8 General discussion

The major pathologic events characterizing temporomandibular joint osteoarthritis include synovitis and internal derangements, giving rise to pain and restricted mobility of the temporomandibular joint, possibly resulting in mandibular function impairment. The treatment is predominantly symptomatic, and several treatment modalities have been reported to be successful in reducing pain and increasing mobility.
Fletcher and Sackett (2000) distinguish three levels of evidence from randomized clinical trials for the efficacy of a treatment modality, i.e. evidence from meta-analyses and systematic reviews of randomized clinical trials, evidence from one or more large randomized clinical trials, and evidence from small randomized clinical trials, respectively.
From our review (chapter 2), it has become apparent that the evidence for the beneficial effects of physical therapy, arthrocentesis and arthroscopic surgery is incomplete. No properly conducted clinical trials are available to perform a meta-analysis. In addition, it became apparent that a wide variety of outcome measures have been used for evaluating treatment modalities for temporomandibular joint disorders. This variety makes comparison between studies difficult, if not impossible. Moreover, the validity and reliability of a considerable number of these measures has not been sufficiently established. Thus, there is not only a need for properly designed and conducted trials, but also for outcome measures that can be generally applied to enable future comparison between studies and meta-analysis of different trials.
Since the major aim of treating temporomandibular joint osteoarthritis, especially when associated with non-reducing disk displacement, is to improve mandibular function by reducing pain and restriction of motion, the logical primary outcome measure would be ‘mandibular function impairment’. In order to judge whether a treatment modality is successful or not, one should define a proper threshold for the primary outcome measure. In this thesis, we have postulated to apply the smallest detectable difference for this purpose.
A major advantage of the smallest detectable difference is that it exceeds the natural course of the disease, the biological variance of the outcome variable and measurement error and is generalizable to all patients that comprise the sample. In temporomandibular joint osteoarthritis accompanied by a disk derangement impaired, mandibular function is largely due to pain and restricted mouth opening (De Laat *et al*, 1993;

Stegenga *et al*, 1993). Because of the interrelationship between pain, range of opening and mandibular function, the biological variation of each of these variables is influenced by the other variables. It has been made plausible that the natural course of the disease contributes to changes within treatment as well as in no-treatment groups (Lundh *et al.* 1992; Sato *et al.* 1997), and this aspect should, therefore, be incorporated in the outcome measure.

The smallest detectable difference is a measure for statistically significant change that can be used in the individual patient as a measure of responsiveness. Responsiveness is defined as the ability of an outcome instrument to detect clinically important changes in a specific condition. However, there is no consensus on the appropriate strategy to quantify responsiveness (Bennekom *et al*. 1996). Most of the approaches are based on the average change in scores relative to baseline in self rated clinically stable and improved patients. These responsiveness ratios are unitless ratios of the variances of only one sources of variation (measurement days) in a changed and a non-changed group of patients (Bronfort and Bouter, 1999). The smallest detectable difference, however, is based on the analyses of error variance of different sources of variation (observer, days and repetitions and their interactions) causing variance around the observed (change) scores. The higher the variance (large SDD' s), the less responsive the outcome variable is expected to be. Although responsiveness is not the primary scope of this thesis, by increasing the number of repetitions the smallest detectable difference is reduced and therefore responsiveness of the outcome variable is likely improved.

Based on their magnitude, as described in chapter 4,5 and 6, we intuitively considered the smallest detectable difference as clinically relevant. Although a mean between-groups difference in mouth opening of, for example, 4 mm may be statistically significant, its clinical relevance is far from obvious. Methods used to quantify clinical significance require the use of at least one valid external criterion of improvement (Bronfort and Bouter, 1999). It is obvious that different observers, different measurement days and different repetitions cause variance around the first and the second observation of the subject. We realize that if the smallest detectable difference is substantially reduced by minimizing this variation, for instance by increasing the number of repetitions, we could get in conflict with the minimally clinically important difference. The minimally clinically important difference is defined as the smallest difference that patients would consider beneficial and which would form the basis for changing therapeutic management (Jaetschke *et al,* 1989). However, the difference that a patient would consider beneficial is a highly individual response on a therapeutic intervention and is not necessarily covered by one of the outcome variables or their effect sizes. Prior to the treatment we measured patients' expectations by requiring them the following question on a visual analogue scale as: "How much change in pain is needed to be satisfied with the treatment result?" Furthermore, following treatment, we assessed patients pain response shift with the question: "How much pain did you have before treatment?". A ratio between these two measures could have lead to a measure of individual responsiveness or clinical significance. However, the size of our randomized clinical trial do not allow extensive statistical analyses to further investigate the clinical significance of the smallest detectable difference.

The random designs used in generalizability and decision studies to estimate the smallest detectable difference of the outcome measurement instrument allow generalization to all the facets involved, i.e., patients, observers, measurement days and repetitions. By contrast, classical reliability assessment and a responsiveness ratio of an outcome instrument only allows generalization to the observers involved in the clinical trial. Because of this generalization, the smallest detectable difference is relatively large compared with the measurement error according to the classical approach, especially in case of only one repetition of the measurement as is the case in most of the clinical trials. In a generalizability study, the standard error of measurement (i.c. absolute error variance) is a cumulation of error variances of all facets and interactions involved. The standard error of measurement in the classical approach strongly depends on the observed reliability and variance within the experimental design and is, therefore, not generalizable. Furthermore a statistically significant difference between groups has been achieved by a specific treatment team, the skills of which are not generalizable to other (surgical) teams. The smallest detectable difference, on the other hand, is generalizable because of the use of a random model. In a random model, all facets (observer, days and repetitions) contributing to the error variance are randomly chosen so generalization is allowed to all of these facets.

The smallest detectable difference applied in a clinical trial provides firm scientific arguments with respect to the clinical decision about treatment. The evidence provided by a randomized clinical trial (or, even better, by a meta-analysis from a systematic review of randomized clinical trials) not only provides information about the efficacy of the treatment in a group of similar patients, but also provides information about how much improvement can be expected. The smallest detectable difference also provides a uniform criterion to judge the outcome of a treatment on an individual level (i.e., successful vs. non-successful). Thus, with the smallest detectable difference a uniform, statistically based ‘threshold of efficacy’ is available for wide application.

Generalizability and decision studies also have limitations that must be recognized. The sample, based on which the smallest detectable difference is estimated, must be homogeneous with respect to the patients for whom the intervention is indicated. In our sample, patients were tested on prognostic relevant factors and pre-treatment group differences before randomization with respect to age and gender, mouth opening, pain and function impairment. No statistical significant differences between the treatment groups and the sample we used for estimating the smallest detectable difference, were found.

The stability of variance components is a major limitation of most of the studies based on the generalizability theory. We used 25 patients in the generalizability and decision studies, while Cronbach *et al.* 1972 recommended a sample size of at least 100 patients. Although we found small standard errors of the estimates (unpublished) and high generalizability coefficients, small sample sizes remain a matter of concern (Roebroeck *et al*, 1993).

Another point of concern is that patients with restricted mobility due to internal joint derangements may have a smaller biological variance in mouth opening than patients with, for example, myofascial pain. In fact, we are dealing with different diagnostic entities among patients with temporomandibular joint disorders. For all of these different diagnostic groups, separate SDDs should be estimated, primarily for the most relevant outcome variable(s) associated with the particular disorders, but also for more common outcome variables of different disorders, such as maximal mouth opening.

The SDDs found for the different variables are relatively large. The magnitude of the SDDs is attributed to its generalizability. A well trained observer may claim to perform a measurement far more precise than do ‘all the clinicians’. By measuring more precise, less measurement error and thus a smaller SDD is obtained. An assigned SDD, based on a fixed observer and calculated out of the correlation coefficient of his personal repeated measurement results and the variance of these results in an individual patient, could be appropriate to calculate a ‘personalized

SDD' for each clinician. This approach still needs to be investigated. So far, it is not clear whether repeated measurements by an assigned observer in one individual are generalizable to the universe of all possible repetitions in that individual. In our samples the main effects of the observers were zero or very small and analyses as a mixed design with random facets days and repetitions and fixed observers did not reduce the smallest detectable difference. The smallest detectable difference, as presented, is based on the results of a group analyses of repeated measurements. Although generalization is allowed to all facets included (i.e. subjects, observers, days and repetitions), no information is available about an individual SDD based on repeated observations in individual subjects and individual observers.

On the basis of the results described in this thesis it is recommended to apply the smallest detectable difference of outcome variables because it supports a reasoned selection of treatment modalities. As we have illustrated in chapters 4 and 5, a fictive patient with impaired mandibular function due to pain and limited mouth opening improved on mouth opening but not on pain and function impairment. The decision to prescribe medication or perform arthrocentesis to relief the pain and to improve mandibular function is based on proper criteria of reliable change rather than on the belief of patient and clinician.

The literature still lacks properly conducted clinical trials evaluating common treatment modalities for patients with temporomandibular joint osteoarthritis. Available (non-randomized and non-controlled) studies use a wide variation of outcome measures to decide on the success of treatment. The smallest detectable difference appears to be a valuable contribution to evidence based medicine and clinical decision making in good clinical practice and research.

The large SDDs analysed in our study were based on repeated measurements in groups of patients. To obtain smaller or even personalized SDDs (i.e., for each clinician his own SDD) rather than a generalizable SDD (i.e., one SDD for all clinicians), further research is necessary. Future investigations to obtain personalized SDDs should focus on repeated measurement by assigned observers in an individual to analyse whether these repetitions of that individual are representative for the universe (population) of all repetitions of that individual.

Further research is needed to investigate the SDD for its responsiveness, and clinical significance. We recommend to include analyses of patients' expectations of pain reduction and patients perceived pain history because we belief in the interaction between disappointment (high expectations versus bad remembering) and true results of efficacy. Reliability analysed in a generalizability and a decision study has turned out to be complementary to the classical approach of reliability assessment. This classical approach does not inform the researcher or the clinician about the amount of error variation attributed to different sources of variation (such as observers, days and repetitions) and their interactions. Knowledge of the amount of variation of all facets in a measurement design obviously improves the reliability, and these are exclusively analysed in a generalizability study. Where correlation coefficients inform about linearity of association and not about measurement error, the smallest detectable differences, recommended as a 'new' measure of reliability depends on the amount of measurement error and is expressed in the same unit as the measurement device tested. Different SDD's should be established for different patient categories. To analyse exclusively the relative effects of non-invasive and minimally invasive treatment modalities in patients with temporomandibular joint osteoarthritis associated with non-reducing disk displacement, large sample sizes are inevitable. It is questionable whether these amount of patients are available within one center, making a multicenter trial a necessary alternative. The efficiency in addressing specific methodological issues related to multicenter trials, such as homogeneity of patient characteristics and diagnosis, could be enhanced by a database management information system based on Internet technology.

Recommendations

Based on this thesis, the following recommendations are suggested:

- Implementation of the smallest detectable difference in clinical practice for the benefits of patients and clinicians information about real expectations and percentages of success.
- To continue research on personalized smallest detectable differences (i.e. each clinician his/her own smallest detectable difference).
- To investigate responsiveness and the minimally clinically important difference in relation to the smallest detectable difference.
- To develop a Relational Database Management System based on Internet technology for the purpose of a structural standardized data collection. Large multi-center randomized clinical trials may

be supported by such a system and the smallest detectable difference could be integrated as a reliable criterion of success.

References

Bennekom CAM, Jelles F, Lankhorst GJ, Bouter LM (1996). Responsiveness of the Rehabilitation Activities Profile and the Barthel Index. J. Clin. Epidemiol 40 (1); 39-44.

Bronfort G, Bouter LM (1999). Responsiveness of general health status in chronic low back pain: a comparison of the COOP Charts and SF-36. Pain 83; 201 - 209.

Cronbach, JL, Gleser, GC, Nanda, H, Rajaratman, N (1972): The dependability of behavioural measurements: Theory of generalizability for scores and profiles. New York: John Wiley and Sons.

De Laat A, Horvath M, Bossuyt M, Fossion E, Baert AL (1993). Myogenous or arthrogenous limitation of mouth opening: correlations between clinical findings, MRI, and clinical outcome. *J Orofac Pain* 7:150-155.

Fletcher S, Sackett DL (2000) Levels of Evidence.
http://cebm.jr2.ox.ac.uk/docs/levels.html

Jaeschke R, Singer J, Guyatt GH.(1989) Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials. Dec;10(4):407-15.

Lundh H, Westesson PL, Erikkson L, Brooks SL. (1992) Temporomandibular joint disk displacements without reduction. Treatment with flat occlusal splint versus no treatment. *Oral Surg Oral Med Oral Pathol*; 73:655-658

Roebroeck M, Harlaar J, Lankhorst GJ (1993). The application of generalizability theory to reliability assessment: an illustration using isometric force measurements. *Phys Ther* 6:386-401.

Sato S, Goto S, Kawamura H. Motegi K (1997). The natural course of nonreducing disc displacement of the temporomandibular joint: relationship of clinical findings at initial visit to outcome after 12 months without treatment. *J Orofac Pain* ; 11:315-320

Stegenga B, de Bont LGM, de Leeuw R, Boering G (1993). Assessment of mandibular function impairment associated with temporomandibular joint osteoarthrosis and internal derangement. *J Orofac Pain* 7:183-195.