3 Smallest detectable difference in outcome variables related to painful restriction of the temporomandibular joint

Kropmans ThJB, Dijkstra PU, Stegenga B, Stewart R, de Bont LGM
Journal of Dental Research 1999; 78 (3): 784 - 789
Reprinted with permission

Introduction

The selection of the most favorable therapeutic intervention for disorders characterized by pain and restricted movement of the temporomandibular joint is basically a matter of clinical decision making, which at present cannot be entirely based on scientific information (Kropmans et al., 1999). Restriction of mouth opening, pain, and function impairment are commonly used outcome variables in these disorders (Carlsson et al., 1980; Kopp and Wennerberg, 1983; Agerberg, 1987; Dworkin et al., 1988; Dworkin et al., 1990; Dahlstrom, 1993; Stegenga et al., 1993a; Chung, 1993; Lobbezoo-Scholte, 1994).

In outcome research, group means are generally compared and statistically analyzed. Statistically significant differences between pre- and post-therapeutic outcome variables are considered to reflect the clinical effects of the intervention. However, group mean comparison provides only average information about the effects of the intervention for individual patients of the samples. Moreover, the individual outcome of the variable ‘maximal mouth opening’ in patients with temporomandibular joint disorders may be biased by the interfering effects of pain. In addition, statistical significance does not necessarily imply clinical relevance (Jacobson et al., 1984). Observed improvement between pre- and post-therapeutic outcome variables not only reflects a therapeutic effect, but also the natural course of the disorder, biological fluctuation of the variables measured, and inconsistency in measurement (Dijkstra et al., 1995). In order to determine whether a change in an outcome variable is the result of a specific intervention, knowledge about the factors contributing to the variation in observed measurement scores (i.e., variation within subjects, variation between measurement occasions, and a residual variation) is necessary (Mitchell, 1979; Roebroeck et al., 1993; Lund et al., 1995). Classical psychometric analysis of repeated measurements is used to assess consistency in measurement results, which is currently quantified.
as a correlation coefficient, such as Pearson's \( r \), Kappa, or an intra-class correlation coefficient, depending on the type of data (Cronbach \textit{et al}, 1972; Shrout and Fleiss, 1979). However, a correlation coefficient does not provide information about systematic differences between observed measurement results (Bland and Altman, 1986). In case of low correlation, factors contributing to these systematic differences become even more important (Stratford, 1989; Lyons, 1991). An important shortcoming of the classical methods used in the current literature concerning reliability of outcome variables of temporomandibular joint disorders is that they do not provide information about changes in the variables. In the literature related to other fields of research than that of the temporomandibular joint, the standard error of the measurement procedure, the 95% confidence interval, and the smallest detectable difference have been proposed to be the appropriate measures for assessing changes in outcome variables (Cronbach \textit{et al}, 1972; Brennan, 1983; Roebroeck \textit{et al}, 1993).

The purpose of this study was to calculate these measures for assessing change in maximal mouth opening, pain intensity, and function impairment measurement. These findings, in turn, will provide a foundation on which to base clinical decision making in patients with complaints of the temporomandibular joint.

\section*{Materials and methods}

\textbf{Theoretical background}
Repeated measurement of a variable provides ‘observed scores’ with a mean and a standard deviation (i.e. descriptive statistics). Mean and standard deviation of these repeated measurements determine the distribution and the borders between which the true score is to be expected (Cronbach \textit{et al}, 1972; Veldhuijzen \textit{et al}, 1993). The observed score is in fact the sum of the unknown true score and a random ‘error’. In a population, the variance of the observed score (\( \sigma^2_o \)) is the sum of the variance of the true score of that population (\( \sigma^2_T \)) and the variance of the error score (\( \sigma^2_E \)). Consistency in measurement results is generally expressed as a correlation coefficient defined as \( r = \sigma^2_T / \sigma^2_O \).

Substituting \( \sigma^2_T + \sigma^2_E \) for \( \sigma^2_O \) (\( r = \sigma^2_T / (\sigma^2_T + \sigma^2_E ) \)), illustrates that the variance of the error score determines the reliability of the measurement (Mitchell, 1979; Roebroeck \textit{et al}, 1993; Veldhuijzen \textit{et al}, 1993). The standard error of a measurement procedure (SEM, \( r \)) is calculated from the known sample standard deviation (SD) and the correlation.
coefficient (r) of the measurement procedure used for that sample: 
$$SEM_p = SD \times \sqrt{1-r} \ (Cronbach \ et \ al, \ 1972; \ Mitchell, \ 1979; \ Ottenbacher \ et \ al, \ 1988; \ Veldhuijzen \ et \ al, \ 1993).$$ 
The corresponding 95% confidence interval, in which the true score (drawn from a normally distributed population) is expected, is $$[\pm 1.96 \times SEM_p]$$. The broader the limits of the 95% confidence interval, the less confident the estimation of the true score, and, as a consequence, the less confident the detection of real change due to intervention. Thus, knowledge about the standard error of the measurement procedure is necessary to decide whether such a change has occurred (Ottenbacher \ et \ al, \ 1988; Hayes, \ 1992).

Moreover, when analyzing a difference between two consecutive observations, one has to consider the standard error of the measurement procedure of the observed score for both the first $$SEM_{p(first)}$$ and the second $$SEM_{p(second)}$$ observations. The smallest detectable difference is known as the measure of statistically significant change between two independently obtained measurements. Given a probability value of $$\alpha = 0.05$$ as indication for statistical significance, the smallest detectable difference is estimated as $$1.96 \times \sqrt{SEM_{p(first)}^2 + SEM_{p(second)}^2} \ (McNemar, \ 1969; \ Guyatt \ et \ al, \ 1987)$$. Assuming that the standard error of the measurement procedure of the observed score of the first and the second observations are equal, the smallest detectable difference is $$1.96 \times \sqrt{2} \times SEM_p$$. To detect a statistically significant change between two separate observations this change must be at least the smallest detectable difference of the measurement procedure.

**Outcome variables**

Maximal mouth opening is the maximal inter-incisal distance and is usually measured with a millimeter-ruler (Murakami \ et \ al, \ 1987; Nitzan \ et \ al, \ 1990; Nitzan \ et \ al, \ 1991; McCain \ et \ al, \ 1992; Stegenga \ et \ al, \ 1993b; Van Sickels and Dolezal, 1994). Only a few studies have been published in which correlation coefficients and standard deviations of the samples were presented. Out of the correlation coefficients and standard deviations published by Wood and Branco (1979), Agerberg (1987) and Stegenga \ et \ al. \ (1993b) we calculated the standard error of the measurement procedure, the 95% confidence interval, and the smallest detectable difference for maximal mouth opening. According to the International Association for the Study of Pain, pain is defined as a subjective unpleasant sensory and emotional experience with actual or potential tissue damage or described in terms of such damage (Merkskey, 1979). Because pain is a sensation with biological
components, it can be influenced by movement, pressure, temperature, emotions, social surroundings and rest. The McGill Pain Questionnaire is a widely used instrument to assess pain and its different dimensions (Melzack, 1975; van der Kloot et al., 1995). The perceived intensity of pain is subjective and therefore differs between and within subjects; it can be estimated on a visual analogue scale (Jensen, 1986; McCormack et al., 1988; Jensen, 1993; Mantha, 1993). A visual analogue scale is a 100 mm line with two opposing descriptions of pain representing the absence of pain and the worst imaginable pain, at each end of the line (Scott, 1976; Jensen, 1986; Vallerand, 1991; Stegenga et al., 1993b; Kuwahara et al., 1994). Although widely used, correlation coefficients and standard deviations of pain scores on a visual analogue scale have not been reported in the temporomandibular joint literature. The Dutch language version of the McGill pain questionnaire (van der Kloot et al., 1995) assesses, besides pain adjectives, three different intensities of pain of the last week on a visual analogue scale: the actual pain, the minimal pain and the maximal pain. Out of mean correlation coefficients and standard deviations published by van der Kloot et al. (1995), we estimated the standard error of the measurement procedure, the 95% confidence interval and the smallest detectable difference for the scores. Problems with eating, chewing, speech, and laughter (i.e. mandibular function impairment) are related to a painful restriction of the temporomandibular joint (Stegenga et al., 1993b). The only available valid and reliable instrument to assess mandibular function impairment is the Mandibular Function Impairment Questionnaire which scores perceived difficulty of representative mandibular functions in relation to jaw complaints. Reliability of the Mandibular Function Impairment Questionnaire was tested comparing a verbal and a non-verbal version of the Mandibular Function Impairment Questionnaire. The total function impairment score is the summation of the scores of 17 representative items on a Likert scale, each scoring from 0-4. From the correlation coefficient and the standard deviation of the sample published by Stegenga et al. (1993b) we calculated the standard error of the measurement procedure, the 95% confidence interval, and the smallest detectable difference of the function impairment score.

**Patient studies**

Analyzing the reliability of maximal mouth opening measurement in normal subjects was based on the studies of Wood and Branco (1979) (n=19) and Agerberg (1987) (n=67). Stegenga et al. (1993b) tested the
reliability of maximal mouth opening in a group of patients with complaints of the temporomandibular joint (n=46). The reliability of scores of pain intensity on a visual analogue scale in the study of van der Kloot et al. (1995) was based on a heterogeneous group of pain patients (n=92). The reliability of the Mandibular Function Impairment Questionnaire was based on a group of patients with complaints of the temporomandibular joint described by Stegenga et al. (1993b) (n=86). Demographic findings, correlation coefficients and standard deviations of these samples are listed in table 3.1. The data out of previous studies were taken for secondary analysis and the protocol was approved by the medical ethic commission of our institution.

Table 3.1 The demographic findings of the subjects presented by different authors

<table>
<thead>
<tr>
<th>Authors</th>
<th>Subjects</th>
<th>Gender</th>
<th>age yrs</th>
<th>MMO mm</th>
<th>VAS mm</th>
<th>MFI units</th>
<th>r</th>
<th>sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wood and Branco (1979)</td>
<td>healthy</td>
<td>9/10</td>
<td>23.6</td>
<td>53.7</td>
<td>.96</td>
<td>7.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agerberg (1987)</td>
<td>healthy</td>
<td>24/43</td>
<td>20</td>
<td>35.7</td>
<td>.87</td>
<td>5.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stegenga et al. (1993b)</td>
<td>TMJ patients</td>
<td>4/24</td>
<td>28</td>
<td>44.5</td>
<td>.97</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>van der Kloot et al. (1995)</td>
<td>pain pat.</td>
<td>66/26</td>
<td>45</td>
<td>32.0</td>
<td>.82</td>
<td>24.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>actual minimal</td>
<td></td>
<td></td>
<td>17.6</td>
<td>.76</td>
<td>15.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>maximal</td>
<td></td>
<td></td>
<td>69.3</td>
<td>.88</td>
<td>23.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stegenga et al. (1993b)</td>
<td>TMJ patients</td>
<td>71/9</td>
<td>25.2</td>
<td>23.1</td>
<td>.95</td>
<td>12.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TMJ = temporomandibular joint; actual, minimal and maximal pain; MMO=maximal mouth opening; MFI=mandibular function impairment

Results

The smallest detectable difference of maximal mouth opening in healthy subjects varies between 4 and 5 mm. These scores are based on a standard error of the measurement procedure in case of one repetition on one occasion of 1.6 mm (Wood and Branco, 1979) versus 2.1 mm and 1.9 mm in the study of Agerberg (1987) and Stegenga et al. (1993b), respectively. Statistically significant change between two independently obtained measurement scores, the smallest detectable difference, is calculated as $1.96 \times \sqrt{2} \times 1.6 \text{ mm}$, $2.1 \text{ mm}$ and $1.9 \text{ mm}$, respectively,
depending on data presented in each reliability study. The standard error of the measurement procedures, the corresponding 95% confidence intervals and the smallest detectable differences of the outcome variables are listed in table 3.2.

<table>
<thead>
<tr>
<th>Outcome variable</th>
<th>Occ</th>
<th>Obs</th>
<th>Rep</th>
<th>SEM</th>
<th>95% CI</th>
<th>SDD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximal mouth opening</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wood en Branco (1979)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1.6</td>
<td>[±3]</td>
<td>4 mm</td>
</tr>
<tr>
<td>mean score</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1.0</td>
<td>[±2]</td>
<td>3 mm</td>
</tr>
<tr>
<td>Agerberg (1987)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2.1</td>
<td>[±4]</td>
<td>5 mm</td>
</tr>
<tr>
<td>Stegenga (1987)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1.9</td>
<td>[±4]</td>
<td>5 mm</td>
</tr>
<tr>
<td>Pain scores on visual analogue scores</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>van der Kloot et al. (1995)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VAS actual pain 0 - 100 mm</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td>[±20]</td>
<td>28 mm</td>
</tr>
<tr>
<td>VAS minimal pain 0 - 100 mm</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>[±16]</td>
<td>22 mm</td>
</tr>
<tr>
<td>VAS maximal pain 0 - 100 mm</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>[±16]</td>
<td>22 mm</td>
</tr>
<tr>
<td>Total Mandibular Function Impairment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stegenga et al. (1993b)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>[±6]</td>
<td>8 units</td>
</tr>
</tbody>
</table>

The smallest detectable difference improved from 4 to 3 mm in the study of Wood and Branco (1979) after three repetitions. The smallest detectable difference is 5 mm for the group of patients with complaints of the temporomandibular joint in the study of Stegenga et al. (1993). The standard error of the measurement procedure for pain scores on a visual analogue scale as used by van der Kloot et al. (1995) was 10 mm for actual pain versus 8 mm for both ‘minimal’ and ‘maximal’ pain. The corresponding smallest detectable difference for actual pain was 28 mm, and 22 mm for both ‘minimal’ pain and ‘maximal’ pain, respectively. The standard error of the measurement procedure of Total Function Impairment as measured with the Mandibular Function Impairment Questionnaire, published by Stegenga et al. (1993b) was 3 units. The corresponding smallest detectable difference, calculated from the inter-method correlation coefficient and the standard deviation, was 8 units.
Discussion

An improvement in maximal mouth opening of at least 5 mm is needed to detect a statistically significant change in maximal mouth opening. These results confirm those of Agerberg et al. (1987). However, Wood and Branco (1979) and Agerberg et al. (1987) measured healthy subjects. In a group of patients with complaints of the temporomandibular joint, Stegenga et al. (1993b) reported a mean range of opening after passive stretch of 44.5 mm with a standard deviation of 11 mm. Descriptive statistics suggest that a considerable part of that sample was restricted in maximal mouth opening and a considerable part was not. The smallest detectable difference in their sample was also 5 mm. Thus, the smallest detectable difference of 5 mm may be considered generalizable to the population of healthy subjects and to patients with restricted mouth opening.

Dworkin et al. (1988) presented maximal assisted mouth opening measurements of seven observers (mean 53.1 mm, standard deviation of 1.2 mm). The scientific relevance of the latter reliability study, however, is limited because the standard deviation presented in this study was calculated from only one measure of central tendency i.e. the mean value of each single observer. The small inter-observer standard deviation presented in this study does not represent the variance in maximal assisted mouth opening of the subjects within each sample and is therefore artificial.

Since maximal mouth opening is a non-stable biological variable, its behavior is also influenced by repeated measurements (Agerberg, 1987; Mezitis et al., 1989). However, if measured repeatedly, the standard error of the measurement procedure decreases by increasing the number of repetitions (Table 3.2) (Cronbach et al., 1972). The influence of repeated measurements of this variable in patients with painful and restricted temporomandibular joints is not known. Therefore, a future reliability study of repeated measurement of maximal mouth opening in patients with painful and restricted temporomandibular joints is necessary.

The reliability of ‘pain scores’ in patients with painfully restricted temporomandibular joints was assessed by Dworkin et al. (1990). However, only an intraclass correlation coefficient was reported; the variance of the method used was not (Stratford, 1989; Bland and Altman, 1986; Dworkin et al., 1990). Van der Kloot et al. (1995) used
the visual analogue scales in a heterogeneous group of non-temporomandibular joint pain patients. An increase or decrease in pain intensity of at least 28 mm on a visual analogue scale for ‘actual’ pain is needed to detect statistical significant change over time. The results of this study are generalizable to a heterogeneous group of pain patients. However, it only provides an indication as to how much change is needed in a painful and restricted temporomandibular joint to detect a significant difference.

In the present study, we calculated the smallest detectable difference for minimal, maximal and actual pain intensities (Table 3.2). The left and right extremes of a 0-100 mm scale represent smallest detectable differences that differ from the middle segment. These differences may be explained, in part, by previous reports that a visual analogue scale is not scoring equal intervals (Ohnhaus, 1975; Price et al., 1983). The choice of scoring interval or ratio for scores on a visual analogue scale seems to be based on convention and convenience rather than on scientific research (McCormack et al., 1988). Group comparisons based on mean scores on visual analogue scales may lead to unreliable conclusions about change in pain because of the different standard error of the measurement procedures, the 95% confidence intervals and the smallest detectable differences corresponding with different pain intensities. For group comparison actual, minimal, and maximal pain should be registered on separate visual analogue scales. The corresponding smallest detectable differences for different pain intensities and for different groups of pain patients should be known to decide about change in experienced pain.

Instruments measuring function impairment are scarce (Schiffman, 1992; Murakami, 1992; Stegenga et al., 1993). The Mandibular Function Impairment Questionnaire is a valid and reliable instrument to measure mandibular function impairment (Stegenga et al., 1993b). An improvement of 8 units on this 0 to 68 units scaled questionnaire appears to be needed for statistical significance. For this study we used the inter-method correlation coefficient to estimate the smallest detectable difference of the Mandibular Function Impairment Questionnaire. Scientifically correct information about the measurement error of the Mandibular Function Impairment Questionnaire should be obtained from the inter and intra-observer variation of this measure, however. Instability and inconsistency of each separate variable (i.e. maximal mouth opening, pain and mandibular function impairment) has been previously reported (Price et al., 1983; Agerberg, 1987; Mezitis et al.,
Maximal mouth opening, pain and mandibular function impairment are obviously inter-related (Stegenga et al., 1989; de Bont and Stegenga, 1993). However, the exact influence of pain on mouth opening and function impairment has not yet been quantified. Analysis of repeated measurement scores should provide information about the factors contributing to the variation in observed scores in patients with painful and restricted temporomandibular joints.

A statistically significant difference between pre and post intervention groups is of limited importance when this difference is smaller than the smallest detectable difference of the outcome variable. The smallest detectable difference of maximal mouth opening in normal subjects is 5 mm, thus the maximal mouth opening has to decrease at least 5 mm to indicate the possibility of a temporomandibular joint disorder. The smallest detectable difference of an outcome variable provides the clinician with knowledge regarding the level of improvement necessary for being therapeutically successful in a specific group of patients. For individual patients the smallest detectable difference provides average information on which to base clinical decision making with regard to the ‘success’ of a therapeutic intervention can be based. The smallest detectable difference of different outcome variables in patients with painful and restricted temporomandibular joints have not yet been quantified. Therefore, more explicit research is needed in this patient population. Since the smallest detectable difference is a measure of statistical as well as clinical relevance this will lead to improved clinical decision making in patients with painful and restricted temporomandibular joints.
References


Chapter 3: smallest detectable difference 38


*Chapter 3: smallest detectable difference* 40


