Outcome Instruments: Rationale for Their Use

By Rudolf W. Poolman, MD, PhD, Marc F. Swiontkowski, MD, Jeremy C.T. Fairbank, MD, FRCS, Emil H. Schemitsch, MD, FRCSC, Sheila Sprague, MSc, and Henrica C.W. de Vet, PhD

The number of outcome instruments available for use in orthopaedic observational studies has increased dramatically in recent years. Properly developed and tested outcome instruments provide a very useful tool for orthopaedic research. Criteria have been proposed to assess the measurement properties and quality of health-status instruments. Unfortunately, not all instruments are developed with use of strict quality criteria. In this article, we discuss these quality criteria and provide the reader with a tool to help select the most appropriate instrument for use in an observational study. We also review the steps for future use of outcome instruments, including the standardization of their use in orthopaedic research.

Introduction

The number of outcome instruments available has increased substantially since the last outcome symposium was published in *The Journal of Bone and Joint Surgery*[1]. The proliferating number of scales that are available complicates the choice faced by investigators during selection of the most appropriate instrument[2-4]. Quality criteria have recently been proposed to assess the measurement properties of health-status instruments[2,5].

Outcome instruments can be used for several purposes, including the evaluation of patients in clinical practice[6] or clinical research[1]. Because the focus of this supplement is on nonrandomized studies, we will refer to outcome instruments that are useful in a clinical research setting rather than on those that are useful in the assessment of individual patients during daily clinical work[4]. However, an extremely important use of these tools is to improve care. This can be accomplished by routinely instituting the use of validated instruments in daily clinical practice and by modifying treatment protocols on the basis of the results.

Clinical outcomes, such as those obtained through the imaging of fractures and through physical examination to assess range of motion[3,7,8] (traditionally described as *objective* or *hard* measures), are valuable in orthopaedic research. Unfortunately, these measures can be subject to interrater disagreement and they often do not provide definitive answers about whether an intervention is useful from a patient's perspective[9]. Furthermore, objective measures may also correlate poorly with a patient's own feelings of wellness[9]. In addition to objective measures, the orthopaedic literature often includes patient-reported, or *subjective* or *soft* outcome measures, such as the measurement of pain or satisfaction following a procedure[3,10]. Well-designed patient-reported instruments have undergone rigorous testing and may be better validated and have greater reproducibility than the so-called objective outcomes[3]. Thus, outcome objectivity is not determined by whether a clinician measures a parameter directly, but rather it is dependent on the reliability or reproducibility of a finding[11]. It cannot be stated that, by definition, objective measures are better than subjective measures or the other way around.

The most important feature of outcome instruments is their ability to test whether treatment is effective in improving symptoms or function from a patient's point of view[9]. Many funding agencies and research ethic boards insist that a quality-of-life instrument (instruments that capture a wider perspective of the state of well-being) be included in the design of proposed clinical studies[9]. In an effort to standardize outcome measurement and reporting in arthroplasty research, *The Journal of Bone and Joint Surgery* strongly encourages inclusion of the Western Ontario and McMaster University Osteoarthritis Index (WOMAC) in studies that report the results of hip and knee arthroplasty.

In this article, we will discuss types of outcome instruments, suggest a conceptual model that will help eliminate confusing terminology such as "hard" and "soft," discuss how to locate the appropriate outcome instruments, describe a method to evaluate outcome instruments on the basis of quality criteria, and discuss the potential pitfalls in utilizing outcome instruments for research purposes. Moreover, we will only refer to validated outcome instruments.

The Conceptual Model of Patient Outcomes as Proposed by Wilson and Cleary

Wilson and Cleary proposed a classification scheme for different measures of health outcome (Fig. 1)[12]. They conceptualized five levels of outcomes: 1) biological and phys-

Disclosure: In support of their research for or preparation of this work, one or more of the authors received, in any one year, outside funding or grants in excess of $10,000 from the National Institute of Child Health and Human Development (NICHD). Neither they nor a member of their immediate families received payments or other benefits or a commitment or agreement to provide such benefits from a commercial entity.

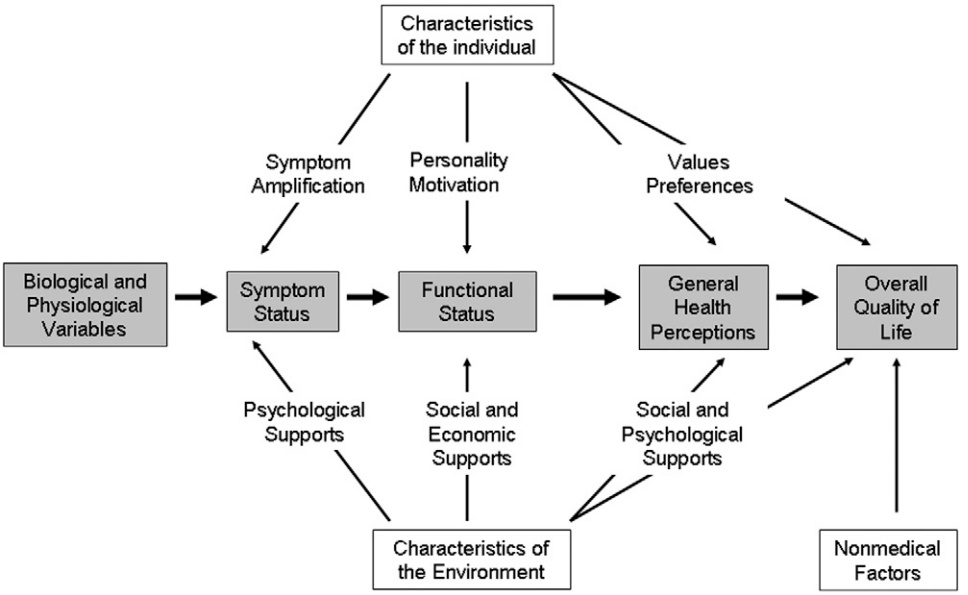

Fig. 1

The Wilson and Cleary conceptual model of patient outcomes. (Reprinted, with permission, from: Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. JAMA. 1995;273:60. Copyright ©1995, American Medical Association. All rights reserved.)

iological variables (level one), 2) symptom status (level two), 3) functional status (level three), 4) general health perceptions (level four), and 5) overall quality of life (level five)[12]. They stated that the concepts of this model (i.e., the concepts encountered when following the levels from left to right across Fig. 1) become increasingly integrated and increasingly difficult to define and measure[12]. Furthermore, they explained that, as one draws closer to level five of the model (the right side of Fig. 1), there are an increasing number of inputs, such as characteristics of the individual (symptom amplification, personality and motivation, and values and preferences), characteristics of the environment (psychological supports, social and economic supports, and social and psychological supports), and nonmedical factors that cannot be controlled by clinicians or by the health-care system as it is traditionally defined[12].

Biological and physiological factors (level one) are commonly conceptualized, measured, and applied in daily clinical practice[12]. Examples include the diagnosis of osteoarthritis, fracture, or osteosarcoma; increased activity on a bone scan; or a meniscal tear seen on a magnetic resonance image.

The focus then moves from specific cells or organs to the organism as a whole when symptoms are assessed (level two)[12]. At this level, both biological and physiological factors play a role. Patients can have physiological abnormalities, such as osteoporosis, without having symptoms. On the other hand, symptoms may not correlate to radiographic findings, such as the lack of radiographic abnormalities in the presence of back pain. Thus, we need to be careful about the way that we interpret measures of biological function, as factors other than biology may influence how a symptom presents. However, symptoms are often the primary reason why patients seek medical attention, and therefore we need to further explore the inconsistent relationship between biological factors and symptoms[12].

Functional status (level three) is also listed as an outcome in the model proposed by Wilson and Cleary. Symptoms, such as pain, have an impact on the functional status of a patient; however, the personality and motivations of the patient are also influential. Additionally, the social environment plays an important role. Functional status is an important outcome to be measured in observational studies. Relationships between symptoms and function clearly exist, but variations in functioning cannot always be fully explained. Therefore, all five levels of outcome need to be measured in an observational study.

General health perceptions (level four) are an assimilation of the previously described levels of outcome. These general health perceptions have been shown to be allied with biological and physiological factors; however, many other factors, including patient expectations, affect a patient's perception of health, resulting in large variations within each stage of clinical severity.

Finally, overall quality of life (level five) is the broadest outcome. This includes a wide range of nonmedical factors as well as experiences and feelings that humans have. Interestingly, functional status was not as strongly associated with happiness as one might expect[12]. It is hypothesized that patients may be able to adapt to their impairment[13].

In contrast to the levels of evidence, a hierarchy in quality does not exist in the outcome levels. A higher level merely illustrates the complexity of the measurement, and the different levels of outcome should coexist in observational studies.

Types of Outcome Instruments

Following the model as described above, we can identify several outcome instruments. Radiographs or other types of imaging and laboratory tests are the outcome measures that focus on biological and physiological factors in orthopaedic research. Symptoms in the orthopaedic patient are frequently pain, stiffness, and loss of strength, and these symptoms are usually captured during physical examination. To structure this examination, clinician-based outcome instruments (such as the Knee Society score) were developed[14]. Furthermore, grip dynamometers, the Jebsen-Taylor Test of Hand Function[15], and the "Get-Up and Go" test[16] are some additional examples of tools that are helpful during the physical examination.

Mixed Clinician-Based and Functional Outcome Instruments

Frequently utilized in past and present orthopaedic research are mixed clinician-based and functional outcome instruments (these instruments are comprised of questions answered by patients and physical examination performed by the outcome assessor). Often, these instruments are not validated and their outcomes are frequently dependent on the technique of the administrator. Range of motion is often a component of mixed scoring systems such as the Harris hip score and the Constant shoulder assessment[17,18]. If physical examination or clinical tests are a part of a scoring system, there is a high risk of interobserver variability[4]. Furthermore, the weight applied to subscales or items that are often spread over different levels of the model by Wilson and Cleary may be arbitrary. In addition, if scores such as pain and hip flexion are combined into one single numerical measure, this single number may yield non-informative conclusions[4,19]. For example, a patient may have an increased range of motion but still have pain or, alternatively, the pain may have subsided but the stiffness still exists. In addition, the patient may have a small amount of improvement in both pain and stiffness scores. Summarizing these findings in a single number is not helpful to investigators or clinicians and will not discriminate between improving and worsening patients.

Apart from mixed clinician-based instruments, other combinations are possible. The biological and physiological variables, symptom status, and functional status levels in the model by Wilson and Cleary are commonly combined. The symptom status, general health perceptions, and overall quality-of-life levels are also possible combinations in mixed outcome instruments. Use of the labels of the level may clarify matters better than use of the term *mixed clinician-based instruments,* as both biological and physiological variables and functional status instruments may be clinician based.

We recommend not using mixed clinical and functional outcome instruments because validated outcome instruments that will yield reproducible results are now available.

System-Specific Outcome Instruments and Disease-Specific Instruments

System-specific instruments are developed to evaluate conditions related to one body region or, more specifically, one joint. A system-specific instrument covers several diseases related to the given body region. For example, several patient-reported outcome instruments are devoted to the knee[20]. Condition or disease-specific instruments evaluate the well-being of patients with a specific disease. For example, a patient population with osteoarthritis of the hip can be evaluated with a system-specific instrument for the hip, but they can also be evaluated with a disease-specific instrument that focuses on osteoarthritis. Another example comes from carpal tunnel syndrome research. The Boston Carpal Tunnel Questionnaire is valid for the population, has good reliability, and is responsive[21]. We encourage the use of a disease-specific instrument, when available, to improve sensitivity to change. Symptom or disease-specific instruments typically have a higher sensitivity than that obtainable with generic quality-of-life instruments. Measurement properties are discussed in more detail below.

General Health-Related Quality-of-Life Instruments

Health-related quality of life is a multifactorial concept[12]. It comprises physical, mental, and social factors[9]. These factors, when combined, describe a person's health and deal with a broad range of daily activities, such as work, hobbies, and social interactions (Wilson and Cleary level four). Thus, health-related quality of life is how a person's health affects his or her ability to carry out normal social and physical activities. General health-related quality-of-life instruments are designed to measure this wide range of health status. In orthopaedics, the Sickness Impact Profile, the Nottingham Health Profile, and the Short Form-36 (SF-36)[22] are most frequently used[4]. More recently, the Short Form-12 (SF-12) has been developed[23]. The SF-12 comes with the advantage of being able to produce the two summary scales (initially developed from the SF-36) with considerable accuracy but with far less respondent burden[24]. Accordingly, the SF-12 may be an outcome instrument of choice for a situation in which a short, generic measure providing summary information on physical and mental health status is mandatory[24]. Moreover, the Musculoskeletal Function Assessment (MFA) instrument and, more recently, the Short Musculoskeletal Function Assessment (SMFA) instrument have been developed[25,26]. These two instruments have demonstrated good validity and are commonly used in musculoskeletal research[25,26].

Overall Quality-of-Life Instruments

Happiness or satisfaction during daily activities or tasks is often as important to individuals as their ability to participate in these daily activities[9,12]. To capture the patients' overall quality of life (Wilson and Cleary level five)[12], quality-of-life outcomes have been developed. An example includes the Quality of Well-Being (QWB) questionnaire[27].

Finding and Selecting an Outcome Instrument

Selecting an outcome instrument starts with the proposed research question[9]. After carefully framing the objective of the study, the next step is to locate instruments most suitable to evaluate the intervention under investigation and the patient

population to be included in the study. One of the most effective methods of selecting an appropriate instrument is to consult with experienced musculoskeletal clinical researchers. Ideally, the selected outcome instruments will evaluate all five levels of outcome as proposed by Wilson and Cleary. For example, if a study is investigating a new surgical device for arthroscopic shoulder stabilization and the prevention of recurrent anterior shoulder dislocation, it should start with a suitable imaging technique, such as magnetic resonance imaging. The outcome instrument should be a system-specific (shoulder) and condition-specific (instability) instrument. Both a general quality-of-life instrument and an overall quality-of-life instrument could be included to help policy makers compare the outcomes of different conditions[28]. However, evaluating the level-five status of a patient according to the classification system of Wilson and Cleary is not deemed necessary when the aim of clinical researchers is to improve clinical practice with regard to a specific condition.

The next step is to conduct a literature search. Consulting a librarian is often beneficial and can improve the efficiency and the results of a literature search. Using the "Clinical Queries" feature in PubMed to find the most up-to-date systematic reviews that evaluated outcome instruments is often helpful. As of this writing, specific search strategies[29], such as those developed for randomized controlled trials, are currently lacking and are undergoing further evaluation. Table I lists the key contemporary systematic reviews that are relevant to musculoskeletal outcome instruments. Once an outcome instrument has been selected from a systematic review, one can apply quality criteria as described below to select the most appropriate instrument. In our example, two reviews could guide us[30,31]. Furthermore, several researchers have published compendiums of instruments to aid in the selection of outcome instruments[32]. Additionally, the use of validated instruments in a daily clinical setting has been shown to improve care. Current guidelines recommend that practitioners who would like to improve patient care should assess patient outcomes with use of a validated condition-specific instrument and focus on the functional aspects of the disease or injury secondarily[25].

Quality Criteria for Outcome Instruments

Any new outcome instrument needs to be developed with use of strict methodological safeguards. Suggested quality criteria are mostly opinion based because there is no empirical evidence in this field to support explicit quality criteria[2,9]. Terwee et al. proposed a checklist of quality criteria to evaluate the methodological soundness of patient-reported outcome instruments[2]. These criteria include: content validity, internal consistency, criterion validity, construct validity, reproducibility (agreement and reliability), responsiveness, floor and ceiling effects, and interpretability[2]. Below, we describe these quality criteria for patient outcome instruments in more detail[2].

Content Validity

Content validity is the extent to which the domain of interest is comprehensively sampled by the items in the instrument[2]. The developers of an outcome instrument should consider that the *measurement aim* of the instrument is important and that the content is tailored to the population of interest. For example, in a study evaluating patients with instability of the knee, a question about return to sports may be relevant. However, another instrument focusing on the knee may have relevant questions with respect to osteoarthritis (i.e., pain or stiffness), but these questions may be less relevant to a patient who has painless giving-way and who is unable to return to sports[33]. Therefore, the investigators have to choose an instrument that is relevant to the aim of the outcome to be measured, given the patient population of interest.

Internal Consistency

Internal consistency is the extent to which items in a scale or subscale are intercorrelated, thus measuring the same construct[2]. Different items or subscales in the instrument can ask the same questions in a slightly different manner to truly capture the respondent's opinion or level of function. The Cronbach alpha is considered an adequate measure of internal consistency, and it should be calculated for each scale or subscale separately[2]. A low Cronbach alpha indicates a lack of correlation between the items in a scale, which makes summarizing the items unjustified. A very high Cronbach alpha reflects high correlations among the items in the scale, which indicate the redundancy of one or more items[2].

Criterion Validity

Criterion validity is the extent to which scores on a particular instrument relate to a so-called gold standard[2]. If a gold standard is available, the outcome instrument can be compared with this standard. However, as a gold standard is frequently unavailable, construct validity has to be assessed.

Construct Validity

Construct validity refers to the extent to which scores on a particular instrument relate to other measures in a manner that is consistent with theoretically derived hypotheses concerning the concepts that are being measured[2]. In their paper on quality criteria, Terwee et al. explained: Without specific a priori hypotheses, for example, about expected correlations between changes in instruments measuring the same concept, or expected differences in changes between "known" groups, "the risk of bias is high because retrospectively it is tempting to think up alternative explanations for low correlations instead of concluding that the questionnaire may not be valid."[2]

Reproducibility

Reproducibility refers to the degree to which repeated measurements (test-retest) in steady populations provide similar answers. Reproducibility is built on agreement and reliability. Agreement is the extent to which the scores on repeated measures are close to each other (absolute measurement error). Reliability is the extent to which patients can be distinguished from each other, despite measurement errors (relative measurement error).

TABLE I Systematic Reviews of Quality-of-Life Instruments in Orthopaedic Research

Topic	Reference
Upper extremity	Dziedzic KS, Thomas E, Hay EM. A systematic search and critical review of measures of disability for use in a population survey of hand osteoarthritis (OA). Osteoarthritis Cartilage. 2005;13:1-12.
	Bot SD, Terwee CB, van der Windt DA, Bouter LM, Dekker J, de Vet HC. Clinimetric evaluation of shoulder disability questionnaires: a systematic review of the literature. Ann Rheum Dis. 2004;63:335-41.
	Michener LA, Leggin BG. A review of self-report scales for the assessment of functional limitation and disability of the shoulder. J Hand Ther. 2001;14:68-76.
	Salerno DF, Copley-Merriman C, Taylor TN, Shinogle J, Schulz RM. A review of functional status measures for workers with upper extremity disorders. Occup Environ Med. 2002;59:664-70.
	Bialocerkowski AE, Grimmer KA, Bain GI. A systematic review of the content and quality of wrist outcome instruments. Int J Qual Health Care. 2000;12:149-57.
	Keskula DR, Lott J. Defining and measuring functional limitations and disability in the athletic shoulder. J Sport Rehabil. 2001;10:221-31.
	Dowrick AS, Gabbe BJ, Williamson OD, Cameron PA. Outcome instruments for the assessment of the upper extremity following trauma: a review. Injury. 2005;36:468-76.
Lower extremity	Drake BG, Callahan CM, Dittus RS, Wright JG. Global rating systems used in assessing knee arthroplasty outcomes. J Arthroplasty. 1994;9:409-17.
	Garratt AM, Brealey S, Gillespie WJ, DAMASK Trial Team. Patient-assessed health instruments for the knee: a structured review. Rheumatology (Oxford). 2004;43:1414-23.
	Terwee CB, Mokkink LB, Steultjens MP, Dekker J. Performance-based methods for measuring the physical function of patients with osteoarthritis of the hip or knee: a systematic review of measurement properties. Rheumatology (Oxford). 2006;45:890-902.
	Haywood KL, Hargreaves J, Lamb SE. Multi-item outcome measures for lateral ligament injury of the ankle: a structured review. J Eval Clin Pract. 2004;10:339-52.
	Eechaute C, Vaes P, Van Aerschot L, Asman S, Duquet W. The clinimetric qualities of patient-assessed instruments for measuring chronic ankle instability: a systematic review. BMC Musculoskelet Disord. 2007;8:6.
Spine	Hallin P, Sullivan M, Kreuter M. Spinal cord injury and quality of life measures: a review of instrument psychometric quality. Spinal Cord. 2000;38:509-23.
	Zanoli G, Strömqvist B, Padua R, Romanini E. Lessons learned searching for a HRQoL instrument to assess the results of treatment in persons with lumbar disorders. Spine. 2000;25:3178-85.
	Grotle M, Brox JI, Vøllestad NK. Concurrent comparison of responsiveness in pain and functional status measurements used for patients with low back pain. Spine. 2004;29:E492-501.
	Moreau CE, Green BN, Johnson CD, Moreau SR. Isometric back extension endurance tests: a review of the literature. J Manipulative Physiol Ther. 2001;24:110-22.
	Kettler A, Wilke HJ. Review of existing grading systems for cervical or lumbar disc and facet joint degeneration. Eur Spine J. 2006;15:705-18.
	Costa LO, Maher CG, Latimer J. Self-report outcome measures for low back pain: searching for international cross-cultural adaptations. Spine. 2007;32:1028-37.
General osteoarthritis	Veenhof C, Bijlsma JW, van den Ende CH, van Dijk GM, Pisters MF, Dekker J. Psychometric evaluation of osteoarthritis questionnaires: a systematic review of the literature. Arthritis Rheum. 2006;55:480-92.
	Sun Y, Stürmer T, Günther KP, Brenner H. Reliability and validity of clinical outcome measurements of osteoarthritis of the hip and knee—a review of the literature. Clin Rheumatol. 1997;16:185-98.

Responsiveness

Responsiveness is the ability of an instrument to detect clinically important changes over time[2]. Terwee et al. offered further clarification of this concept when they said: "We consider responsiveness to be a measure of longitudinal validity. In analogy to construct validity, longitudinal validity should be assessed by testing predefined hypotheses, e.g., about expected correlations between changes in measures, or expected differences in changes between 'known' groups. This shows the ability of a questionnaire to measure changes if they really have happened. Furthermore, the instrument should be able to distinguish clinically important change from measurement error."[2]

Floor and Ceiling Effects

Floor and ceiling effects describe the number of respondents who achieved the lowest or highest possible score[2]. If a patient

scores the highest possible score (that is, best health status) preoperatively, the outcome instrument cannot measure improvement after the treatment. This ceiling effect may be observed when outcome instruments that were designed to evaluate patients with severe disease are used to evaluate patients who are more fit. For example, an instrument designed to evaluate elderly patients with osteoarthritis is less suitable for evaluating improvement after treatment in a group of athletes. This example stresses the importance of choosing an instrument according to the population for which it was originally designed.

Interpretability

Interpretability is the degree to which one can assign qualitative meaning to quantitative scores[2]. A difference of five points between two treatment groups on a scale ranging from 0 to 100 points may be significant, but this does not mean that such a small difference is also clinically relevant. An instrument with a high degree of reproducibility may be able to detect small changes[34]; however, investigators should decide a priori on the clinical relevance of the spectrum of that scale. On a 100-point scale, are patients with a 5-point higher score really experiencing an important change, or is a change in health status as reflected in a difference of 20 points the minimally important change[35,36]? This raises an important issue, especially since calculations regarding sample size are often performed on the best estimates of "important treatment effects." Not only is the choice of a primary end point important, but the relevance of change on a scale should also be known[35,36].

Pitfalls Involved with the Use of Outcome Instruments in Orthopaedic Research

Other Methodological Considerations

If multiple outcome instruments are used in an observational study, then positive results can be an effect of chance alone[37]. The occurrence of an erroneous false-positive conclusion is designated as a type-I or alpha error. Therefore, investigators should carefully choose the most appropriate outcome instrument a priori. We advise researchers to consider choosing an instrument for each outcome level unless the intervention is specifically focused on one or two levels. For example, if the intervention is aimed to help patients cope with their symptoms, it would not be necessary to make use of the Wilson and Cleary classification system to measure biological and physiological variables. The advice to cover several levels will result in multiple end points. Five well-defined outcome measures will strengthen a study, thus covering all aspects of orthopaedic outcome measurement. It is important, however, that the investigators define their primary outcome of interest, as multiple end points increase the risk of having false-positive results[37]. If multiple outcomes are used, a statistical correction, (that is, the Bonferroni method) is recommended[37,38]. For instance, if an investigator plans to conduct five tests of significance on five different outcome measures and report significance at the $p < 0.05$ level, the effective probability value is not 0.05 but rather can be approximated by a rule of thumb ($5 \text{ outcomes} \times 0.05 = 0.25$). Using this rule of thumb, there is a 25% chance of having a false-positive result among the five outcome measures. To limit such incorrect conclusions (alpha errors), the alpha level of significance can be adjusted from 0.05 to 0.01 ($0.05 \div 5 \text{ outcomes} = 0.01$)[37]. Data dredging (this process refers to multiple testing until one test shows positive results without the use of an a priori hypothesis) can be a problem in retrospective case series, in which protocols can be adjusted to the investigator's convenience after analyzing the data, thus resulting in flawed conclusions.

Another hazard exists when only the parts of the instrument that show a significant result after statistical testing are reported, not the complete instrument. Instruments are designed to be used in their original form, since the instrument was validated as such. Modification of validated outcome instruments without revalidation may result in flawed conclusions[39,40]. After modification of an outcome instrument, the instrument should undergo the same rigorous validation process as the original instrument did[39]. Simply adding or removing questions for one's own convenience violates the validity of an outcome instrument and is strongly discouraged. Orthopaedic research will have more impact if a core set of unmodified outcome instruments are used in all observational studies.

Application of an outcome instrument needs to be accomplished by blinded or independent study personnel to help reduce bias. Although it is the patient who provides the outcome that is assessed, there is always a risk that an interviewer may have guided the responses of the patient[39]. Unblinded outcome assessment may yield a threefold increase in treatment effect[39]. In addition, if patients believe that the information that they give on a questionnaire about symptoms or function will be reported to their treating surgeon, they often will not provide accurate information because they may appreciate the surgeon's efforts to help them and because they may not want to disappoint. This is why it is so critical to use blinding techniques or enlist the services of someone who is independent of the treating team to administer the instrument.

Cultural and Linguistic Considerations for a Specific Population

A patient from the United States may respond differently to certain questions than a patient from the United Kingdom would, even though both are native English speakers[4]. Depending on the geographic locations involved, words can have subtly different meanings. Therefore, it is very important to consider the population on which the outcome instrument was originally tested[41].

Language barriers are another difficulty in clinimetrics (a methodological discipline that focuses on the quality of clinical measurements, e.g., diagnostic characteristics and disease outcomes). Simply translating a previously validated instrument is insufficient[42]. The translated instrument needs to be

TABLE II Modes of Administration of Outcome Instruments[9]*

Mode of Administration	Advantages	Disadvantages
Interviewer	Maximal response rate	Costly
	Can clarify questions	Interviewer bias
	Higher completion rate	Reporting bias
	Control over who is the respondent	Characteristics of the interviewer (voice inflections, age, race, or gender) may introduce bias
	Control over the order of questions	
Telephone	Greater response rate than that obtained with mail-out	Excludes those without access to a telephone
	Relatively inexpensive	Voice inflections of the interviewer may introduce bias
	Relatively quick data collection	
	Interviewer can probe for incomplete answers	
	Data collector can get clarification for ambiguous answers	
Mail	Relatively inexpensive	Response rates generally low
	No bias introduced through the interviewer	Possibility of bias because of nonresponse
	May reach more respondents	No control over who is the respondent
	Respondents can take time to locate certain information	May misunderstand the question
		May miss questions (incomplete)
		Questionnaire may be lost in the mail
		Excludes illiterate, less-educated, disabled, and/or non-English speaking populations
Computer-based (including interactive voice-response technology)	Consistent presentation	Demands subject sit or stand in front of a screen
	Prompts for omissions	Demands some computer skills
	Can be web based	
	Reliable scoring	
	Easy transfer to database	
Self	Maximal response rate	May misunderstand the question
	Inexpensive	May miss questions (incomplete)
Proxy	Can collect information on patients who otherwise are not represented	Response may differ from that of target respondent

*Reproduced, with modification, from: Jackowski D, Guyatt G. A guide to health measurement. Clin Orthop Relat Res. 2003;413:86. Reprinted with permission of Wolters Kluwer Health and Lippincott Williams and Wilkins.

revalidated before it can be utilized in clinical or research settings. The Disabilities of the Arm, Shoulder and Hand (DASH) outcome measure is an example of a validated instrument that has been translated into several languages[43-46]. Other translated and validated instruments focus on hip and knee osteoarthritis (e.g., the Hip Disability and Osteoarthritis Outcome Score [HOOS][47,48], the Knee Injury and Osteoarthritis Outcome Score [KOOS][49], and the Oxford 12-Item Knee Questionnaire[50]). These issues are becoming increasingly important for multicenter cohort studies.

Mental conditions, such as posttraumatic conditions, may complicate the ability of patients to fill out instruments. Injury may also physically impair the patient's hand function and thus the patient's ability to respond to a mailed instrument requiring a written response. Jackowski and Guyatt summarized methods to administer outcome instruments (Table II)[9].

The Future of Outcome Measurement

In the future, the development of a computer-based outcome assessment that incorporates the item response theory may bring much progress[51], but, until then, the standardization of outcome instrument use in orthopaedic research will be a giant leap forward. Ideally, a core set of validated patient-reported outcome instruments should be used in orthopaedic research. We should take a lesson from rheumatology research, in which standardization of outcome measurement has brought much success[52]. Additionally, pain

researchers have put substantial efforts into producing consensus outcome measures[53]. Reaching consensus is not an easy process, however, and the results from consensus meetings need to be subject to ongoing debate[54]. Still, investigators could use these core instruments and then choose additional outcome instruments, if needed, to suit the unique nature of their study. The use of universal, previously validated outcome instruments will facilitate a comparison of results of different studies and will also facilitate subsequent meta-analysis[55]. Before embarking on the development of yet another outcome instrument, investigators should perform a systematic review to evaluate the current state of the art[56]. In terms of improving patient care, there is much more to be gained from utilizing the currently available instruments in clinical outcomes research than from developing a new instrument.

Summary

Validated outcome instruments are a useful tool in orthopaedic research. Unfortunately, not all instruments have been developed with use of strict quality criteria. The conceptual model of patient outcomes that was proposed by Wilson and Cleary can help in choosing an instrument to measure all five levels of outcome[15]. The proposed criteria of Terwee et al. may help in future validation of outcome instruments[2]. As the number of validated outcome instruments continues to increase, a method of standardization, based on evidence and international consensus, could be helpful in future orthopaedic research and may improve the quality of patient care. ■

NOTE: The authors thank Daniel Vena for the editing and proofreading of this paper.

Rudolf W. Poolman, MD, PhD
Department of Orthopaedic Surgery, Onze Lieve Vrouwe Gasthuis, Postbus 95500, 1090 HM Amsterdam, The Netherlands.
E-mail address: poolman@trauma.nl

Marc F. Swiontkowski, MD
Department of Orthopaedic Surgery, University of Minnesota Medical Center, 2450 Riverside Avenue South, Minneapolis, MN 55454

Jeremy C.T. Fairbank, MD, FRCS
Department of Orthopaedic Surgery, Nuffield Orthopaedic Centre, Windmill Road, Headington, Oxford OX3 7LD, United Kingdom

Emil H. Schemitsch, MD, FRCSC
Division of Orthopaedics, Department of Surgery, St. Michael's Hospital, 55 Queen Street East, Suite 800, Toronto, ON M5C 1R6, Canada

Sheila Sprague, MSc
Department of Clinical Epidemiology and Biostatistics, McMaster University, 293 Wellington Street North, Suite 110, Hamilton, ON L8L 8E7, Canada

Henrica C.W. de Vet, PhD
EMGO Institute, VU University Medical Center, Van der Boechorststraat 7, 1081 BT Amsterdam, The Netherlands

References

1. Swiontkowski MF, Buckwalter JA, Keller RB, Haralson R. The outcomes movement in orthopaedic surgery: where we are and where we should go. J Bone Joint Surg Am. 1999;81:732-40.

2. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, Bouter LM, de Vet HC. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60:34-42.

3. Pynsent PB. Choosing an outcome measure. J Bone Joint Surg Br. 2001;83:792-4.

4. Pynsent PB, Fairbank JCT, Carr AJ, editors. Outcome measures in orthopaedics and orthopaedic trauma. 2nd ed. New York: Oxford University Press; 2004.

5. Assessing health status and quality-of-life instruments: attributes and review criteria. Qual Life Res. 2002;11:193-205.

6. Marshall S, Haywood K, Fitzpatrick R. Impact of patient-reported outcome measures on routine practice: a structured review. J Eval Clin Pract. 2006;12:559-68.

7. Poolman RW, Hanel DP, Mann FA, Ponsen KJ, Marti RK, Roolker L. Trans-Atlantic hospital agreement in reading first day radiographs of clinically suspected scaphoid fractures. Arch Orthop Trauma Surg. 2002;122:373-8.

8. Bhandari M, Guyatt GH, Swiontkowski MF, Tornetta P 3rd, Sprague S, Schemitsch EH. A lack of consensus in the assessment of fracture healing among orthopaedic surgeons. J OrthopTrauma. 2002;16:562-6.

9. Jackowski D, Guyatt G. A guide to health measurement. Clin Orthop Relat Res. 2003;413:80-9.

10. Boutron I, Tubach F, Giraudeau B, Ravaud P. Blinding was judged more difficult to achieve and maintain in nonpharmacologic than pharmacologic trials. J Clin Epidemiol. 2004;57:543-50.

11. Suk M, Hanson BP, Norvell DC, Helfet DL. The AO handbook of musculoskeletal outcomes measures and instruments. Davos, Switzerland: Thieme; 2005.

12. Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. JAMA. 1995;273:59-65.

13. Johnson SR, Archibald A, Davis AM, Badley E, Wright JG, Hawker GA. Is self-reported improvement in osteoarthritis pain and disability reflected in objective measures? J Rheumatol. 2007;34:159-64.

14. Insall JN, Dorr LD, Scott RD, Scott WN. Rationale of the Knee Society clinical rating system. Clin Orthop Relat Res. 1989;248:13-4.

15. Jebsen RH, Taylor N, Trieschmann RB, Trotter MJ, Howard LA. An objective and standardized test of hand function. Arch Phys Med Rehabil. 1969;50:311-9.

16. Mathias S, Nayak US, Isaacs B. Balance in elderly patients: the ''get-up and go'' test. Arch Phys Med Rehabil. 1986;67:387-9.

17. Harris WH. Traumatic arthritis of the hip after dislocation and acetabular fractures: treatment by mold arthroplasty. An end-result study using a new method of result evaluation. J Bone Joint Surg Am. 1969;51:737-55.

18. Constant CR, Murley AH. A clinical method of functional assessment of the shoulder. Clin Orthop Relat Res. 1987;214:160-4.

19. Bryant MJ, Kernohan WG, Nixon JR, Mollan RA. A statistical analysis of hip scores. J Bone Joint Surg Br. 1993;75:705-9.

20. Garratt AM, Brealey S, Gillespie WJ; DAMASK Trial Team. Patient-assessed health instruments for the knee: a structured review. Rheumatology (Oxford). 2004;43:1414-23.

21. Jerosch-Herold C, Leite JC, Song F. A systematic review of outcomes assessed in randomized controlled trials of surgical interventions for carpal tunnel syndrome using the International Classification of Functioning, Disability and Health (ICF) as a reference tool. BMC Musculoskelet Disord. 2006;7:96.

22. Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care. 1992;30:473-83.

23. Ware JE Jr, Kosinski M, Keller SD. A 12-item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. Med Care. 1996;34:220-33.

24. Jenkinson C, Layte R. Development and testing of the UK SF-12 (short form health survey). J Health Serv Res Policy. 1997;2:14-8.

25. Barei DP, Agel J, Swiontkowski MF. Current utilization, interpretation, and recommendations: the musculoskeletal function assessments (MFA/SMFA). J Orthop Trauma. 2007;21:738-42.

26. Swiontkowski MF, Engelberg R, Martin DP, Agel J. Short musculoskeletal function assessment questionnaire: validity, reliability, and responsiveness. J Bone Joint Surg Am. 1999;81:1245-60.

27. Kaplan RM, Alcaraz JE, Anderson JP, Weisman M. Quality-adjusted life years lost to arthritis: effects of gender, race, and social class. Arthritis Care Res. 1996;9:473-82.

28. Dijksman LM, Poolman RW, Bhandari M, Goeree R, Tarride JE; International Evidence-Based Orthopaedic Surgery Working Group. Money matters: what to look for in an economic analysis. Acta Orthop. 2008;79:1-11.

29. Haynes RB, McKibbon KA, Wilczynski NL, Walter SD, Werre SR; Hedges Team. Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey. BMJ. 2005;330:1179.

30. Bot SD, Terwee CB, van der Windt DA, Bouter LM, Dekker J, de Vet HC. Clinimetric evaluation of shoulder disability questionnaires: a systematic review of the literature. Ann Rheum Dis. 2004;63:335-41.

31. Michener LA, Leggin BG. A review of self-report scales for the assessment of functional limitation and disability of the shoulder. J Hand Ther. 2001;14:68-76.

32. Agel J, Swiontkowski MF. Guide to outcomes instruments for musculoskeletal trauma research. J Orthop Trauma. 2006;20(8 Suppl):S1-146.

33. Zarins B. Are validated questionnaires valid? J Bone Joint Surg Am. 2005; 87:1671-2.

34. Sierevelt IN, van Oldenrijk J, Poolman RW. Is statistical significance clinically important?—A guide to judge the clinical relevance of study findings. J Long Term Eff Med Implants. 2007;17:173-9.

35. de Vet HC, Terwee CB, Ostelo RW, Beckerman H, Knol DL, Bouter LM. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. Health Qual Life Outcomes. 2006;4:54.

36. Poolman RW, Kerkhoffs GM, Struijs PA, Bhandari M, International Evidence-Based Orthopedic Surgery Working Group. Don't be misled by the orthopedic literature: tips for critical appraisal. Acta Orthop. 2007;78:162-71.

37. Bhandari M, Whang W, Kuo JC, Devereaux PJ, Sprague S, Tornetta P 3rd. The risk of false-positive results in orthopaedic surgical trials. Clin Orthop Relat Res. 2003;413:63-9.

38. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ. 1995;310:170.

39. Poolman RW, Struijs PA, Krips R, Sierevelt IN, Marti RK, Farrokhyar F, Bhandari M. Reporting of outcomes in orthopaedic randomized trials: does blinding of outcome assessors matter? J Bone Joint Surg Am. 2007;89:550-8.

40. Harvie P, Pollard TC, Chennagiri RJ, Carr AJ. The use of outcome scores in surgery of the shoulder. J Bone Joint Surg Br. 2005;87:151-4.

41. Johnson JA, Luo N, Shaw JW, Kind P, Coons SJ. Valuations of EQ-5D health states: are the United States and United Kingdom different? Med Care. 2005;43:221-8.

42. McKenna SP, Doward LC. The translation and cultural adaptation of patient-reported outcome measures. Value Health. 2005;8:89-91.

43. Veehof MM, Sleegers EJ, van Veldhoven NH, Schuurman AH, van Meeteren NL. Psychometric qualities of the Dutch language version of the Disabilities of the Arm, Shoulder, and Hand questionnaire (DASH-DLV). J Hand Ther. 2002;15:347-54.

44. Dubert T, Voche P, Dumontier C, Dinh A. [The DASH questionnaire. French translation of a trans-cultural adaptation]. Chir Main. 2001;20:294-302. French.

45. Rosales RS, Delgado EB, Diez de la Lastra-Bosch I. Evaluation of the Spanish version of the DASH and carpal tunnel syndrome health-related quality-of-life instruments: cross-cultural adaptation process and reliability. J Hand Surg [Am]. 2002;27:334-43.

46. Offenbacher M, Ewert T, Sangha O, Stucki G. Validation of a German version of the 'Disabilities of Arm, Shoulder and Hand' questionnaire (DASH-G). Z Rheumatol. 2003;62:168-77.

47. de Groot IB, Reijman M, Terwee CB, Bierma-Zeinstra SM, Favejee M, Roos EM, Verhaar JA. Validation of the Dutch version of the Hip Disability and Osteoarthritis Outcome Score. Osteoarthritis Cartilage. 2007;15:104-9.

48. van Oldenrijk J, Sierevelt IN, Haverkamp D, Harmse IW, Poolman RW. Re: validation of the Dutch version of the Hip Disability and Osteoarthritis Outcome Score (HOOS). Osteoarthritis Cartilage. 2009;17:133-4.

49. de Groot IB, Favejee MM, Reijman M, Verhaar JA, Terwee CB. The Dutch version of the Knee Injury and Osteoarthritis Outcome Score: a validation study. Health Qual Life Outcomes. 2008;26:6-16.

50. Haverkamp D, Breugem SJ, Sierevelt IN, Blankevoort L, van Dijk CN. Translation and validation of the Dutch version of the Oxford 12-item knee questionnaire for knee arthroplasty. Acta Orthop. 2005;76:347-52.

51. Fliege H, Becker J, Walter OB, Bjorner JB, Klapp BF, Rose M. Development of a computer-adaptive test for depression (D-CAT). Qual Life Res. 2005;14:2277-91.

52. Singh JA, Solomon DH, Dougados M, Felson D, Hawker G, Katz P, Paulus H, Siegel J, Wallace C. Contributions of OMERACT to rheumatic disease research. Arthritis Rheum. 2007;57:186.

53. Dworkin RH, Turk DC, Farrar JT, Haythornthwaite JA, Jensen MP, Katz NP, Kerns RD, Stucki G, Allen RR, Bellamy N, Carr DB, Chandler J, Cowan P, Dionne R, Galer BS, Hertz S, Jadad AR, Kramer LD, Manning DC, Martin S, McCormick CG, McDermott MP, McGrath P, Quessy S, Rappaport BA, Robbins W, Robinson JP, Rothman M, Royal MA, Simon L, Stauffer JW, Stein W, Tollett J, Wernicke J, Witter J; IMMPACT. Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. Pain. 2005;113:9-19.

54. McQuay H. Consensus on outcome measures for chronic pain trials. Pain. 2005;113:1-2.

55. Bombardier C. Outcome assessments in the evaluation of treatment of spinal disorders: summary and general recommendations. Spine. 2000;25:3100-3.

56. de Vet HC, Terwee CB, Bouter LM. Current challenges in clinimetrics. J Clin Epidemiol. 2003;56:1137-41.