm{VAS})),$$

then

$$\begin{aligned}\sigma^2(\mathrm{VAS}_2 - \mathrm{VAS}_1) &= 2 \times \sigma^2(\mathrm{VAS}) - 2 \times r \times \sigma^2(\mathrm{VAS}) \\ &= 2 \times \sigma^2(\mathrm{VAS}) \times (1 - r)\end{aligned}$$

Taking the square root of each side gives for data from a sample:

$$\begin{aligned}\mathrm{SDd} &= \sqrt{}(\mathrm{SD}^2(\mathrm{VAS}_2 - \mathrm{VAS}_1)) = \mathrm{SD} \times \sqrt{2} \times (1 - r) \\ &= \sqrt{2} \times \mathrm{SEMp}.\end{aligned}$$

Hence, the two definitions for SDD are equivalent when the variances at time 1 and time 2 are equal. Because these variances are very similar indeed (cf. columns SD_1 and SD_2 in Table 2), only minor differences occur between the two methods when the data from the present paper are used. An interclass correlation coefficient (ICC; the ratio of the variance between subjects and the total variance) is generally accepted in the medical literature as the preferred method of quantifying reliability (Streiner and Norman, 2004). Because systematic differences between duplicate scores are lacking, the values of any type of ICC are nearly identical to Pearson's correlation coefficient for the data from the present study. Whereas ICC is a measure of the accuracy of an instrument with respect to variation in the measurement values between subjects, SEMp and SDD are related to within-subject variability of a measurement. SEMp and SDD therefore better capture the essence of the reproducibility of an instrument (Beckerman et al., 2001, see also de Vet et al., 2006 for the relationship and difference between ICC and SEMp as parameters of reliability and agreement, respectively). The definition of SDD by SEMp (including the use of ICC as a reliability coefficient) is more general because it can be applied to situations involving more than two repeated measures. Furthermore, it can be calculated for different study samples, i.e. by using the SD-value from one study sample and a reliability coefficient from another sample. In the most general case, calculations of SDD are based on a decision study following a generalizability study in which multiple factors are determined under several measurement conditions (e.g. observers, repetitions, sessions, occasions; Kropmans et al., 2002; Streiner and Norman, 2004).

References

Beckerman H, Roebroeck ME, Lankhorst GJ, Becher JG, Bezemer PD, Verbeek ALM. Smallest real difference, a link between reproducibility and responsiveness. Qual Life Res 2001;10:571–8.

Blanchard EB, Hillhouse J, Appelbaum KA, Jaccard J. What is an adequate length of baseline in research and clinical practice with chronic headache? Biofeedb Self-Regul 1984;12:323–9.

Chen ACN, Dworkin SF, Haug J, Gehrig J. Human pain responsivety in a tonic pain model: psychological determinants. Pain 1989;37:143–60.

Clark CT. Interocclusal appliance therapy. In: Mohl ND, Zarb GA, Carlsson GE, Rugh JD, editors. A textbook of occlusion. Chicago: Quintessence; 1988. p. 271–84.

Cronbach JL, Gleser GC, Nanda H, Rajaratman N. The dependability of behavioral measurements: theory of generalizability for scores and profiles. New York: Wiley; 1972.

Dao TTT, Lavigne GJ, Charbonneau JS, Feine JS, Lund JP. The efficacy of oral splints in the treatment of myofacial pain of the jaw muscles: a controlled clinical trial. Pain 1994;56:85–94.

de Vet HCW, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol 2006;59:1033–9.

Dworkin RH, Turk DC, Farrar JT, Haythornthwaite JA, Jensen MP, Katz NP, et al. Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. Pain 2005;113:9–19.

Ekberg EC, Vallon D, Nilner M. Oclusal appliance therapy in patients with temporomandibular disorders: a double-blind controlled study in a short-term perspective. Acta Odontol Scand 1998;56:122–8.

Farrar JT, Young Jr JP, LaMoreaux L, Werth JL, Poole RM. Clinical importance of changes in chronic pain intensity measured on a 11-point numerical pain rating scale. Pain 2001;94:149–58.

Gracely RH, Kwilosz DM. The descriptor differential scale: applying psychophysical principles to clinical pain assessment. Pain 1988;35:279–88.

Greenwood LF. Masticatory muscle disorders. In: Zarb GA, Carlsson GE, Sessle BJ, Mohl ND, editors. Temporomandibular joint and masticatory muscle disorders. Copenhagen: Munksgaard; 1994. p. 256–70.

Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chron Dis 1987;40:171–8.

Jacobson NS, Follette WC, Revenstorf D. Psychotherapy outcome research: methods for reporting variability and evaluating clinical significance. Behav Therapy 1984;15:336–52.

Jensen, M. The validity and reliability of pain measures for use in clinical trials in adults: review paper written for the initiative on methods, measurement, and pain assessment in clinical trials (IMMPACT) meeting, April 12–13 2003. Available from: www.immpact.org/meetings.html, IMMPACT-II page.

Kropmans ThJB, Dijkstra PU, Stegenga B, Stewart R, de Bont LGM. Smallest detectable difference in outcome variables related to painful restriction of the temporomandibular joint. J Dent Res 1999a;78:784–9.

Kropmans ThJB, Dijkstra PU, Stegenga B, van Veen A, de Bont LGM. The smallest detectable difference of mandibular function impairment in patients with a painfully restricted temporomandibular joint. J Dent Res 1999b;78:1445–9.

Kropmans ThJB, Dijkstra PU, Stegenga B, Stewart R, de Bont LGM. Repeated pain assessment in temporomandibular joint patients; decision making using uni- and multidimensional pain scales. Clin J Pain 2002;18:107–15.

Morley S. Vivid memory for 'everyday' pains. Pain 1993;55: 55–62.

Norušis MJ. SPSS 9.0. Guide to data analysis. Upper Saddle River, NJ: Prentice-Hall; 1999, p. 431–39.

Portney LG, Watkins MP. Foundations of clinical research. Applications to practice. New Jersey: Prentice-Hall; 2000.

Okeson JP. Orofacial pain: guidelines for assessment, classification, and management. Illinois: The American Academy of Orofacial Pain, Quintessence; 1996.

Raphael KG, Marbach JJ. A year of chronic TMPDS: evaluating patient's pain patterns. JADA 1992;123:53–8.

Rudy ThE, Turk DC, Kubinski JA, Zaki HS. Differential treatment responses of TMD patients as a function of psychological characteristics. Pain 1995;61:103–12.

Streiner DL, Norman GR. Health measurement scales, a practical guide to their development and use. Oxford: Oxford University Press; 2004.

van Grootel RJ, van der Glas HW, Buchner R, de Leeuw JRJ, Passchier J. Patterns of pain variation related to myogenous temporomandibular disorders. Clin J Pain 2005;21: 154–65.

Vendrig AA, Lousberg R. Within-person relationships among pain intensity, mood and physical activity in chronic pain: a naturalistic approach. Pain 1997;73:71–6.