The handle http://hdl.handle.net/1887/28732 holds various files of this Leiden University dissertation.

**Author:** Marks, Miriam  
**Title:** Treatment of patients with hand osteoarthritis: outcome measures, patient satisfaction, and economic evaluation  
**Issue Date:** 2014-09-11
ABSTRACT
The objective was to identify all outcome measures used in studies on trapeziometacarpal osteoarthritis (TMC OA) and evaluate their measurement properties. In a two-step systematic literature review, we first identified studies including TMC OA patients and extracted all outcome measures. They were categorized according to the Outcome Measures in Rheumatology (OMERACT) core set for OA including five dimensions: pain, physical function, global assessment, imaging, and quality of life (QoL). Secondly, we retrieved articles on the measurement properties of the identified outcome measures for TMC OA patients. First, 316 articles including 101 different outcome measures were identified, addressing the OMERACT pain and function domains most frequently but under-representing QoL. Second, 12 articles investigating measurement properties of 12 outcome measures were identified. The methodological quality of these studies was poor to fair, implying that based on the literature no recommendations to use any of the outcome measures can yet be made.
INTRODUCTION

Numerous studies have evaluated conservative and surgical treatments for patients with trapeziometacarpal osteoarthritis (TMC OA), with both approaches generally found to be effective in reducing pain and increasing function. Several specific sets of outcome measures, known as core sets, are considered relevant to the best way of measuring treatment outcomes for TMC OA. Angst et al. proposed a core set to assess outcomes after resection interposition arthroplasty of the TMC joint; this consisted of the Short Form 36 (SF-36), Disabilities of the Arm, Shoulder and Hand questionnaire (DASH) or the Patient-Rated Wrist Evaluation (PRWE), and a customized form including assessment of range of motion (ROM), strength, and other clinical tests. Although showing good construct validity in this particular study, the reliability of the customized form and responsiveness of the whole set have not been investigated.

Three other core sets are available; they do not recommend specific outcome measures but rather areas that are relevant for patients with conditions affecting the hand. Based on the International Classification of Functioning, Disability and Health (ICF), a comprehensive and brief core set have been developed and validated to assess patients with any hand condition. These two detailed and complex core sets are known mainly to hand therapists and are not widely implemented in clinical practice. A simpler, more general core set of OA outcome measures (hip, knee, hand) was developed at the Outcome Measures in Rheumatology (OMERACT) III conference. It is intended to serve as an international standard for clinical trials. This set contains the domains ‘pain’, ‘physical function’, ‘patient’s global assessment’, ‘joint imaging’, and ‘quality of life (QoL)’, but, like the ICF core set, it does not comprise specific outcome measures.

In research and daily practice, decisions for treatments are made, amongst others, based on the results of health status questionnaires. Before such an instrument may be implemented, its measurement properties, such as reliability, validity, and responsiveness, should be assessed and considered adequate for the target population. It is important to use reliable and valid outcome measures in order to avoid biased results and conclusions. Quality criteria for evaluating measurement properties of health status questionnaires have been introduced and are widely accepted. However, these criteria do consider the outcome measure itself but not the methodological quality of the study. To evaluate whether a study on a specific outcome measure is of good methodological quality, the COnsensus-based Standards for the selection of health status Measurement INstruments (COSMIN) checklist has been recently developed.

Given the high prevalence of TMC OA and the many available treatment methods, a standardized assessment is essential for comparing the interventions and providing evidence of best practice. So far, it remains unclear to what extent researchers are using valid and reliable assessment tools in TMC OA studies and whether these meet the recommendations of the core sets mentioned previously. Furthermore, the methodological quality of studies investigating measurement properties of outcome measures for hand patients has not been investigated yet.

In order to identify suitable outcome measures and to make recommendations for outcome measures to be used for patients with TMC OA, our objectives were to (1) identify all subjective and objective outcome measures used in clinical trials of conservative and surgical treatments of TMC OA; (2) to relate them to the OMERACT core set; and (3) to evaluate the measurement
properties of standardized outcome measures employed in patients with TMC OA as well as the methodological quality of these studies.

METHODS

We performed this systematic literature review in accordance with the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) statement for developing study protocols and reporting systematic reviews22, 23. The review protocol was registered in the Netherlands National Trial Register (no. 2602).

**Step 1: Literature search for TMC OA studies**

An experienced librarian performed the search for articles published up to November 2010 in the following databases: PubMed, EMBASE, Web of Science, Cochrane Library, CINAHL, Academic Search Premier, ScienceDirect, and PEDro. The following search strategy was applied to PubMed and optimized for the other databases: (“trapeziometacarpal osteoarthritis” OR “trapezio metacarpal osteoarthritis” OR “carpometacarpal osteoarthritis” OR “carpo metacarpal osteoarthritis” OR “thumb osteoarthritis” OR ((Osteoarthritis OR Osteoarthroses OR Osteoarthritides OR Osteoarthritis OR Osteoartrosis OR Osteoarthritis OR “Degenerative Arthritis” OR “Degenerative Arthritis” OR Arthrosis[tw] OR Arthroses) AND (Carpometacarpal Joints OR Carpometacarpal Joint OR carpometacarpal OR trapeziometacarpal OR thumb OR thumbs OR “thumb base” OR carpometacarpal” OR ((Metacarpus OR Metacarpal) AND (Carpal OR Carpus OR Carpo OR Carpi)) OR ((Trapezium OR Trapezoid OR Trapezium OR Trapezial) AND (Metacarpus OR Metacarpal OR Carpal OR Carpus OR Carpo OR Carpi))))). As language restriction is unreliable or not possible in all databases, we conducted the search without any such restrictions.

Inclusion criteria for the review were (a) clinical study involving a minimum of 10 people with TMC OA who had received any conservative or surgical treatment for TMC OA; (b) study designs including all randomized controlled trials and observational (prospective or retrospective) studies; (c) studies in which the effectiveness of the treatment was evaluated with at least one outcome measure; (d) the paper was written in English or German.

Exclusion criteria were (a) studies investigating patients with generalized OA; (b) studies in which the results of patients with TMC OA could not be separated from those of patients with other conditions; (c) reviews, case reports, post-mortem and veterinary studies, and conference abstracts not published as full journal articles, because they lacked full information about the study design; and (d) studies not in English or German, as we had no reliable translators.

Two independent reviewers reviewed the titles and abstracts that had been identified. The full texts of the selected abstracts were retrieved and again analyzed independently by two of the authors. Consensus on inclusion of the studies was reached by discussion.

We checked the references of the included articles to find other suitable papers and subjected them to a similar selection process.

Data for the following variables were extracted using a predefined form: authors, publication year, number of patients, patient demographics, intervention, and follow-up period, as well as all objective and subjective outcome measures used in the studies. The level of evidence was determined using the slightly modified rating scheme described by Wright et al. 24.
Some authors analyzed the same study population more than once and presented their findings in several publications. These articles were analyzed as individual studies, in case inadequate descriptions of the study populations made it impossible to identify the overlap.

We classified all concepts included in the outcome measures, such as pain or strength, according to the five domains of the OMERACT core set (‘pain’, ‘physical function’, ‘patient’s global assessment’, ‘joint imaging’, ‘QoL’). We chose this core set as the reference tool because it is simpler and better known to hand surgeons than the complex ICF concept. For the purpose of this study, the domain ‘function’ included isolated functions of the hand (such as extending the thumb) and activities of daily living (ADLs) making use of the hand; ‘global assessment’ was defined as an overall assessment of the hand condition, including treatment satisfaction, symptom improvement, and disease activity. ‘QoL’ was defined as a multidimensional appraisal of various aspects of health, including pain and function. ‘Imaging’ included all techniques such as radiography or magnetic resonance imaging. Given that some outcome measures cover more than one concept, each item, element, or dimension of a combined outcome measure was analyzed separately to assign it to several corresponding OMERACT domains.

Step 2: Measurement properties

In the period up to April 2012, we performed a second literature search on the measurement properties of the identified outcome measures for TMC OA in the databases mentioned previously. We applied the same strategy as in step 1 adding the following terms on measurement properties: ...

We included studies with a population of at least 50% of the patients suffering TMC OA or analyzing patients with TMC OA as a subgroup, and evaluating any measurement property of an outcome measure revealed in step 1, regardless of whether the investigation of measurement properties was the primary objective of the study or only mentioned tangentially. Studies investigating patients with hand OA, for example, where the proportion of TMC OA patients was less than 50%, were excluded.

The following eight measurement properties of the outcome measures were rated according to the criteria developed by Terwee et al., which we slightly modified for our purpose (see definitions in Appendix 1): internal consistency, content validity, criterion validity, construct validity, reproducibility (agreement and reliability), responsiveness, floor or ceiling effects, and interpretability. Two reviewers independently extracted all these data and results were graded as positive (+), doubtful (?), or poor (-). As several studies investigated the same tool, the different studies were synthesized using the rating achieved by most of the articles.

The methodological quality of the articles reporting on the measurement properties of outcome measures was rated on a 4-point scale according to the COSMIN checklist. This checklist is used to assess whether a study on a specific outcome measure tool meets the standards for good methodological quality. A score is calculated for each of nine standards (COSMIN boxes A-I) which somewhat differ from those criteria of Terwee: A. internal consistency, B. reliability,
C. measurement error, D. content validity, E. structural validity, F. hypotheses testing, G. cross-cultural validity, H. criterion validity, and I. responsiveness. There are two additional boxes given; the generalizability and interpretability box. The corresponding 15 items are intended to be used as data extraction forms to extract all data on study characteristics and interpretability issues (e.g., norm scores, floor/ceiling effects, and relevancy for subgroups). An assessments of the statistical methods used in articles based on the Item Response Theory (IRT) (box general requirements for studies that applied IRT Models) was not performed, as this procedure was not used in any of the included studies. Each standard (box) included various items (number ranging from 5 to 18 per box). An overall quality score for that standard was obtained by taking the worst rating of any item (worst score counts principle). The resulting rating could be excellent, good, fair, or poor. There is no formal interpretation of how to combine the measurement property scores (Terwee et al.’s checklist) and the methodological quality scores of studies according to the COSMIN checklist. The COSMIN group stated that the quality of an instrument under investigation remains unclear if the methodological quality of a study is inadequate. For that reason, in the present study, we considered the measurement properties of a tool to be equivocal if the methodological quality of the related studies was rated as poor, irrespective of its rating on the Terwee scale.

RESULTS

Step 1: Literature search for TMC OA studies

Our initial search identified 2979 articles. After removing duplicates, checking references and the two-phase review process, we finally included 316 articles (Figure 1, references in Appendix 2) investigating 13,231 patients (Table 1). Forty-five articles from 17 different research groups reported on patients who had also been subjects in other studies included in our review. Four articles reported on 273 patients affected by hand OA, but the precise number of patients with TMC OA could not be determined. Different surgical procedures were investigated in 268 articles, while conservative treatments were studied in 66 papers. The methodological quality of most of the articles was low: 244 were level IV studies.

In total, we identified 101 different outcome measures, not counting 22 ways to examine radiographs and the self-developed instruments which were excluded.

These 101 outcome measures addressed the OMERACT domain ‘pain’ in 298 articles, ‘physical function’ in 303, ‘global assessment’ in 187, ‘imaging’ in 213 and ‘QoL’ in 13 (Table 2). A visual analogue scale (VAS) was most often applied (n = 93) in the domain ‘pain’. ‘Physical function’ most frequently included measurement of muscle strength and ROM. Grip strength (n = 218) was the most commonly assessed measure of strength, often using a dynamometer (n = 122). Thumb ROM was most often based on abduction (n = 179), in most cases not stating the method used to measure it (n = 114) but sometimes mentioning use of a goniometer (n = 26).

‘Global assessment’ was done primarily by evaluating treatment satisfaction (n = 160), using nine different tools. ‘Imaging’ consisted mainly of rating the stage of OA on the radiographs (n = 160), most frequently using the Eaton classification (n = 132). The Colville questionnaire was used to evaluate ‘QoL’ in five of the 13 articles investigating this dimension.

Twenty-one different standardized questionnaires were used; the DASH was the most common, having been applied in 46 articles.
Step 2: Measurement properties

The second literature search yielded 538 articles, of which we included 12 in the final analysis (Figure 2, Table 3).

These articles examined the measurement properties of 12 outcome measures specifically in patients with TMC OA (Table 4). The DASH and the PRWE were the ones most extensively studied. None of the studies examined all eight measurement properties. Positive ratings (+) were seen for the DASH, quickDASH, Australian / Canadian Osteoarthritis Hand Index (AUSCAN), and Nelson Score. In contrast, the Eaton classification, CMC grind test, and
### Table 1: Characteristics of the 316 clinical studies on TMC OA. Due to inadequate descriptions, not all variables could be extracted from all studies.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Sum of all studies (% of all articles)</th>
<th>Median (range) per study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
<td>2000 (1968-2010)</td>
<td></td>
</tr>
<tr>
<td>Patients</td>
<td>13 231</td>
<td>32 (10-315)</td>
</tr>
<tr>
<td>Females (%) of population</td>
<td>8855 (83.2)</td>
<td>26 (0-162)</td>
</tr>
<tr>
<td>Males (%) of population</td>
<td>1784 (16.8)</td>
<td>5 (0-38)</td>
</tr>
<tr>
<td>Hands</td>
<td>12 521</td>
<td>34 (0-315)</td>
</tr>
<tr>
<td>Age (years)</td>
<td>59.1 (33.7-74.6)</td>
<td></td>
</tr>
<tr>
<td>Follow up (years)</td>
<td>2.9 (0.04-16.4)</td>
<td></td>
</tr>
<tr>
<td>Level of evidence (%)</td>
<td>4 (1-4)</td>
<td></td>
</tr>
<tr>
<td>Level I</td>
<td>33 (10)</td>
<td></td>
</tr>
<tr>
<td>Level II</td>
<td>13 (4)</td>
<td></td>
</tr>
<tr>
<td>Level III</td>
<td>26 (8)</td>
<td></td>
</tr>
<tr>
<td>Level IV</td>
<td>244 (77)</td>
<td></td>
</tr>
<tr>
<td>Intervention</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Implant arthroplasty</td>
<td>92</td>
<td></td>
</tr>
<tr>
<td>Trapeziectomy + ligament reconstruction + tendon interposition</td>
<td>67</td>
<td></td>
</tr>
<tr>
<td>Trapeziectomy + tendon interposition</td>
<td>49</td>
<td></td>
</tr>
<tr>
<td>Trapeziectomy</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>Arthrodesy</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>Injection</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>Splint</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>Trapeziectomy + interposition with various material</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Various surgical interventions^h</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>Trapeziectomy + ligament reconstruction</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Various conservative treatments^i</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Osteotomy</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Physical/occupational therapy</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Drugs</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Unspecified conservative treatments</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

^a taken from 315 articles  
^b taken from 259 articles  
^c taken from 270 articles  
^d taken from 273 articles  
^e taken from 287 articles  
^f due to rounding errors, the sum of the percentages may be less than 100%  
^g more than one intervention per study possible  
^h including unspecified surgical interventions, different surgical interventions in one study group, tendon interposition without trapeziectomy, debridement, synovectomy or denervation  
^i including laser therapy, iontophoresis, radiation therapy, leech therapy, nettle sting, acupuncture, phonophoresis
Table 2 Concepts and outcome measures used in 316 articles about TMC OA categorized according to the OMERACT core set. The OMERACT domain is given in capital letters. Furthermore, the outcome measures are arranged according to whether they are specific for the hand / upper extremity or if they are generic outcome measures.

<table>
<thead>
<tr>
<th>OMERACT domain and outcome measure</th>
<th>Articles (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PAIN</strong></td>
<td>298*</td>
</tr>
<tr>
<td>Hand specific</td>
<td></td>
</tr>
<tr>
<td>Visual Analogue Scale (VAS)</td>
<td>93</td>
</tr>
<tr>
<td>Likert scale(s)</td>
<td>48</td>
</tr>
<tr>
<td>Joint tenderness</td>
<td>23</td>
</tr>
<tr>
<td>Carpometacarpal grind test</td>
<td>14</td>
</tr>
<tr>
<td>Alnot classification</td>
<td>5</td>
</tr>
<tr>
<td>Self-developed questionnaire for hand pain</td>
<td>111</td>
</tr>
<tr>
<td>Generic</td>
<td></td>
</tr>
<tr>
<td>Intake of analgesics</td>
<td>27</td>
</tr>
<tr>
<td>McGill Pain Questionnaire</td>
<td>1</td>
</tr>
<tr>
<td><strong>PHYSICAL FUNCTION</strong></td>
<td>303*</td>
</tr>
<tr>
<td>Hand specific</td>
<td></td>
</tr>
<tr>
<td>Strength</td>
<td>267</td>
</tr>
<tr>
<td>Range of motion</td>
<td>223</td>
</tr>
<tr>
<td>Stability</td>
<td>42</td>
</tr>
<tr>
<td>Dexterity</td>
<td>30</td>
</tr>
<tr>
<td>Sensibility</td>
<td>25</td>
</tr>
<tr>
<td>Subjective hand function</td>
<td>24</td>
</tr>
<tr>
<td>Stiffness</td>
<td>19</td>
</tr>
<tr>
<td>Wound healing</td>
<td>9</td>
</tr>
<tr>
<td>Self-developed function tests</td>
<td>9</td>
</tr>
<tr>
<td>Pegboard tests</td>
<td>8</td>
</tr>
<tr>
<td>Functional Index of Hand OA (FIHOA) / Dreiser index</td>
<td>5</td>
</tr>
<tr>
<td>Jebsen-Taylor Test</td>
<td>4</td>
</tr>
<tr>
<td>Muscle outline</td>
<td>2</td>
</tr>
<tr>
<td>Cochin Scale</td>
<td>2</td>
</tr>
<tr>
<td>Sollerman Hand Function Test</td>
<td>2</td>
</tr>
<tr>
<td>Green Test</td>
<td>1</td>
</tr>
<tr>
<td>Upper extremity specific</td>
<td></td>
</tr>
<tr>
<td>Activities of daily living (ADLs) - Self-developed questionnaire</td>
<td>75</td>
</tr>
<tr>
<td>Activities of daily living (ADLs) - Method not specified</td>
<td>25</td>
</tr>
<tr>
<td>Hand Functional Index (HFI) of the Keitel Functional Test (KFT)</td>
<td>2</td>
</tr>
<tr>
<td>Abilhand</td>
<td>1</td>
</tr>
<tr>
<td>Generic</td>
<td></td>
</tr>
<tr>
<td>Sleep disturbance</td>
<td>2</td>
</tr>
<tr>
<td>Fatigue</td>
<td>1</td>
</tr>
</tbody>
</table>
### Table 2 (Continued)

<table>
<thead>
<tr>
<th>OMERACT domain and outcome measure</th>
<th>Articles (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GLOBAL ASSESSMENT</strong></td>
<td>187*</td>
</tr>
<tr>
<td>Treatment satisfaction</td>
<td>160</td>
</tr>
<tr>
<td>Subjective result</td>
<td>22</td>
</tr>
<tr>
<td>Self-developed questionnaire</td>
<td>4</td>
</tr>
<tr>
<td>Disease activity</td>
<td>1</td>
</tr>
<tr>
<td><strong>IMAGING</strong></td>
<td>213*</td>
</tr>
<tr>
<td>Stage of thumb OA (radiographs)</td>
<td>160</td>
</tr>
<tr>
<td>Scapho-metacarpal distance</td>
<td>104</td>
</tr>
<tr>
<td><strong>QUALITY OF LIFE</strong></td>
<td>13*</td>
</tr>
<tr>
<td>Colville Questionnaire</td>
<td>5</td>
</tr>
<tr>
<td>Arthritis Impact Measurement 2 (AIMS2)</td>
<td>3</td>
</tr>
<tr>
<td>Medical Outcomes Study Short Form 36 (SF-36)</td>
<td>4</td>
</tr>
<tr>
<td>Medical Outcomes Study Short Form 12 (SF12)</td>
<td>1</td>
</tr>
<tr>
<td><strong>PAIN + FUNCTION</strong></td>
<td>291*</td>
</tr>
<tr>
<td>Hand specific</td>
<td></td>
</tr>
<tr>
<td>Australian/Canadian Osteoarthritis Hand Index (AUSCAN)</td>
<td>5</td>
</tr>
<tr>
<td>Patient-Rated Wrist Evaluation (PRWE)</td>
<td>4</td>
</tr>
<tr>
<td>Sequential Occupational Dexterity Assessment (SODA)</td>
<td>3</td>
</tr>
<tr>
<td>Nelson Score</td>
<td>1</td>
</tr>
<tr>
<td>Upper extremity specific</td>
<td></td>
</tr>
<tr>
<td>Disabilities of the arm, shoulder and hand questionnaire (DASH)</td>
<td>46</td>
</tr>
<tr>
<td>QuickDASH</td>
<td>3</td>
</tr>
<tr>
<td>Generic</td>
<td></td>
</tr>
<tr>
<td>Health Assessment Questionnaire - Disability Index (HAQ-DI)</td>
<td>3</td>
</tr>
<tr>
<td>Western Ontario McMaster Universities Osteoarthritis Index (WOMAC)</td>
<td>1</td>
</tr>
<tr>
<td><strong>PAIN + FUNCTION + GLOBAL</strong></td>
<td>173*</td>
</tr>
<tr>
<td>Hand specific</td>
<td></td>
</tr>
<tr>
<td>Buck Gramko Scale</td>
<td>10</td>
</tr>
<tr>
<td>Michigan Hand Questionnaire (MHQ)</td>
<td>4</td>
</tr>
<tr>
<td>Patient Evaluation Measure (PEM)</td>
<td>1</td>
</tr>
<tr>
<td>OMERACT-OARSI Response Index</td>
<td>1</td>
</tr>
<tr>
<td><strong>OMERACT NOT ASSIGNABLE</strong></td>
<td></td>
</tr>
<tr>
<td>Hand specific</td>
<td></td>
</tr>
<tr>
<td>Hand appearance</td>
<td>43</td>
</tr>
<tr>
<td>Crepitus</td>
<td>24</td>
</tr>
<tr>
<td>Thumb shortening</td>
<td>10</td>
</tr>
<tr>
<td>Confidence with hand use</td>
<td>1</td>
</tr>
</tbody>
</table>
Table 2 (Continued)

<table>
<thead>
<tr>
<th>OMERACT domain and outcome measure</th>
<th>Articles (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Generic</strong></td>
<td></td>
</tr>
<tr>
<td>Complications</td>
<td>234</td>
</tr>
<tr>
<td>Return to work</td>
<td>69</td>
</tr>
<tr>
<td>Comfort with device</td>
<td>6</td>
</tr>
<tr>
<td>Laboratory results</td>
<td>2</td>
</tr>
<tr>
<td>Met expectations</td>
<td>2</td>
</tr>
<tr>
<td>Intake of hormones</td>
<td>2</td>
</tr>
<tr>
<td>Pain Anxiety Symptoms Scale (PASS)</td>
<td>1</td>
</tr>
<tr>
<td>Center for the Epidemiological Study of Depression instrument (CES-D)</td>
<td>1</td>
</tr>
<tr>
<td>Pain Catastrophizing Scale (PCS)</td>
<td>1</td>
</tr>
</tbody>
</table>

* Number of articles covering this domain. Although pain, for example, might be evaluated by more than one outcome measure, this value does not necessarily reflect the sum of the instruments given below.

Hand Functional Index of the Keitel Functional Test (HFI / KFT)* rated poorly. Ratings for the PRWE6, 29, 31 and SF-366, 29-31 were equivocal.

The methodological quality of these articles, rated according to the COSMIN checklist was generally fair to poor and most of the measurement properties have not been investigated (Figure 3, Table 5). The positive results of the DASH were weakened by the poor methodological quality of the studies investigating its responsiveness32, while the overall quality of the study considering the Nelson score was also rated as poor33.

**DISCUSSION**

In our review of the outcome measures used in TMC OA studies, we identified 316 papers. We found a wide variety of outcome measures, with pain and function being the most frequent and QoL underrepresented. Studies rarely examined the measurement properties of outcome measures specifically for patients with TMC OA, and the methodological quality of those that did so was fair, so that no recommendations for the use of any outcome measure can be made.

The heterogeneity of the outcome measures employed raises serious issues about the statistical comparison of different interventions, as shown in a recent systematic review of the surgical management of TMC OA2. This concerns not only studies on patients with TMC OA but also studies on hand OA, where many different outcome measures have also been used40. The finding that numerous tools (some self-developed) were used to assess the effectiveness of treatment highlights the need to develop homogeneous, standardized, and validated outcome measures for patients with TMC OA, in order to facilitate comparisons of patient populations and the outcomes of different surgical and non-surgical procedures.

Apart from the variety, we also found that specific aspects of outcome were not covered equally. The OMERACT core set includes the assessment of QoL as a strongly recommended module11, but only few studies on TMC OA include it. Given that hand OA greatly affects the
quality of life\textsuperscript{40}, several authors recommend using a generic outcome measure such as the SF-36 to evaluate QoL in patients with hand disorders\textsuperscript{6, 40-42}. The observed predominance of objective measures (such as muscle strength and ROM) performed by healthcare providers shows that many researchers still do not make the subjective patient perspective their primary focus. This implies underrepresentation of concepts such as psychological consequences, aesthetic changes, and effects on leisure activities, which are important to patients with hand OA\textsuperscript{43}.

The measurement properties of the DASH and PRWE were the most extensively examined ones in patients with TMC OA. Overall, the DASH was rated more favourably than the PRWE, especially

Figure 2  Study selection process for step 2
regarding responsiveness, and floor and ceiling effects. It should be noted, however, that the methodological quality of the studies on the responsiveness of DASH was low. If the study methodology is of poor quality, the quality of the instrument remains equivocal. Furthermore, the specificity and sensitivity of this tool in these particular patients remains questionable because the score is influenced by function/dysfunction of the elbow and shoulder. For this reason, it might be better to use a hand-specific questionnaire such as the AUSCAN or Nelson score. The AUSCAN has only been examined for construct validity in patients with TMC OA, which does not permit any firm conclusions on its overall value in this patient group. The reliability and responsiveness of the AUSCAN were, however, found to be satisfactory for patients with general hand OA. Apart from its measurement properties, other characteristics of a questionnaire such as feasibility and associated costs have to be considered. While the DASH is freely available, the AUSCAN has to be purchased. The Nelson Score, a questionnaire specifically designed to assess the outcome following TMC OA surgery, has so far only been applied by the developers themselves. Interpretation of their findings is further hampered by the poor methodological quality of the study including assessing only 36 patients. The Eaton classification to assess the stage of OA is the only imaging method that has been studied for reliability in patients with TMC OA. Although its reliability remains questionable, it seems to be the best method of staging currently available. The patient’s global assessment was done primarily by evaluating patient satisfaction. Researchers used several instruments, such as a VAS, Likert scale is, and different questionnaires, all of which still have to be tested for their measurement properties in patients with TMC OA. To date, there is no validated instrument available in hand surgery to measure patient satisfaction, which might be due to the numerous health-related, personal, and environmental factors influencing patient satisfaction. The present review yielded equivocal ratings regarding construct validity and floor effect for the SF-36 with respect to QoL, and its responsiveness has not been investigated for TMC OA patients. Although other researchers have found a relatively low sensitivity to change in patients with carpal tunnel syndrome and distal radius fractures, a generic instrument to
### Table 3: Study characteristics of the 12 included articles about measurement properties of outcome measures for patients with TMC OA according to the COSMIN generalizability and interpretability boxes.

<table>
<thead>
<tr>
<th>Study</th>
<th>Outcome measure</th>
<th>No. of patients with TMC OA</th>
<th>Age, years (mean, SD)</th>
<th>Sex (% females)</th>
</tr>
</thead>
<tbody>
<tr>
<td>John et al., 2008*</td>
<td>PRWE, DASH, SF-36</td>
<td>103 (112 cases)</td>
<td>68 ± 9.8</td>
<td>83</td>
</tr>
<tr>
<td>Angst et al., 2009*</td>
<td>quickDASH, DASH, SF-36</td>
<td>103 (112 cases)</td>
<td>68 ± 9.8</td>
<td>83</td>
</tr>
<tr>
<td>Angst et al., 2005*</td>
<td>PRWE, DASH, SF-36, HFI/KFT</td>
<td>103 (112 cases)</td>
<td>68 ± 9.8</td>
<td>83</td>
</tr>
<tr>
<td>MacDermid et al., 2007</td>
<td>PRWE, DASH, AUSCAN, SF-36</td>
<td>120</td>
<td>65 ± 8.1</td>
<td>82</td>
</tr>
<tr>
<td>De Smet, 2004</td>
<td>DASH</td>
<td>15</td>
<td>56 (median)</td>
<td>93</td>
</tr>
<tr>
<td>Niekel et al., 2009</td>
<td>DASH, quickDASH, CES-D, PCS, PASS</td>
<td>107</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Citron et al., 2007</td>
<td>Nelson Score</td>
<td>36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dela Rosa et al., 2004</td>
<td>Eaton classification</td>
<td>30 (40 cases)</td>
<td>59</td>
<td>87</td>
</tr>
<tr>
<td>Kubik III and Lubahn, 2002</td>
<td>Eaton classification</td>
<td>40</td>
<td>60</td>
<td>83</td>
</tr>
<tr>
<td>Hansen et al., 2012</td>
<td>Eaton classification</td>
<td>43 (50 cases)</td>
<td>60</td>
<td>72</td>
</tr>
<tr>
<td>Spaans et al., 2011</td>
<td>Eaton classification</td>
<td>40 cases</td>
<td>60</td>
<td>73</td>
</tr>
<tr>
<td>Merritt et al., 2010</td>
<td>CMC grind test</td>
<td>54 (70 cases)</td>
<td>60 ± 13.4</td>
<td>85</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Study</th>
<th>Outcome measures</th>
<th>Missing items</th>
<th>Handling of missing items</th>
<th>Distribution of the scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>John et al., 2008*</td>
<td>PRWE, DASH, SF-36</td>
<td>PRWE ≤ 12%</td>
<td>PRWE: ≤ 12%</td>
<td>PRWE: left skewed</td>
</tr>
<tr>
<td>Angst et al., 2009*</td>
<td>quickDASH, DASH, SF-36</td>
<td>PRWE ≤ 14%; DASH ≤ 17%; SF-36 ≤ 6%; HFI/KFT = 7%</td>
<td>Non-parametric</td>
<td></td>
</tr>
<tr>
<td>Angst et al., 2005*</td>
<td>PRWE, DASH, SF-36, HFI/KFT</td>
<td>PRWE ≤ 14%; DASH ≤ 17%; SF-36 ≤ 6%; HFI/KFT = 7%</td>
<td>Non-parametric</td>
<td></td>
</tr>
<tr>
<td>MacDermid et al., 2007</td>
<td>PRWE, DASH, AUSCAN, SF-36</td>
<td>SF-36: normal distribution; other instruments: non-parametric</td>
<td></td>
<td></td>
</tr>
<tr>
<td>De Smet, 2004</td>
<td>DASH</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Niekel et al., 2009</td>
<td>DASH, quickDASH, CES-D, PCS, PASS</td>
<td>Non-parametric</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Citron et al., 2007</td>
<td>Nelson Score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dela Rosa et al., 2004</td>
<td>Eaton classification</td>
<td>N/A</td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>Kubik III and Lubahn, 2002</td>
<td>Eaton classification</td>
<td>N/A</td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>Hansen et al., 2012</td>
<td>Eaton classification</td>
<td>N/A</td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>Spaans et al., 2011</td>
<td>Eaton classification</td>
<td>N/A</td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>Merritt et al., 2010</td>
<td>CMC grind test</td>
<td>N/A</td>
<td></td>
<td>N/A</td>
</tr>
</tbody>
</table>

* These articles report on the same cohort

AUSCAN = Australian/Canadian Osteoarthritis Hand Index; CES-D: Center for the Epidemiological Study of Depression instrument; DASH = Disabilities of the Arm, Shoulder and Hand questionnaire;
Table 3

Study characteristics of the 12 included articles about measurement properties of outcome measures for patients with TMC OA according to the COSMIN generalizability and interpretability boxes.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Setting</th>
<th>Countries</th>
<th>Language</th>
<th>Patient selection</th>
<th>Response rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resection Interposition Arthroplasty</td>
<td>Hospital</td>
<td>Switzerland</td>
<td>German</td>
<td>Consecutive</td>
<td>72</td>
</tr>
<tr>
<td>Resection Interposition Arthroplasty</td>
<td>Hospital</td>
<td>Switzerland</td>
<td>German</td>
<td>Consecutive</td>
<td>72</td>
</tr>
<tr>
<td>Resection Interposition Arthroplasty</td>
<td>Hospital</td>
<td>Switzerland</td>
<td>German</td>
<td>Consecutive</td>
<td>72</td>
</tr>
<tr>
<td>Resection Interposition Arthroplasty</td>
<td>Hospital</td>
<td>Canada</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Surgery</td>
<td>Hospital</td>
<td>Belgium</td>
<td></td>
<td>Consecutive</td>
<td>27</td>
</tr>
<tr>
<td>None described</td>
<td>Hospital</td>
<td>The Netherlands</td>
<td></td>
<td>Convenience</td>
<td></td>
</tr>
<tr>
<td>Surgery</td>
<td>Outpatient clinic</td>
<td>UK</td>
<td>English</td>
<td>Consecutive</td>
<td></td>
</tr>
<tr>
<td>None described</td>
<td>Hospital</td>
<td>USA</td>
<td>N/A</td>
<td>Random</td>
<td>N/A</td>
</tr>
<tr>
<td>None described</td>
<td>Hospital</td>
<td>USA</td>
<td>N/A</td>
<td>Random</td>
<td>N/A</td>
</tr>
<tr>
<td>Pre-operative analysis</td>
<td>Hospital</td>
<td>Denmark</td>
<td>N/A</td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>Various</td>
<td>Hospital</td>
<td>The Netherlands</td>
<td>N/A</td>
<td>Convenience</td>
<td>N/A</td>
</tr>
<tr>
<td>None</td>
<td>Private orthopaedic clinic, occupational medicine clinic, general community</td>
<td>USA</td>
<td>N/A</td>
<td>Convenience</td>
<td>N/A</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Floor effect</th>
<th>Ceiling effect</th>
<th>Scores for relevant (sub) groups</th>
<th>MIC or MID</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRWE: 16-24%</td>
<td>PRWE ≥ 24%; DASH = 0%; SF-36 = 16%; HFI/KFT = 0%</td>
<td>quickDASH/ DASH: 4.5-11.3</td>
<td></td>
</tr>
<tr>
<td>PRWE = 0%; DASH = 20%; SF-36 = 83%; HFI/KFT = 16%</td>
<td>Norm data given for SF-36 and DASH</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Existent for PRWE, DASH, and AUSCAN</td>
<td>Scores for patients with solely hand OA compared to patients with hand OA and OA at other joints</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not existent</td>
<td>Scores for other hand disorders included</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

HFI/KFT: Hand Functional Index of the Keitel Functional Test; N/A: not applicable; PASS: Pain Anxiety Symptoms Scale; PCS: Pain Catastrophizing Scale; PRWE: Patient-Rated Wrist Evaluation; SF-36: Medical Outcomes Study Short Form 36
Table 4 Rating of the measurement properties of outcome measures tested for patients with TMC OA according to Terwee et al. They were graded as positive (+), doubtful (?), or poor (-). Blank boxes show that these properties have not been investigated.

<table>
<thead>
<tr>
<th>Outcome measure instrument</th>
<th>Internal Consistency</th>
<th>Content validity</th>
<th>Criterion validity</th>
<th>Construct validity</th>
<th>Reproducibility</th>
<th>Responsiveness</th>
<th>Floor or ceiling effect</th>
<th>Interpretability</th>
</tr>
</thead>
<tbody>
<tr>
<td>DASH</td>
<td>+ 29</td>
<td>+ 30</td>
<td>+ 6.28</td>
<td>+ 31.32</td>
<td>+ 6</td>
<td>+ 6</td>
<td>? 30</td>
<td></td>
</tr>
<tr>
<td>AUSCAN</td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td>?</td>
<td></td>
</tr>
<tr>
<td>PRWE</td>
<td>+ 28</td>
<td>+ 28</td>
<td>- 28</td>
<td>- 6.28</td>
<td>- 6.28</td>
<td>? 30</td>
<td>?</td>
<td></td>
</tr>
<tr>
<td>Nelson Score</td>
<td>+</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SF-36</td>
<td>+ 29</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>- 6</td>
<td>- 6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eaton classification (stage of OA)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carpometacarpal grind test</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>quickDASH</td>
<td>+ 29</td>
<td>+ 20.38</td>
<td>?</td>
<td>?</td>
<td>2.3438</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CES-D</td>
<td>?</td>
<td>?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCS</td>
<td>?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PASS</td>
<td>?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HF/KFT</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

AUSCAN = Australian/Canadian Osteoarthritis Hand Index; CES-D: Center for the Epidemiological Study of Depression instrument; DASH = Disabilities of the Arm, Shoulder and Hand questionnaire; HF/KFT: Hand Functional Index of the Keitel Functional Test; PASS: Pain Anxiety Symptoms Scale; PCS: Pain Catastrophizing Scale; PRWE: Patient-Rated Wrist Evaluation; SF-36: Medical Outcomes Study Short Form 36
Table 5: Methodological quality of studies investigating outcome measure instruments in patients with TMC OA. In accordance with the COSMIN checklist, the method of investigating each measurement property was rated excellent, good, fair, or poor.

<table>
<thead>
<tr>
<th>Author</th>
<th>Outcome measure</th>
<th>Internal consistency</th>
<th>Reliability</th>
<th>Measurement error</th>
<th>Content validity</th>
<th>Structural validity</th>
<th>Hypotheses testing</th>
<th>Cross-cultural validity</th>
<th>Criterion validity</th>
<th>Responsiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td>John et al., 2008*</td>
<td>PRWE, DASH, SF-36</td>
<td>Poor</td>
<td>Fair</td>
<td>Fair</td>
<td>Fair</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td></td>
</tr>
<tr>
<td>Angst et al., 2009*</td>
<td>quickDASH, DASH, SF-36</td>
<td>Fair</td>
<td></td>
<td></td>
<td>Fair</td>
<td>Fair</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Angst et al., 2005*</td>
<td>PRWE, DASH, SF-36, HFI/KFT</td>
<td>Poor</td>
<td>Fair</td>
<td></td>
<td>Poor</td>
<td>Fair</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MacDermid et al., 2007*</td>
<td>PRWE, DASH, AUSCAN, SF-36</td>
<td>Fair</td>
<td></td>
<td></td>
<td>Fair</td>
<td>Fair</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>De Smet, 2004</td>
<td>DASH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Niekel et al., 2009</td>
<td>DASH, quickDASH, CES-D, PCS, PASS</td>
<td>Fair</td>
<td></td>
<td></td>
<td>Poor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Poor</td>
</tr>
<tr>
<td>Citron et al., 2007</td>
<td>Nelson Score, DASH</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td></td>
</tr>
<tr>
<td>Dela Rosa et al., 2004</td>
<td>Eaton classification</td>
<td></td>
<td></td>
<td></td>
<td>Fair</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kubik III and Lubahn, 2002*</td>
<td>Eaton classification</td>
<td></td>
<td></td>
<td></td>
<td>Fair</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hansen et al., 2012</td>
<td>Eaton classification</td>
<td></td>
<td></td>
<td></td>
<td>Good</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanns et al., 2011</td>
<td>Eaton classification</td>
<td></td>
<td></td>
<td></td>
<td>Fair</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Merritt et al., 2010</td>
<td>CMC grind test</td>
<td></td>
<td></td>
<td></td>
<td>Fair</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Fair</td>
</tr>
</tbody>
</table>

*These articles report on the same cohort.

AUSCAN = Australian/Canadian Osteoarthritis Hand Index; CES-D: Center for the Epidemiological Study of Depression instrument; DASH = Disabilities of the Arm, Shoulder and Hand questionnaire; HFI/KFT: Hand Functional Index of the Keitel Functional Test; PASS: Pain Anxiety Symptoms Scale; PCS: Pain Catastrophizing Scale; PRWE: Patient-Rated Wrist Evaluation; SF-36: Medical Outcomes Study Short Form 36.
measure QoL is recommended because it allows the comparison between different conditions and patient populations\textsuperscript{12}.

Assessing the methodological quality of studies is an important point in systematic reviews. However, there are no uniform guidelines, for how to assess the methodological quality for different types of studies. For randomized controlled trials (RCTs), the Cochrane collaboration recommends its risk of bias tool\textsuperscript{13}. For observational studies, there are various checklists and scores available, but none of these can be recommended to be used as a gold-standard\textsuperscript{14}. Other common checklists, such as the CONSORT\textsuperscript{15}, the PRISMA\textsuperscript{22, 23}, and Strengthening the Reporting of Observational Studies in Epidemiology (STROBE)\textsuperscript{16} statements, are not intended to serve as quality appraisal tools but to guide authors when reporting RCTs, systematic reviews and observational studies, respectively. For grading the methodological quality of studies investigating measurement properties, the COSMIN checklist\textsuperscript{19} is the only available tool, so far.

Our review has certain limitations. As only English and German articles have been included, some studies published in other languages might have been missed. Additionally, many articles lacked information on the study population and methods, making it impossible to determine actual overlap among studies and calculate the exact numbers of patients investigated. Furthermore, the low methodological quality of all the studies, assessed by the COSMIN checklist, prohibits recommendations. The scoring of this tool is rather rigid, giving the overall rating of a specific measurement property as poor even if only one item is scored as such. For each measurement property, the number of missing items and their handling has to be scored. Though this information is lacking in most of the studies, this leads to an overall fair rating, although the study achieved better ratings regarding the other items of that property. For this reason, the methodological quality of the articles might have been underestimated. Another limitation of the study is that we used the Wright classification for rating the levels of evidence. Following our rating, a revised classification for evidence-based medicine was published\textsuperscript{57}. As the primary purpose of our publication was not to report the evidence levels of studies on TMC OA but rather to focus on measurement instruments, it was decided not to repeat the classification.

Based on the results of the present study, no recommendation for a particular outcome measure can be made. A combination of hand-specific questionnaires, which are most suitable for detecting changes in patients with TMC OA, general health status questionnaires, and clinical data are suggested. However, more research on the psychometric properties of outcome measures in methodological sound studies is needed before we can make any firm recommendations about the use of specific tools.
ACKNOWLEDGEMENTS
We thank Stefanie Hensler, Martina Wehrli, and Carina Muoth for assistance in reviewing the articles, Prof Dr Nelissen for his scientific support, and Dr Meryl Clarke for her assistance in preparing the manuscript.

CONFLICT OF INTERESTS
None declared.

FUNDING
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.
REFERENCES


reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ 2009; 339:b2700.


44. MacDermid JC, Tottenham V. Responsiveness of the disability of the arm, shoulder, and hand (DASH) and patient-rated wrist/hand evaluation (PRWHE) in evaluating change after hand therapy. J Hand Ther 2004; 17:18-23.


APPENDIX 1: DEFINITIONS OF THE PSYCHOMETRIC PROPERTIES ACCORDING TO TERWEE ET AL.¹ AND THEIR RATINGS (SLIGHTLY MODIFIED)

Internal consistency
Definition: Internal consistency is a measure of the extent to which items in a questionnaire (sub)scale are correlated (homogeneous), thus measuring the same concept. Internal consistency is an important measurement property for questionnaires that intend to measure a single underlying concept (construct) by using multiple items.

Positive (+) rating: A positive rating was assigned if a factor analysis was performed on adequate sample size (7 × number of items and ≥ 100) or if Cronbach’s alpha was calculated per subscale and was between 0.70 and 0.95.

Doubtful (?) rating: No factor analysis performed, or doubtful design or method.

Poor (-) rating: Cronbach’s alpha was < 0.70 or > 0.95, despite adequate design and method.

Content validity
Definition: Content validity examines the extent to which the concepts of interest are comprehensively represented by the items in the questionnaire.

Positive (+) rating: A positive rating was assigned if a clear description was provided of the measurement aim, the target population, the concepts that were being measured, the item selection and target population, and if the investigators or experts were involved in item selection.

Doubtful (?) rating: A clear description of above-mentioned aspects was lacking or only the target population was involved, or a doubtful design or method was used.

Poor (-) rating: The target population was not involved.

Criterion validity
Definition: Criterion validity is the extent to which scores on a particular questionnaire relate to a gold standard. According to Mokking et al.², there is no gold standard for a health related patient reported outcome. Only if a shortened version is compared to its original long version, can it be considered as the gold standard (e.g. the quickDASH versus the full DASH).

Positive (+) rating: A positive rating was assigned if the correlation with a true gold standard was ≥ 0.70.

Doubtful (?) rating: No convincing arguments that the gold standard is really the gold standard, or a doubtful design was used.

Poor (-) rating: The correlation to the gold standard was < 0.7 despite adequate design and methods.

Construct validity
Definition: Construct validity is the extent to which scores on a particular questionnaire relate to other measures in a manner that is consistent with theoretically derived hypotheses concerning the concepts that are being measured.
Outcome measures for TMC OA

Positive (+) rating: A positive rating was assigned if specific hypotheses were formulated and at least 75% of the results were in accordance with these hypotheses. Though the testing of hypotheses is quite a new approach for testing construct validity, a “+” was also assigned if the target outcome measure correlated ≥ 0.7 with another outcome measure evaluating the same construct.

Doubtful (?) rating: Doubtful design or methods.

Poor (-) rating: Less than 75% of the hypotheses were confirmed, despite adequate design and methods or correlation with another outcome measure evaluating the same construct was < 0.7.

Reproducibility
Reproducibility concerns the degree to which repeated measurements in stable persons (test-retest) provide similar answers, and can be divided into agreement and reliability:

Agreement
Definition: Agreement concerns the absolute measurement error, which means how close the scores on repeated measures are, expressed in the unit of the measurement scale at issue. Small measurement error is required for evaluative purposes in which one wants to distinguish clinically important changes from measurement error.

Positive (+) rating: A positive rating was assigned if the minimal important change (MIC) was smaller than the smallest detectable change (SDC), if the MIC was outside the limits of agreement (LOA) or if convincing arguments that agreement is acceptable were given. In addition to this definition by Terwee et al., a “+” was also assigned if the Intraclass correlation coefficient (ICC) or kappa was ≥ 0.7.

Doubtful (?) rating: Doubtful design or methods, or MIC not defined.

Poor (-) rating: The MIC was greater than the SDC or the MIC was inside the LOA, despite adequate design and methods.

Reliability
Definition: Reliability is the extent to which patients can be distinguished from each other, despite measurement errors (relative measurement error).

Positive (+) rating: A positive rating was assigned if the ICC or weighted Kappa was ≥ 0.70.

Doubtful (?) rating: Doubtful design or methods (e.g. time interval not mentioned).

Poor (-) rating: ICC or weighted Kappa was < 0.70 despite adequate design and methods.

Responsiveness
Definition: Responsiveness is the ability of a questionnaire to detect clinically important changes over time.

Positive (+) rating: A positive rating was assigned if the SDC had been calculated, if the SDC was smaller than the MIC or the MIC laid outside the LOA, or if the response ratio (RR) was greater than 1.96 or the area under curve (AUC) was greater than 0.7. Though many researchers calculate effect sizes (ES) or standardized response means (SRM), a “+” was assigned if these figures were ≥ 0.8.
Doubtful (?) rating: Doubtful design or methods.

Poor (-) rating: A poor rating was assigned if the SDC was greater than the MIC, if the MIC was equal or laid inside the LOA, if the RR was smaller than 1.96 or the area under curve (AUC) was smaller than 0.7, or if ES or SRMs were smaller than 0.7.

**Floor or ceiling effects**

Definition: Floor or ceiling effects are considered to be present if more than 15% of respondents achieved the lowest or highest possible score, respectively.

Positive (+) rating: A positive rating was assigned if less than 15% of the respondents achieved the highest or lowest possible scores.

Doubtful (?) rating: Doubtful design or methods.

Poor (-) rating: More than 15% of the respondents achieved the lowest or highest possible score.

**Interpretability**

Definition: Interpretability is the degree to which one can assign qualitative meaning to quantitative scores. Investigators should provide information about what (change in) score would be clinically meaningful.

Positive (+) rating: A positive rating was assigned if mean and standard deviation (SD) scores were presented for at least four relevant subgroups of patients, and if MIC was defined.

Doubtful (?) rating: Doubtful design or methods, less than four subgroups, or if MIC was not defined.

**References**


APPENDIX 2: INCLUDED ARTICLES


Brand J, Gaulke R, Geerling J, Mellor R, Krettek C. [Suspension arthroplasty of the thumb carpometacarpal


Outcome measures for TMC OA


Helbig B. Indication and results of the treatment with the francobal endoplastic in case of arthrosis of the carpometacarpal joint of the thumb. [German]. Orthopädische Praxis. 1989;25:318-324.


Vadstrup LS, Schou L, Boeckstyns ME. Basal joint osteoarthritis of the thumb treated with Weilby
OUTCOME MEASURES FOR TMC OA


