The use and reporting of WOMAC in the assessment of the benefit of physical therapies for the pain of osteoarthritis of the knee: findings from a systematic review of clinical trials

Nerys F. Woolacott¹, Mark S. Corbett¹ and Stephen J. C. Rice¹

Abstract

Objective. For the purposes of meta-analysis and network meta-analysis, the use of standard outcome measures is ideal. In OA research, the WOMAC was developed as an OA-specific measure of disability. It includes a pain subscale. In 1994 a consensus meeting recommended the use of WOMAC as a primary measure of efficacy in OA. In the context of a review of the efficacy of physical interventions for the relief of the pain of OA of the knee, we investigated the use of WOMAC.

Methods. A systematic review (December 2009-January 2010) identified trials that used the WOMAC outcome. These were investigated for correct use and clear reporting of the WOMAC pain subscale and the WOMAC index.

Results. The WOMAC pain subscale was used in 45% of the 134 trials. Reporting of the exact method of administering the WOMAC pain subscale was poor in many cases: in 53% of trials the reporting of the type of WOMAC scale used was inadequate; the score range was reported ambiguously in 38% of trials, with a further 10% being completely unclear. Similar less than optimal reporting of the WOMAC index was found.

Conclusion. Poor reporting of both the WOMAC pain subscale and the WOMAC index resulted in significant uncertainty in the interpretation of the results of individual trials and limited their contribution to evidence synthesis. Improved adherence with the standard use of the WOMAC scoring system, with clear reporting of it in trials of OA of the knee should be encouraged.

Key words: WOMAC, pain, osteoarthritis of the knee, assessment, systematic review.

Introduction

For the purposes of meta-analysis and network meta-analysis, the use of standard outcome measures is ideal. In the field of OA research, the WOMAC was developed as an OA-specific measure of disability [1]. In 1994 a consensus meeting recommended the use of WOMAC as a primary measure of efficacy in OA trials [2].

WOMAC is a self-administered health status measure that assesses the dimensions of pain, stiffness and function (either separately or as an overall index) in patients with OA of the hip or knee; it is available in 5-point Likert, 11-point numerical rating and 100-mm visual analogue scale (VAS) formats [3]. Under each dimension there are a number of questions designed to assess the clinical severity of the disease (5 questions for pain, 2 questions for stiffness and 17 questions for physical function). The five pain questions reflect pain experienced on five different activities: the five situations are walking on a flat surface, going up or down stairs, at night while in bed, sitting or lying, and standing upright. The patient’s response to each question produces a score that is then summed to derive an aggregated score for each dimension. It produces three subscale scores (pain, stiffness and physical function) and a total score (WOMAC index) that reflects disability overall.
The WOMAC pain score range is variously reported and includes VAS 0–10 scale (commonly reported as a 0–50 range), VAS 0–100 scale (commonly reported as a 0–500 range), an 11-box numerical rating scale (NRS) (commonly reported as 0–50 range) or a Likert scale (commonly reported as a 0–20 range). The overall WOMAC score (index) is determined by summing the scores across the three dimensions and the score ranges include 0–240 (derived from the VAS 0–10 or NRS scale), or 0–2400 (derived from the VAS 0–100) or 0–96 (derived from a 0–4 Likert scale). A number of various transformations and modifications are reported in the literature.

We conducted a systematic review of physical therapies for the relief of pain associated with OA of the knee, which included only trials that reported pain as an outcome [4]. It was hoped and anticipated that the majority of studies meeting the eligibility criteria in our review would use the WOMAC pain subscale to assess pain, thereby minimizing a source of heterogeneity within the data available for synthesis. During the review it became apparent that there were ambiguities across the literature in the ways in which WOMAC was used and reported, which meant we had to resort to making assumptions when extracting data and using the standardized mean difference (SMDs) to analyse the data. Within the context of investigating the efficacy of physical interventions for the relief of the pain of OA of the knee, this article summarizes the use of the WOMAC pain subscale and the WOMAC index in studies of OA of the knee and makes suggestions for the future reporting of its use and the generated results.

Methods

The systematic review processes and methods of analysis were specified in advance and documented in a protocol, and details are given in a full report [4]. The systematic review was conducted following the general principles recommended in the Centre for Reviews and Dissemination’s (CRD) guidance [5] and the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement [6]. Briefly, we searched 15 electronic databases from inception to June 2010 for randomized controlled trials (RCTs) of named physical therapies in OA of the knee that reported pain as an outcome using any scale. No language or date restrictions were applied.

Bibliographies of all relevant reviews and guidelines were checked for further potentially relevant studies, and internet searches were made of websites relating to OA. Two reviewers independently screened all abstracts, and then all relevant full papers, with disagreements resolved by discussion, or by a third reviewer when necessary. The named physical therapies were acupuncture, balneotherapy, braces, aerobic exercise, muscle strengthening exercise, heat treatment, ice/cooling treatment, insoles, interferential therapy, laser/light therapy, manual therapy, neuromuscular electrical stimulation, pulsed electrical stimulation, pulsed electromagnetic fields, static magnets, Tai Chi, transcutaneous electrical nerve stimulation (TENS) and weight loss.

<table>
<thead>
<tr>
<th>Form of WOMAC used</th>
<th>WOMAC pain score range</th>
<th>WOMAC index score range</th>
</tr>
</thead>
<tbody>
<tr>
<td>VAS 0–10</td>
<td>0–50</td>
<td>0–240</td>
</tr>
<tr>
<td>VAS 0–100</td>
<td>0–500</td>
<td>0–2400</td>
</tr>
<tr>
<td>NRS 0–10</td>
<td>0–50</td>
<td>0–240</td>
</tr>
<tr>
<td>Likert scale (0–4)</td>
<td>0–20</td>
<td>0–96</td>
</tr>
</tbody>
</table>

Data extraction and assessment of trial quality

Data extraction and quality assessments were performed by one reviewer and checked by a second reviewer. Using a standardized data extraction form, data were extracted on population characteristics, intervention parameters, study quality [graded as excellent, good, satisfactory or poor (see supplementary Appendix A, available as supplementary data at Rheumatology Online)] and the pain outcome measure used. For those trials that utilized WOMAC, further details were extracted: which scale was used (Likert/VAS 0–10, VAS 0–100/NRS); whether the WOMAC pain subscale or the WOMAC index was used; and whether any modifications were reported. In the light of inconsistencies and lack of clarity identified during the review, the WOMAC outcome details were re-examined by a third reviewer (N.F.W.) and further information was extracted as necessary to address the following questions:

(i) Was it clear that all 5 (24) assessments had been conducted?
(ii) Was the score range clear?
(iii) Were details reported on how the final score had been calculated (sum or average or transformation to 0–100 scale)?
(iv) Were baseline scores reported (and approximate baseline score)?

In addition, the scale used and the score range that could be deduced from the information provided in the paper was recorded, and the ease of identification of these was categorized as being either clearly stated in the paper (Stated), required assumptions to be made (assumed) or unclear. All information that could support any assumptions, including baseline score (or, where not reported, follow-up scores), was also recorded. The expected range of scores for the WOMAC pain subscale and the WOMAC index are given in Table 1.

Results

A total of 134 original trials formed the basis of the review. Pain was measured using a variety of scales, with the WOMAC pain subscale being reported in 60 (45%) of the trials. The WOMAC index was reported in only 31 (23%) trials. Four trials reported the WOMAC index but not the WOMAC pain scale. The reporting of the...
WOMAC pain subscale and WOMAC index across the trials is summarized in Table 2. The full list of trials, assumptions and results of the quality assessment are given in supplementary Tables S1 and S2, available as supplementary data at Rheumatology Online.

WOMAC pain subscale

Reporting of the exact method used in administering the WOMAC pain subscale scoring was poor in many cases (Table 3). In 21 of 60 trials, deciding whether a Likert scale or VAS was used had to be deduced from other information given in the paper—usually the score range or the actual baseline or follow-up data. In a further 11 cases, reporting was so poor that it was not possible to deduce the scoring method. Thus, in 32 (53%) of 60 papers the reporting of the WOMAC scale used was inadequate.

The number of different score ranges used in the sample of trials was larger than we had expected (Table 3). Variations on the standard WOMAC Likert pain score range (0/20) included 0/10 or 25 derived when the Likert scale is counted from 1 (no pain) to 5 (extreme pain). When such a variation is used without full explanation it can be confusing: for example, a baseline score of 14 would seem very high if it were assumed it was out of a maximum score of 20 (as would be the case from a standard Likert 0–4) rather than 25 [7]. Other misleading assumptions can be made between the scales: for example, a baseline score of 7 is low if out of 20 (Likert scale), but high if actually derived from a VAS or NRS (0–100) and thus out of 10 [8]. Another common variation on the standard score ranges was a 0/100 score range, generated either when a VAS 100 scale is averaged across the five items rather than summed, or when other scales are standardized to a 100 scale [9]. Overall, the score range was reported ambiguously in 23 (38%) of the 60 trials and in a further 6 it was completely unclear, giving a total of 29 (48%) trials with inadequate reporting of WOMAC pain results.

To explore the strength of the assumptions required to determine the score range, we further examined the basis for the assumptions. The results are summarized in Table 4. Although enough information about the scale used or other information (such as use of the WOMAC index scale) helped to support the conclusions reached regarding the score range in a number of studies, almost half of the assumptions (comprising 18% of all trials reporting WOMAC pain subscale) were based on the baseline score alone. Where no information is given other than the baseline value it can be difficult to make the correct deduction: a baseline score of 7 could be from a VAS 10 scale reported as the mean of the five pain assessments or could be from a Likert 0–4 scale out of 20 [10]. Overall, only 15 (25%) trials reported unambiguously both the scale and score range for their use of the WOMAC pain subscale [11–25]. Of these 15, 53% were rated good or satisfactory [11, 15–18, 21–22, 24] compared with 24% of all trials reviewed.

WOMAC index

As for the WOMAC pain subscale, reporting of the WOMAC index was less than optimal in a large proportion...
of studies. In 10 (32%) trials the scale used could only be
deducted with the use of at least one assumption and in a
further 13 (42%) cases it could not be deduced at all. Thus
in 23 of 31 (74%) trials the reporting of the scale used was
inadequate (Table 5).

The number of different score ranges was high. Only 12
trials used the standard 0–40, 0–100 or 0–240 ranges. In
a high proportion of trials (32%) reporting of the score
range was unclear; for most of these (9/10), while it ap-
peared that the score range was either 0–40 or 0–100, it
was not possible to determine which. To explore the
strength of the assumptions required to determine the
score range, we further examined the basis for the as-
sumptions (Table 6). Of the eight trials for which some
assumption had to be made, three had to have the as-
sumption based on baseline value alone (10% of all trials
reporting WOMAC). Only four (13%) trials reported unam-
biguously both the scale and score range for their use of
the WOMAC index [17, 21, 26–27], two of which were
rated satisfactory [17, 21] and two were rated poor quality
studies [26, 27].

**WOMAC function and stiffness subscales**

The WOMAC subscales of stiffness and function were not
included in our review: our primary outcome was pain,
and the broader effect of the interventions on OA of the
knee were to be captured by the WOMAC index. For the
purpose of the analysis presented in this article, it is un-
likely that the results for the function or stiffness subscales
would be reported differently to the pain subscale, and
just one trial that did not report WOMAC pain scale data
or the WOMAC index reported the function or stiffness
subscales [28]. This article reported only the WOMAC
function subscale, and did so clearly reporting that a 5-
point Likert scale had been used normalized to 100 (score
range 0–100).

<table>
<thead>
<tr>
<th>Information basis</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clearly stated in paper</td>
<td>31a (52)</td>
</tr>
<tr>
<td>Uninterpretable score range</td>
<td>2b (3)</td>
</tr>
<tr>
<td>Unclear (insufficient information to permit assumptions)</td>
<td>6b (10)</td>
</tr>
<tr>
<td>Assumption made based on:</td>
<td></td>
</tr>
<tr>
<td>Baseline value and type of scale used</td>
<td>9 (15)</td>
</tr>
<tr>
<td>Baseline value and other information (other than scale used)</td>
<td>3 (6)</td>
</tr>
<tr>
<td>Baseline value alone</td>
<td>11 (18)</td>
</tr>
</tbody>
</table>

Data given as number of trials (total n = 60).

aIncludes one stated score range that was not interpretable. bTable totals 62 because it includes two
stated score ranges that were not interpretable.

<table>
<thead>
<tr>
<th>Table 5 WOMAC index: scales and score ranges used in trials and whether assumptions were required to determine these</th>
</tr>
</thead>
<tbody>
<tr>
<td>No assumption needed</td>
</tr>
<tr>
<td>----------------------</td>
</tr>
<tr>
<td>Scale used</td>
</tr>
<tr>
<td>Likert scale</td>
</tr>
<tr>
<td>VAS</td>
</tr>
<tr>
<td>Unclear</td>
</tr>
<tr>
<td>Score range</td>
</tr>
<tr>
<td>0–4</td>
</tr>
<tr>
<td>0–10</td>
</tr>
<tr>
<td>0–96</td>
</tr>
<tr>
<td>0–100</td>
</tr>
<tr>
<td>0–240</td>
</tr>
<tr>
<td>0–2400</td>
</tr>
<tr>
<td>5–120</td>
</tr>
<tr>
<td>26–130</td>
</tr>
<tr>
<td>Unclear</td>
</tr>
<tr>
<td>Unclear but probably out of 96 or 100</td>
</tr>
</tbody>
</table>

Data given as number of trials (total n = 31).
TABLE 6 Sources of information regarding score ranges for the WOMAC index in the sample of trials

<table>
<thead>
<tr>
<th>Information basis</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clearly stated in paper</td>
<td>13 (42%)</td>
</tr>
<tr>
<td>Uninterpretable score range</td>
<td>2 (10%)</td>
</tr>
<tr>
<td>Unclear (insufficient information to permit assumptions)</td>
<td>1 (3%)</td>
</tr>
<tr>
<td>Unclear (but probably either 0–96 or 0–100 based on baseline score)</td>
<td>9 (29%)</td>
</tr>
<tr>
<td>Assumption made based on:</td>
<td></td>
</tr>
<tr>
<td>Baseline value and scale used</td>
<td>3 (10%)</td>
</tr>
<tr>
<td>Baseline value and other information (other than scale used)</td>
<td>2 (6%)</td>
</tr>
<tr>
<td>Baseline value alone</td>
<td>3 (10%)</td>
</tr>
</tbody>
</table>

Data given as number of trials (total n = 31).

*aIncludes two uninterpretable score ranges. bTable totals 62 because these two counted twice (see the matter given in the table body for footnote a).

Discussion

Despite the fact that WOMAC was identified as a key primary effectiveness measure in OA trials back in 1994, only 45% of the trials identified in our review of physical therapies for the relief of pain due to OA used the WOMAC pain subscale and 23% used the WOMAC index. Furthermore, reporting of the WOMAC pain subscale was less than ideal across the trials with only 25% of these trials clearly reporting the scale used and the score range. Similarly only 13% reported clearly the WOMAC index scale and score range. Clear reporting was associated with better study quality. Using the baseline score and other information allowed the type of WOMAC pain subscale and score range to be deduced for a further 35 and 38% of trials, respectively. However, the type of scale used and the score range could not be deduced at all in 18 and 10%, respectively. Similarly, for the WOMAC index, assumptions had to be used to deduce the scale and scores in 32 and 26% of trials, respectively, but a significant number remained unclear.

The question that can be asked is ‘Does it matter if the reader and researchers using the work do not know which form of WOMAC was used?’ Although the use of the Likert scale scores as simple numeric data is questionable [29] and there is some evidence that VAS scales and Likert scales can generate different results [30], it has been shown that for single items the correlation between VAS and Likert scores is high [31]. A study of patient assessments of the pain of OA reported no differences between VAS scoring and Likert scoring, including WOMAC [32]. However, there is always the possibility that differences between scales could be a source of variation between trials, and thus information on which scale was used should be reported.

The score range is very important for the interpretation of the results of any trial, alone or in the context of other trials. It becomes even more important when the results are to be combined with those from other trials in an evidence synthesis. While in the reporting of a single trial it may seem sufficient to report only the treatment difference and the statistical significance, all trials should be interpreted within the whole body of relevant evidence. Use of assumptions in deducing which score range was used in individual trials can lead to a more time-consuming and subjective data extraction process, and this can lead to an underestimation of the differences (heterogeneity) between trials. For example, a trial that reported a baseline value of WOMAC pain of 16 could be interpreted as 16/20 (0.80), a high level of baseline pain, or may be assumed to be 16/50 (0.32, a more typical baseline pain level). If that latter apparently reasonable assumption was accepted, but the former was in fact the true level of baseline pain, an opportunity to recognize the existence of heterogeneity among the trials would be lost. When pooling data or drawing inferences from a number of trials it is essential that the trials are similar enough for the results to be meaningful; failure to recognize heterogeneity will mean that the interpretation of the results may be incorrect.

Differences between outcomes can be resolved to some extent by either transforming each scale to the same scale (both the treatment effects and the standard deviations), or through the use of SMDs. An SMD is an index that is comparable across studies with different scales and is created by dividing the mean difference in each study by that study’s standard deviations. However, there are problems associated with both approaches. If the method of transforming the scale is adopted but the scale is incorrectly identified, then both the effect estimate and the variance may be wrong, which will inappropriately weight the trial in a meta-analysis. If the SMD approach is adopted, this can introduce limitations for meta-analysis because ‘change from baseline’ and ‘final values’ data cannot be combined when SMDs are used: as many trial results are presented as one or the other type of data and not both, this limitation can result in the loss of many trials from a meta-analysis. In addition, the final result will be in SMD units rather than the original scale and so can be difficult to interpret. Furthermore, if with SMDs an adjustment for baseline values was made by
including the baseline value as a covariate, this would have to be a proportion of the scale rather than the absolute value if different scales were included in the analysis. Obviously the proportion of the scale would be incorrect were the wrong scale identified. Therefore it is crucial that the data are reported explicitly. A lack of clarity regarding the scales used in the original trials can limit exploration and testing of the results: making it impossible to, for example, conduct a sensitivity analysis to investigate the influence of scale type. Use of a validated assessment tool such as the WOMAC should greatly reduce heterogeneity across trials in evidence synthesis, but this can only be achieved if the tool is used correctly and if variations are reported clearly.

Strengths/limitations

Our investigation into the reporting of the WOMAC pain subscale and WOMAC index was based on a thorough systematic review of all physical interventions for the treatment of pain associated with OA of the knee. As such it was an unbiased sample. However, the findings may not reflect trials of other treatments for OA pain, namely pharmaceuticals or surgery. Many of the trials were of poor quality and it may be that the failures in the reporting of WOMAC in our study are generalizable mainly to small pilot or investigator-led studies. It should also be acknowledged that the limited use of the WOMAC index may partly due to the fact that while studies had to report pain to be included in the review, they did not have to report disability.

Lessons to be learned

Our study found that reporting of the methods (and results) in RCTs using the WOMAC assessment tool lacked clarity. The various versions of WOMAC available are clearly defined and have all been validated and offer investigators a helpful range of options, but it is important that investigators and readers are aware of these variations and they are reported clearly. For the reader and researchers a full description would be ideal. A small number of the trials did include such a description. For example, ‘The WOMAC used in this study was the Likert version 3.1 standardized with English for an American population, consisting of 24 self-administrated questions that were answered for each item on a 5-point Likert scale (none, mild, moderate, severe and extreme). It was reported as three separate subscales: pain, physical function, and stiffness. The WOMAC pain subscale had five questions scored 0 to 4 and was considered invalid if more than one item was missing; hence, it had a range of 0 (no pain) to 20 (maximal pain). In the event of a missing item, the remaining four items were averaged and then multiplied by five.’ [13]. Such clear reporting should be encouraged and not be sacrificed in the drive to reduce word count. As an absolute minimum, the type of WOMAC used (Likert/VAS/NRS) and the score range must be reported.

Rheumatology key messages

- Reporting of both the WOMAC pain subscale and the WOMAC index was poor.
- Poor reporting of WOMAC hampered interpretation of trial results and limited the evidence synthesis.
- Adherence to the standard WOMAC scoring system, and its clear reporting, should be encouraged.

Acknowledgements

We would like to thank Melissa Harden, who conducted the literature searches, Vichithranie Madurasinghe, Russell Slack and Anthony Danso Appiah, who assisted with data extraction, and Hugh McPherson, who acquired the funding.

Funding: This article presents independent research funded by the National Institute for Health Research (NIHR) under Programme Grants for Applied Research (Grant no. RP-PG-0707-10186). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. All researchers were independent of the funders. The funders had no role in study design, data collection, data synthesis, data interpretation or writing the report.

Disclosure statement: The authors have declared no conflicts of interest.

Supplementary data

Supplementary data are available at Rheumatology Online.

References


