A Call for Evidence-based Decision Making When Selecting Outcome Measurement Instruments for Summary of Findings Tables in Systematic Reviews: Results from an OMERACT Working Group

Dorcas E. Beaton, Caroline B. Terwee, Jasvinder A. Singh, Gillian A. Hawker, Donald L. Patrick, Laurie B. Burke, Karine Toupin-April, and Peter S. Tugwell

ABSTRACT. Objective. Systematic reviews often struggle with how to combine information when more than 1 instrument is used across studies being synthesized. Different techniques have been suggested based on frequency of use in the literature, or on consensus. We explore an approach blending 2 initiatives: OMERACT (Outcome Measurement in Rheumatology) and COSMIN (Consensus On Selection of Measurement Instruments), and investigate the effects of an evidence-based measurement approach on selection of outcomes.

Methods. Readings were circulated to attendees registered for a preconference workshop on pain measurement. Three instruments were considered and exercises conducted to engage people in the content and measurement performance of these tools. Consensus was sought that an evidence-based approach could be created for selection of instruments for summary of findings (SoF) tables.

Results. The blending of COSMIN and OMERACT approaches led to an evidence-based approach that depended both on a clear definition of target concept and a review of measurement performance of the instrument. Participants emphasized that conceptual clarity and practical considerations should come before measurement property results.

Conclusion. Evidence-based approaches can be adopted for selection of instruments for SoF tables. A research agenda was formulated. (First Release September 15 2015; J Rheumatol 2015;42:1954–61; doi:10.3899/jrheum.141446)
The outcome measurement instruments used in a study become its metric of benefit (or unintended harm) of an intervention. Through their scope and content, they define what can or cannot be said about the intervention, as well as how accurately and precisely it can be said. The choice of outcomes (target domains), therefore defines how we will understand the effects of treatment\textsuperscript{1,2,3}. Choosing outcome measurement instruments is not a decision to be taken lightly by either the study designer or the systematic reviewer\textsuperscript{4}. Heterogeneity of the outcomes and of the outcome measurement instruments found within systematic reviews leads frustrated reviewers to abandon syntheses, report on only certain instruments (and therefore only certain studies using those instruments), or derive techniques to combine data across outcomes. All these mechanisms diminish confidence in results and introduce a risk of bias related to outcome reporting\textsuperscript{5,6,7,8}. It is clear that summary of findings (SoF) tables found in systematic reviews cannot display all outcomes or outcome measurement instruments fielded in every trial gathered during a systematic review: some priority setting must be done\textsuperscript{9}.

In 2006, Juni described a predefined hierarchy for outcomes instruments to be included in metaanalyses\textsuperscript{10} that was adopted by the musculoskeletal Cochrane group\textsuperscript{9}. In 2012 Juhl and colleagues described another approach to avoid outcome biases\textsuperscript{11}. Selecting all trials that fielded multiple outcomes of either pain or disability in one of the top 10 journals in internal medicine or rheumatology, they created a standardized mean difference (SMD) to summarize the effect detected, and ranked instruments within each study according to the magnitude of the SMD (effect size detected). The ranks for each instrument were averaged across all the studies that fielded it (minimum of 5 studies or it was not considered). This mean rank score was then used to rank across instruments in order to see which one(s) could be included in SoF tables to represent that concept (see Table 1)\textsuperscript{11}.

Juhl’s approach is transparent and logically tries to capture more commonly used outcome measurement instruments in top ranked journals. Top journals are used to assure some level of quality of the outcome measurement instruments, and the effect size serves as a proxy for validity of change and ability to discriminate. However, there are limitations to this approach. First, it is limited to instruments used in several trials, older scales with track records are favored over newer scales that may perform as well or better than the older instruments\textsuperscript{12,13}. Second, it does not make use of a growing body of literature on measurement properties of different instruments that could provide high quality evidence of instrument performance\textsuperscript{8,14,15,16}. Third, as a means of selection, this approach does not address the conceptual focus of an outcome measurement instrument and so risks missing differences in concepts across instruments that may appear, by their title, to be addressing similar targets (such as pain). Finally, it favors those picking up larger relative effect sizes rather than ones with appropriate effect sizes. Larger is assumed to be better when it could be the result of “noise.” An argument has been made that both concept (outcome) and how it is captured (outcome measurement instrument) be considered when selecting methods to present findings in SoF tables\textsuperscript{6,7}.

We describe an alternative approach to prioritizing instruments to be included in SoF tables using an evidence-based approach, with emphasis on the instruments’ conceptual focus and measurement properties. This was developed and refined in conjunction with the preconference workshop on the measurement of pain held ahead of the Outcome Measures in Rheumatology 12 (OMERACT 12) meeting in Budapest in May, 2014.

**Alternative Approach: Organizations**

Consensus-based Standards for the selection of health Measurement INstruments (COSMIN, www.cosmin.nl). The COSMIN initiative was founded in 2005 with the aim of improving selection of health measurement instruments. COSMIN’s efforts include establishing methods for searching literature\textsuperscript{17}, and establishing consensus-based standards for assessing the methodological quality of measurement property studies. Latterly, a Delphi process has been used to reach consensus on how studies of measurement

### Table 1. Description of the 3 multiitem pain measures considered in workshop exercise.

<table>
<thead>
<tr>
<th>Scale (reference)</th>
<th>No. of Items</th>
<th>Response Options</th>
<th>Concepts Assessed</th>
</tr>
</thead>
<tbody>
<tr>
<td>WOMAC Pain\textsuperscript{24} (Western Ontario and McMaster Universities Osteoarthritis Index — pain subscale)</td>
<td>5</td>
<td>5-point Likert or 0–100 visual analog scale. Likert: none, mild, moderate, severe, extreme</td>
<td>Pain during specified daily activities (pain during walking, pain descending stairs, in bed, sitting or lying, and standing) Effect of pain on physical and mental parts of life. Separates intermittent pain and constant pain, sums together for total As per WOMAC plus additional items on pain with twist, bend, and straightening to reduce ceiling effect and catch milder disease</td>
</tr>
<tr>
<td>ICOAP\textsuperscript{26} (Intermittent and Constant Osteoarthritis Pain)</td>
<td>11</td>
<td>5-point Likert: Not at all, mildly, moderately, severely, extremely</td>
<td></td>
</tr>
<tr>
<td>KOOS Pain\textsuperscript{25*} (Knee Injury and Osteoarthritis Score)</td>
<td>9</td>
<td>5-point Likert</td>
<td></td>
</tr>
</tbody>
</table>

* KOOS tool includes and expands upon the pain component of the WOMAC group.
sets) and finding all the candidate measures for a given OMERACT would focus on defining domains (core outcome of that evidence against agreed upon methodological study to create the evidence, and COSMIN defines the quality it has passed the “OMERACT Filter” of truth (validity), candidate outcome measurement instruments to measure each of 9 measurement properties included in systematic reviews. COSMIN and their followers have gone on to recommend methods for summarizing the results of studies of measurement properties into an evidence ranking [i.e., 2 or more “excellent” quality studies supporting a property as the highest level (similar to GRADE)].

OMERACT (www.omeract.org). OMERACT was founded in 1992 with the aim of standardizing the measurement of outcomes in arthritis research. OMERACT is also a consensus-based group in rheumatology that seeks to ensure representation of 4 core areas of interest (death, life impact, resource use/economic impact, and pathophysiological manifestations) in the outcome battery across intervention studies in arthritis. Adverse events and contextual factors must also be considered. Within each core area, specific domains are selected, such as functional status, pain, ability to work, disease activity, and utility. Consensus is achieved at OMERACT meetings as to the appropriateness of the proposed domains in each area. This list of domains becomes the core outcome set. The next step, finding or developing candidate outcome measurement instruments to measure each core domain, is judged, again, by consensus, on evidence that it has passed the “OMERACT Filter” of truth (validity), feasibility (practicality, cost, burden), and discrimination (precision, responsiveness, sensitivity to change in a clinical trial setting and interpretability in responder analyses). Evidences supporting this is gathered from existing literature, or in its absence, from studies designed to address the gap. Multiple studies are required to support each property. OMERACT defines, gathers and creates evidence, but does not have a specific process for systematically reviewing the literature, or defining specific criteria for the strength of that literature.

Both COSMIN and OMERACT seek to put the best instruments into the hands of researchers and clinicians. OMERACT defines the nature of the evidence that needs to be gathered either through literature review or conducting a study to create the evidence, and COSMIN defines the quality of that evidence against agreed upon methodological standards.

An approach blending the strengths of COSMIN and OMERACT would focus on defining domains (core outcome sets) and finding all the candidate measures for a given domain. It would then proceed to review, and if necessary, create, the evidence of the measurement properties that are important for its intended use. This approach would not necessarily prioritize instruments that are more frequently used, and would allow room for emerging measures with good measurement properties and performance to rise to a SoF table. It would also move towards ensuring clarity in the concept of the target outcome domain, and in the concept being quantified by a candidate measure. We anticipate the result will be a body of evidence showing how an instrument is likely to perform in a given context of use, as well as identifying gaps in need of additional study.

Workshop
At the pre-OMERACT workshop on the measurement of pain in clinical trials and systematic reviews, a group of participants (largely clinical or academic researchers with a special interest in measurement methodology) were invited to consider if the same measures would be selected for a SoF table if an evidence-based approach were used in lieu of the Juhl (2012) approach.

All participants received material prior to the workshop on the Juhl approach and on a number of candidate articles to be considered. They received 1 overview article on the measurement of pain in adults with arthritis (Hawker, 2011) as well as articles about 3 pain instruments found in osteoarthritis research: the WOMAC (Western Ontario and McMaster Universities Osteoarthritis Index) Pain scale, the KOOS (Knee injury and Osteoarthritis Outcome Score) Pain Scale and the ICOAP (Intermittent and Constant Osteoarthritis Pain scale). If a future trial fielding all 3 of these scales were created, only the WOMAC Pain scores would be considered in the Juhl-based SoF table. The other measures selected are newer and emerging in the field and are therefore not on the Juhl 2012 list (not yet found in published trials). That said, there is some suggestion in the literature that their responsiveness in clinical trial settings is the same as, if not slightly better than, the WOMAC, and would perhaps be worthy of consideration as new trials arise using them. Concepts and structure of these instruments are described briefly in Table 1.

RESULTS
Participants actively engaged the Juhl article and the importance of criteria for SoF tables for Cochrane reviews given the heterogeneity of outcome instruments found across clinical trials. They also heard a discussion on the role of measurement properties as a source of evidence for the quality of outcome measurement instruments. The discussions and subgroup discussions focused on process and on staging an approach to decide on outcomes instruments to be included using an evidence-based approach. The following describes the overall suggestions to be utilized by the research agenda that emerged.
A 3-phase decision-making process was suggested by the attenders: 1. Ensuring the instrument measures the concept of interest; 2. Considering practical aspects of the outcome measurement instrument; and 3. Gathering high quality evidence of the necessary measurement properties in a similar context of use.

**Ensuring the instrument measures the concept of interest.** The workshop process suggested that the first step in selecting an outcome measurement instrument is to discern whether there is a clear match between the concept quantified in the instrument and the target outcome concept. Through direct comparison and discussion, it was agreed that the instruments we provided for the review captured very different concepts of pain. Pain can be quantified on its intensity, frequency, or its effect on daily activity (i.e., the degree to which pain prevents one from performing a specific activity). Participants felt unable to assess fit and face validity without a clear understanding of both the concept in the instrument to be measured and the concept of the target outcome. This is important because the Juhl approach and a focused review of measurement properties could both miss the subtle but important differences in the concepts that were raised when reviewing the instruments in the workshop.

Defining the target concept and the factors having a direct effect on that concept is part of — in the language of regulators 27,28 — defining the context of use, or the intended use argument, as described in some measurement methodologies 29,30. Knowing what you want to measure, in whom you wish to measure, and what claims will be made from the numeric scores is important to consider up front for both study designers and for systematic reviewers. In our initial discussions about the 3 tools, it became clear that while all 3 tools aim to measure change in pain in persons with knee osteoarthritis, the actual concept of pain varied across instruments. The KOOS and WOMAC place pain experiences within very specific contexts, such as going up stairs or stooping/bending 24,25, which provide a way to monitor pain experiences in very structured situations. The KOOS was developed to address milder knee pain, adding items like pain during twisting of the knee 25. The ICOAP, based on qualitative research, found pain experience in osteoarthritis to be separated into an intermittent type of pain, and a constant/persistent pain 26. ICOAP also asks about the pain experience in broader contexts than the KOOS and WOMAC, for example, pain in sleep, and pain “in activities” (without defining specific activities). Our workshop participants emphasized that there are several different experiences and expressions of pain that are not all the same. They suggested that one should pay careful attention to the concept of pain one wishes to measure and the concept of pain that is being captured in the content and scoring of the candidate instrument, before considering the instrument for SoF tables.

We propose, as has been outlined in some instrument selection guides 31,32, that this type of scrutiny should be an emphasized first step for both study designers and systematic reviewers. The target outcome domain of interest must be articulated and must match the measured domain of each candidate outcome measurement instrument before it is considered to be a serious contender.

**Considering practical aspects of the outcome measurement instrument.** Workshop participants suggested that the practicalites of using an instrument [often called feasibility (OMERACT Filter), clinical utility 33, applicability 34,35, or sensibility 36] should be considered very early in the selection process, and certainly before statistical properties are considered. Practical limitations in an instrument’s use are often insurmountable and will prevent its use 30. Auger 34 suggested such consideration should include domains of patient burden (length, language, response burden), researcher burden (cost, availability, equipment needs, scoring difficulty), distribution of scores, and acceptability of format. Other practical considerations are acceptability to the particular patient group, reading and health literacy levels, content validity, and face validity (also addressed in concept match above). Evaluation can be done by the user team, but is greatly enhanced by patient/respondent input, particularly in patient-reported outcomes 37. Consideration of these practical components early in the selection process is planned for the next version of the COSMIN protocol 38 and is already embedded as a key component of the OMERACT Filter 2.0 39

**Gathering high quality evidence of the necessary measurement properties in a similar context of use.** Once candidate instruments assessing the desired concept have been identified and assessed for feasibility, a full review of the measurement properties needed for a given application can be undertaken. If necessary, additional information is created to fill any gaps. The specific evidence that is needed depends on the context of use (target concept, population, and trial design).

Consistent with the principles of OMERACT, key measurement properties will include truth (content, construct, and criterion validity), and the ability to discriminate in clinical trial settings (precision, test-retest reliability, longitudinal construct validity/ability to detect change that has occurred, and sensitivity to the differences experienced by 2 treatment groups) 30. Evidence from radically different patient populations, or addressing other properties should not be considered in the decision-making process. Thus, the OMERACT filter narrows the type of evidence needed in the decision making, and emphasizes the conceptual and practical considerations as the first steps in deciding on an instrument (see Figure 1) 18,40. Both OMERACT and COSMIN support the need for multiple, high quality studies with consistent evidence of each property in the target population to provide greater confidence in performance.

Consistent with the principles of COSMIN 37, a systematic review of the literature should be conducted, and measure-
Figure 1. Depiction of the decision-making process determining whether a candidate measure fulfills OMERACT Filter requirements for a defined measurement need. Match of concept to the intended need is an essential first step. If a match does not exist, the pursuit of measurement properties is not necessary (from Beaton, et al. Outcome measurement, Ch 30. In: Firestein, et al, eds. Kelley’s textbook of rheumatology, 10th edition. Oxford: Saunders Elsevier, in press; with permission). In Step 2, practical considerations are evaluated; if cost, burden, or equipment needs are prohibitive, it is best to select another tool. Steps 3 and 4 are the compilation of measurement property evidence and would parallel a COSMIN-based systematic review [www.cosmin.nl (accessed August 20, 2014)] and synthesis of the evidence. Step 5 is needed for responder analyses (% responded) and important criteria in patient-centered research.
Research Agenda

1. To continue development of a feasible template/toolkit to assist in defining measurement need and assessing candidate instrument match to that need. To place/reinforce this as the first stage in the selection of an instrument
2. To evaluate whether evidence-based approaches offer different recommendations for measures to be included in SoF tables, versus a Juhl-style approach, and if any different conclusions would be drawn from clinical trials if these were utilized
3. To stratify outcomes based on the quality of the evidence supporting their measurement properties18,45, and to test if this has a differential effect on the results of a systematic review and on techniques used to blend data from different instruments into metaanalyses6,7.

In conclusion, SoF tables in systematic reviews cannot report evidence found across all the various instruments currently being fielded in the literature. There are simply too many. Faced with a growing body of literature making use of different pain and disability instruments in knee osteoarthritis trials, Juhl, et al created a transparent, reproducible means to select and prioritize the outcome measurement instruments to be included in SoF tables11. In the present article, we suggest an alternative, evidence-based approach to prioritize outcomes based on the quality of the (pain) instruments. Here quality is defined through a match of the target concept with the concept being quantified in an instrument (not always clearly articulated by the conceptors), consideration of very practical aspects of instrument use in a study situation (checking content against target concept), and a systematic review of the relevant measurement properties.

Systematic reviews of measurement properties, in turn, consider the methodological quality and risk of bias in looking for high quality evidence from multiple studies before reaching a conclusion about that measurement property. All 3 elements (conceptual match, practicalities, and measurement properties) are critical to the evidence-based approach. Our process combined the experiences of 2 outcome measurement groups, OMERACT and COSMIN, and was based on articles reviewed in our workshop12,13. We believe this approach could lead to additional or different contenders for the SoF list11 because newer measures may have been developed using stronger methodologies, but have insufficient field application to meet Juhl’s recommendations. Our group recommends an evidence-based approach be considered for the selection of outcome measurement instruments, with evidence being derived from high-quality studies of relevant measurement properties of candidate instruments.

ACKNOWLEDGMENT

The authors wish to thank all of the workshop participants in the preconference workshop on pain, Budapest, OMERACT 12 (2014). They also thank Patricia Nedanovski and Taucha Inrig for their assistance in manuscript preparation for submission.

REFERENCES


Boe, et al: Evidence-based outcome selection


consensus on how to select outcome measurement instruments for outcomes included in a 'core outcome set'. Trials 2014;15:247.


