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)

Key Indexing Terms:
EVIDENCE BASED OUTCOME MEASUREMENT REPRODUCIBILITY OF RESULTS
HEALTH STATUS INDICATORS SYSTEMATIC REVIEWS

From the Musculoskeletal Health and Outcomes Research, Li Ka Shing Knowledge Institute, St Michael's Hospital, Toronto; Institute for Work & Health; University of Toronto, Toronto, Canada; VU University Medical Centre, Department of Epidemiology and Biostatistics and EMGO Institute for Health and Care Research, Amsterdam, The Netherlands; Birmingham Veterans Affairs Medical Center and University of Alabama at Birmingham, Birmingham, AL, USA; Women's College Hospital, Institute for Clinical Evaluative Sciences and University of Toronto, Toronto, Canada; Seattle Quality of Life Group\Center for Disability Policy and Research, University of Washington, Seattle, WA; Office of New Drugs, Centre for Drug Evaluation and Research, Food and Drug Administration, Silver Spring, MD, USA; Department of Epidemiology and Community Medicine, Faculty of Medicine, University of Ottawa; Department of Medicine, Faculty of Medicine, Ottawa Hospital Research Institute; Clinical Epidemiology Program, University of Ottawa; Department of Epidemiology and Community Medicine, Faculty of Medicine, Institute of Population Health, Ottawa, Canada.

Supported through the OMERACT premeeting conference on pain measurement. JAS is supported by grants from the Agency for Health Quality and Research Center for Education and Research on Therapeutics (AHRQ CERTs) U19 HS021110, US National Institute of Arthritis, Musculoskeletal and Skin Diseases (NIAMS) P50 AR060772 and U34 AR062891, National Institute of Aging U01 AG018947, National Cancer Institute U10 CA149950, the resources and the use of facilities at the VA Medical Center at Birmingham, Alabama and research contract CE-1304-6631 from the Patient Centered Outcomes Research Institute.

JAS has received research grants from Takeda and Savient and consultant fees from Savient, Takeda, Regeneron, and Allergan. JAS is a member of the American College of Rheumatology's Guidelines Subcommittee of the Quality of Care Committee; and a member of the Veterans Affairs Rheumatology Field Advisory Committee. JAS, DEB, and PT are members of the executive of OMERACT, an organization that develops outcome measures in rheumatology and receives arms-length funding from 36 companies.

D.E. Beaton, BScOT, PhD, Senior Scientist Institute for Work & Health, Scientist St. Michael's Hospital; Associate Professor, University of Toronto; C.B. Terwee, PhD, Senior Epidemiologist, VU University Medical Centre, Department of Epidemiology and Biostatistics and EMGO Institute for Health and Care Research; J.A. Singh, MBBS, MPH, Associate Professor of Medicine, Birmingham Veterans Affairs Medical Center and University of Alabama at Birmingham; G.A. Hawker, MD, MSc, FRCPC, Women's College Hospital, Institute for Clinical Evaluative Sciences and University of Toronto; D.L. Patrick, PhD, Seattle Quality of Life Group\Center for Disability Policy and Research, University of Washington; L.B. Burke, RPh, MPH, Study Endpoints and Label Development, Office of New Drugs, Centre for Drug Evaluation and Research, Food and Drug Administration; K. Toupin-April, BScOT, PhD, Associate Scientist, Assistant Professor, Department of Epidemiology and Community Medicine, Faculty of Medicine, University of Ottawa; P.S. Tugwell, MD, MSc, University of Ottawa, Department of Medicine, Faculty of Medicine, Ottawa Hospital Research Institute, Clinical Epidemiology Program, University of Ottawa, Department of Epidemiology and Community Medicine, Faculty of Medicine, Institute of Population Health.

Address correspondence to Dr. Beaton, St. Michael's Hospital, 30 Bond St, Toronto, ON, M5B 1W8; E-mail: beatond@smh.ca

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[1,2,3]. Choosing outcome measurement instruments is not a decision to be taken lightly by either the study designer or the systematic reviewer[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[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[9].

In 2006, Juni described a predefined hierarchy for outcomes instruments to be included in metaanalyses[10] that was adopted by the musculoskeletal Cochrane group[9]. In 2012 Juhl and colleagues described another approach to avoid outcome biases[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)[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 as or better than the older instruments[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[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[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[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.

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

* KOOS tool includes and expands upon the pain component of the WOMAC group.

properties should be viewed in terms of correct methodology (study design requirements and preferred statistical methods). These standards are included in the COSMIN checklist (dichotomous, or 4-point "excellent" to "poor ratings") for each of 9 measurement properties. Each property is then assessed by 4–18 independent standards. Because of its attention to the use of the correct methodology in measurement property studies, the COSMIN checklist can also be used for designing measurement properties studies, reporting on measurement properties studies, evaluating the quality of submitted or published studies of measurement properties, as well as evaluating the quality of all studies of 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[18] into an evidence ranking [i.e., 2 or more "excellent" quality studies supporting a property as the highest level (similar to GRADE)[19]].

OMERACT (www.omeract.org). OMERACT was founded in 1992 with the aim of standardizing the measurement of outcomes in arthritis research[20]. 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"[1,20,21] of truth (validity), feasibility (practicality, cost, burden), and discrimination (precision, responsiveness, sensitivity to change in a clinical trial setting and interpretability in responder analyses). Evidence 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[22]. 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[11].

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)[23], as well as articles about 3 pain instruments found in osteoarthritis research: the WOMAC (Western Ontario and McMaster Universities Osteoarthritis Index) Pain scale[24], the KOOS (Knee injury and Osteoarthritis Outcome Score) Pain Scale[25] and the ICOAP (Intermittent and Constant Osteoarthritis Pain scale)[26]. 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[11]. 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)[11]. 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[12,13]. 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 practicalities 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[17], a systematic review of the literature should be conducted, and measure-

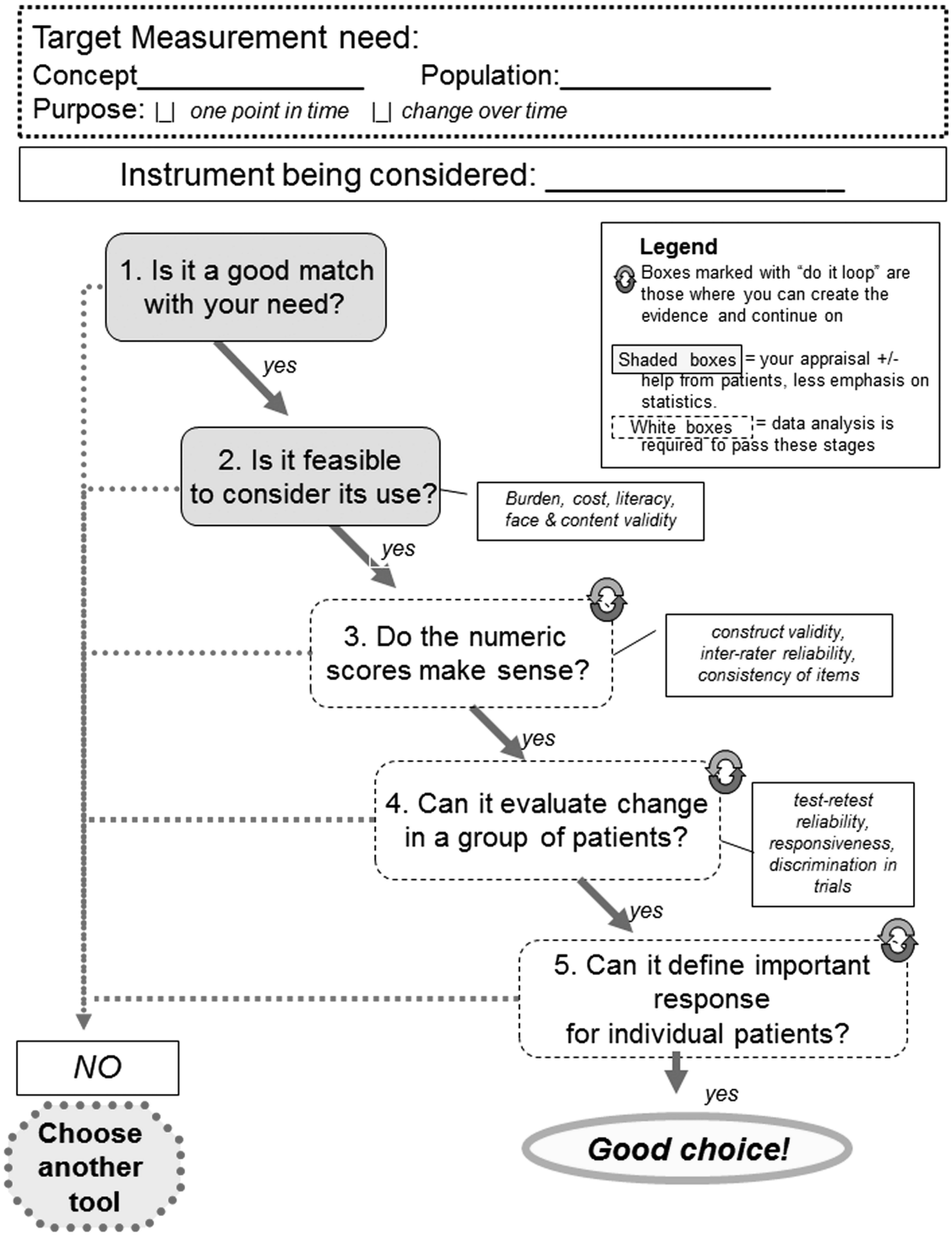

Figure 1. Depiction of the decision-making process determining whether a candidate measure fulfills OMERACT Filter requirements for a defined measurement need. Match of concept to the intended need is an essential first step. If a match does not exist, the pursuit of measurement properties is not necessary (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[40]; with permission). In Step 2, practical considerations are evaluated; if cost, burden, or equiment 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[18]. Step 5 is needed for responder analyses (% responded) and important criteria in patient-centered research.

ment property studies identified should be assessed for their methodological quality[16,41]. In a review of systematic reviews of measurement properties, Mokkink found they lacked the important step of quality appraisal, and often a standardized data synthesis technique was not applied[42]. Search strategies varied greatly, and may not have been thorough enough to capture all relevant studies[42]. COSMIN emphasizes the importance of making a distinction between good quality and lesser quality methods used in measurement property studies, because lower quality studies can lead to the selection of flawed outcome measures in effectiveness studies and SoF tables, and, in turn, produce biased information in a systematic review or metaanalysis. Following the model of best evidence synthesis, if the quality of a study of measurement properties is poor, the quality of the instrument under scrutiny cannot be judged. To improve this situation, COSMIN developed a 10-step protocol for performing systematic reviews of outcome measurement instruments based on general guidelines for systematic reviews of the Cochrane Collaboration (clear research questions, comprehensive search strategies, explicit selection criteria, critical appraisal, clear approaches to synthesis and conclusions)[43]. Guidelines are included at several stages, including performing a systematic literature search (a specific search filter for PubMed was developed[17]), assessing the methodological quality of the included studies (a specific 4-point rating scale version of the COSMIN checklist was developed[16]), and a systematic approach for data synthesis suggested[18,44].

The ability to discern the quality of studies in a review through a detailed appraisal of their methods, and so draw attention to the risk of bias, is a contribution of the COSMIN group to the OMERACT process. The need to clearly define target domains of interest (along with the population and intended use), and the importance of reaching consensus across multiple stakeholders on the domains, quality, and results of pertinent measurement properties, is a contribution of OMERACT to the COSMIN process. OMERACT defines "stakeholders" as researchers, patients, industry, regulators, and clinicians; however, special attention is paid to patient research partners[1]. Together OMERACT and COSMIN processes provide an improved evidence base from which decisions can be made about instruments to be used in interventional studies. This well-known goal of OMERACT here finds synergy not only with COSMIN, but also in the context of systematic reviews for treatment effect, which is a goal of Cochrane Review SoF tables.

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 properties[18,45], 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 metaanalyses[6,7].

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 tables[11]. 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 workshop[12,13]. We believe this approach could lead to additional or different contenders for the SoF list[11] 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

1. Boers M, Kirwan JR, Wells G, Beaton D, Gossec L, D'Agostino MA, et al. Developing core outcome measurement sets for clinical trials: OMERACT filter 2.0. J Clin Epidemiol 2014;67:745-53.
2. Kirwan JR, Boers M, Hewlett S, Beaton D, Bingham CO III, Choy E, et al. Updating the OMERACT filter: core areas as a basis for defining core outcome sets. J Rheumatol 2014;41:994-9.

3. Tugwell P, Boers M, D'Agostino MA, Beaton D, Boonen A, Bingham CO, III, et al. Updating the OMERACT filter: implications of filter 2.0 to select outcome instruments through assessment of "truth": content, face, and construct validity. J Rheumatol 2014;41:1000-4.
4. Clarke M. Standardising outcomes for clinical trials and systematic reviews. Trials 2007;8:39.
5. Boers M, Idzerda L, Kirwan JR, Beaton D, Escorpizo R, Boonen A, et al. Toward a generalized framework of core measurement areas in clinical trials: a position paper for OMERACT 11. J Rheumatol 2014;41:978-85.
6. Johnston BC, Patrick DL, Thorlund K, Busse JW, da Costa BR, Schunemann HJ, et al. Patient-reported outcomes in meta-analyses-part 2: methods for improving interpretability for decision-makers. Health Qual Life Outcomes 2013;11:211.
7. Johnston BC, Patrick DL, Busse JW, Schunemann HJ, Agarwal A, Guyatt GH. Patient-reported outcomes in meta-analyses-Part 1: assessing risk of bias and combining outcomes. Health Qual Life Outcomes 2013;11:109.
8. Gabriel SE, Normand SL. Getting the methods right—the foundation of patient-centered outcomes research. N Engl J Med 2012;367:787-90.
9. Ghogomu EA, Maxwell LJ, Buchbinder R, Rader T, Pardo PJ, Johnston RV, et al. Updated method guidelines for Cochrane musculoskeletal group systematic reviews and metaanalyses. J Rheumatol 2014;41:194-205.
10. Juni P, Reichenbach S, Dieppe P. Osteoarthritis: rational approach to treating the individual. Best Pract Res Clin Rheumatol 2006;20: 721-40.
11. Juhl C, Lund H, Roos EM, Zhang W, Christensen R. A hierarchy of patient-reported outcomes for meta-analysis of knee osteoarthritis trials: empirical evidence from a survey of high impact journals. Arthritis 2012;2012:136245.
12. Davis AM, Lohmander LS, Wong R, Venkataramanan V, Hawker GA. Evaluating the responsiveness of the ICOAP following hip or knee replacement. Osteoarthritis Cartilage 2010;18:1043-5.
13. Risser RC, Hochberg MC, Gaynor PJ, D'Souza DN, Frakes EP. Responsiveness of the Intermittent and Constant Osteoarthritis Pain (ICOAP) scale in a trial of duloxetine for treatment of osteoarthritis knee pain. Osteoarthritis Cartilage 2013;21:691-4.
14. Fava GA, Ruini C, Rafanelli C. Psychometric theory is an obstacle to the progress of clinical research. Psychother Psychosom 2004;73:145-8.
15. Feinstein AR. Clinical biostatistics XLI. Hard science, soft data, and the challenges of choosing clinical variables in research. Clin Pharmacol Ther 1977;22:485-98.
16. Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res 2012;21:651-7.
17. Terwee CB, Jansma EP, Riphagen II, de Vet HC. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. Qual Life Res 2009;18:1115-23.
18. Schellingerhout JM, Verhagen AP, Heymans MW, Koes BW, de Vet HC, Terwee CB. Measurement properties of disease-specific questionnaires in patients with neck pain: a systematic review. Qual Life Res 2012;21:659-70.
19. Guyatt GH, Thorlund K, Oxman AD, Walter SD, Patrick D, Furukawa TA, et al. GRADE guidelines: 13. Preparing summary of findings tables and evidence profiles-continuous outcomes. J Clin Epidemiol 2013;66:173-83.
20. Tugwell P, Boers M, Brooks P, Simon L, Strand V, Idzerda L. OMERACT: an international initiative to improve outcome measurement in rheumatology. Trials 2007;8:38.
21. Boers M, Brooks P, Strand V, Tugwell P. The OMERACT Filter for outcome measures in rheumatology. J Rheumatol 1998;25:198-9.
22. Kane MT. Validation as a pragmatic, scientific activity. J Educ Meas 2013;50:115-22.
23. Hawker GA, Mian S, Kendzerska T, French M. Measures of adult pain: Visual Analog Scale for Pain (VAS Pain), Numeric Rating Scale for Pain (NRS Pain), McGill Pain Questionnaire (MPQ), Short-Form McGill Pain Questionnaire (SF-MPQ), Chronic Pain Grade Scale (CPGS), Short Form-36 Bodily Pain Scale (SF-36 BPS), and Measure of Intermittent and Constant Osteoarthritis Pain (ICOAP). Arthritis Care Res (Hoboken) 2011;63 Suppl 11:240-52.
24. Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. J Rheumatol 1988;15:1833-40.
25. Roos EM, Toksvig-Larsen S. Knee injury and Osteoarthritis Outcome Score (KOOS) - validation and comparison to the WOMAC in total knee replacement. Health Qual Life Outcomes 2003;1:17.
26. Hawker GA, Davis AM, French MR, Cibere J, Jordan JM, March L, et al. Development and preliminary psychometric testing of a new OA pain measure-an OARSI/OMERACT initiative. Osteoarthritis Cartilage 2008;16:409-14.
27. Reeve BB, Wyrwich KW, Wu AW, Velikova G, Terwee CB, Snyder CF, et al. ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. Qual Life Res 2013;22:1889-905.
28. U.S. Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research (CDER). Guidance for industry: Patient-reported outcome measures: Use in medical product development to support labeling claims. 2009. Available from: http://www.fda.gov/downloads/Drugs/Guidances/UCM193282.pdf
29. Kane MT. Validating the interpretations and uses of test scores. J Educ Meas 2013;50:1-73.
30. Beaton DE, Boers M, Tugwell P. Health outcomes assessment. In: Kelley's textbook of rheumatology. 9 ed. Philadelphia: Saunders Elsevier; 2013:462-75.
31. Beaton DE, van Eerd D, Smith P, van der Velde G, Cullen K, Kennedy CA, et al. Minimal change is sensitive, less specific to recovery: a diagnostic testing approach to interpretability. J Clin Epidemiol 2011;64:487-96.
32. Valderas JM, Ferrer M, Mendivil J, Garin O, Rajmil L, Herdman M, et al. Development of EMPRO: a tool for the standardized assessment of patient-reported outcome measures. Value Health 2008;11:700-8.
33. Law M. Measurement in occupational therapy: Scientific criteria for evaluation. Can J Occup Ther 1987;54:133-8.
34. Auger C, Demers L, Swaine B. Making sense of pragmatic criteria for the selection of geriatric rehabilitation measurement tools. Arch Gerontol Geriatr 2006;43:65-83.
35. Auger C, Demers L, Desrosiers J, Giroux F, Ska B, Wolfson C. Applicability of a toolkit for geriatric rehabilitation outcomes. Disabil Rehabil 2007;29:97-109.
36. Feinstein AR, Wells CK, Joyce CM, Josephy BR. The evaluation of sensibility and the role of patient collaboration in clinimetric indexes. Trans Assoc Am Physicians 1985;98:146-9.
37. Tang K, Beaton DE, Lacaille D, Gignac MA, Bombardier C. Sensibility of five at-work productivity measures was endorsed by patients with osteoarthritis or rheumatoid arthritis. J Clin Epidemiol 2013;66:546-56.
38. Prinsen CA, Vohra S, Rose MR, King-Jones S, Ishaque S, Bhaloo Z, et al. Core Outcome Measures in Effectiveness Trials (COMET) initiative: protocol for an international Delphi study to achieve

consensus on how to select outcome measurement instruments for outcomes included in a 'core outcome set'. Trials 2014;15:247.
39. Wells G, Beaton DE, Tugwell P, Boers M, Kirwan JR, Bingham CO, III, et al. Updating the OMERACT filter: discrimination and feasibility. J Rheumatol 2014;41:1005-10.
40. Beaton DE, Boers M, Tugwell P. Outcome measurement, Ch. 30. In: Firestein GS, Budd RC, Gabriel SE, McInnes IB, O'Dell JR, eds. Kelley's textbook of rheumatology, 10th edition. Oxford: Saunders Elsevier; in press.
41. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res 2010;19:539-49.
42. Mokkink LB, Terwee CB, Stratford PW, Alonso J, Patrick DL, Riphagen I, et al. Evaluation of the methodological quality of systematic reviews of health status measurement instruments. Qual Life Res 2009;18:313-33.
43. Clarke M. The Cochrane Collaboration and systematic reviews. Br J Surg 2007;94:391-2.
44. Schellingerhout JM, Heymans MW, Verhagen AP, Lewis M, de Vet HC, Koes BW. Prognosis of patients with nonspecific neck pain: development and external validation of a prediction rule for persistence of complaints. Spine 2010;35:827-35.
45. Kennedy CA, Beaton DE, Smith P, Van Eerd D, Tang K, Inrig T, et al. Measurement properties of the QuickDASH (Disabilities of the Arm, Shoulder and Hand) outcome measure and cross-cultural adaptations of the QuickDASH: a systematic review. Qual Life Res 2013;22:2509-47.