Qual Life Res
DOI 10.1007/s11136-012-0344-y

ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research

Bryce B. Reeve · Kathleen W. Wyrwich · Albert W. Wu · Galina Velikova · Caroline B. Terwee · Claire F. Snyder · Carolyn Schwartz · Dennis A. Revicki · Carol M. Moinpour · Lori D. McLeod · Jessica C. Lyons · William R. Lenderking · Pamela S. Hinds · Ron D. Hays · Joanne Greenhalgh · Richard Gershon · David Feeny · Peter M. Fayers · David Cella · Michael Brundage · Sara Ahmed · Neil K. Aaronson · Zeeshan Butt

Accepted: 17 December 2012

Abstract
Purpose An essential aspect of patient-centered outcomes research (PCOR) and comparative effectiveness research (CER) is the integration of patient perspectives and experiences with clinical data to evaluate interventions. Thus, PCOR and CER require capturing patient-reported outcome (PRO) data appropriately to inform research, healthcare delivery, and policy. This initiative's goal was to identify minimum standards for the design and selection of a PRO measure for use in PCOR and CER.
Methods We performed a literature review to find existing guidelines for the selection of PRO measures. We also conducted an online survey of the International Society for Quality of Life Research (ISOQOL) membership to solicit input on PRO standards. A standard was designated as "recommended" when >50 % respondents endorsed it as "required as a minimum standard."
Results The literature review identified 387 articles. Survey response rate was 120 of 506 ISOQOL members. The respondents had an average of 15 years experience in PRO research, and 89 % felt competent or very competent providing feedback. Final recommendations for PRO measure standards included: documentation of the conceptual and measurement model; evidence for reliability, validity (content validity, construct validity, responsiveness); interpretability of scores; quality translation, and acceptable patient and investigator burden.
Conclusion The development of these minimum measurement standards is intended to promote the appropriate use of PRO measures to inform PCOR and CER, which in turn can improve the effectiveness and efficiency of healthcare delivery. A next step is to expand these

This study was conducted on behalf of the International Society for Quality of Life Research (ISOQOL).

Electronic supplementary material The online version of this article (doi:10.1007/s11136-012-0344-y) contains supplementary material, which is available to authorized users.

B. B. Reeve (✉)
Department of Health Policy and Management, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, 1101-D McGavran-Greenberg Building, 135 Dauer Drive, CB 7411, Chapel Hill, NC 27599-7411, USA
e-mail: bbreeve@email.UNC.edu

B. B. Reeve · J. C. Lyons
Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

K. W. Wyrwich · D. A. Revicki · W. R. Lenderking
United BioSource Corporation, Bethesda, MD, USA

A. W. Wu · C. F. Snyder
Johns Hopkins School of Medicine, Baltimore, MD, USA

G. Velikova · J. Greenhalgh
University of Leeds, Leeds, UK

C. B. Terwee
VU University Medical Center, Amsterdam, The Netherlands

C. Schwartz
DeltaQuest Foundation, Inc., Concord, MA, USA

C. M. Moinpour
Fred Hutchinson Cancer Research Center, Seattle, WA, USA

L. D. McLeod
Research Triangle Institute Health Solutions, Durham, NC, USA

Published online: 04 January 2013

minimum standards to identify best practices for selecting decision-relevant PRO measures.

Keywords Patient-reported outcomes · Comparative effectiveness · Patient-centered outcomes research · Psychometrics · Questionnaire

Introduction

An essential aspect of patient-centered outcomes research (PCOR) and comparative effectiveness research (CER) is the integration of patients' perspectives about their health with clinical and biological data to evaluate the safety and effectiveness of interventions. Such integration recognizes that health-related quality of life (HRQOL) and how it is affected by disease and treatment complements traditional clinical endpoints such as survival or tumor response in cancer. For HRQOL endpoints, it is widely accepted that the patient's report is the best source of information about what he or she is experiencing. The challenge for PCOR and CER is how to best capture patient-reported data in a way that can inform decision making in healthcare delivery, research, and policy settings.

Observational and experimental studies have increasingly included patient-reported outcome (PRO) measures, defined by the Food and Drug Administration (FDA) as "any report of the status of a patient's health condition that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else [1]." Patients can report accurately on a number of domains that are important for evaluating an intervention or disease burden, including symptom experiences (e.g., pain, fatigue, nausea), functional status (e.g., sexual, bowel, or urinary functioning), well-being (e.g., physical, mental, social), quality of life, and satisfaction with care or with a treatment [1–4]. Arguably, patients are the gold standard source of information for assessing such domains. To draw valid research conclusions regarding patient-centered outcomes, PROs must be measured in a standardized way using scales that demonstrate sufficiently robust measurement properties [4–9].

The goal of this study was to identify minimum standards for the selection of PRO measures for use in PCOR and CER. We defined minimum standards such that if a PRO measure did not meet these criteria, it would be judged not suitable for a PCOR study. A central aim in developing this set of standards was to clearly define the critical attributes for judging a PRO measure for a PCOR study. We identified these standards using two complementary approaches. The first was an extensive review of the literature including both published and unpublished guidance documents. The second was to seek input, via a formal survey, from an international group of experts in PRO measurement and PCOR who are members of the International Society for Quality of Life Research (ISOQOL) [10]. Although not the primary objective of this study, our approach allowed us to also identify criteria that were not deemed as a necessary minimum standard, but would rather be considered "best practice" standards for PRO measures.

Identification of minimal standards is a first step toward enabling PCOR and CER to achieve their goals of enhancing healthcare delivery and ultimately improving patients' health and well-being. Access to scientifically sound and decision-relevant PRO measures will allow investigators to collect empirical evidence on the differential benefits of interventions from the patients' perspective [6, 9, 11, 12]. This information can then be disseminated to patients, providers, and policy makers to provide a richer perspective on the impact of interventions on patients' lives using endpoints that are meaningful to them [13].

P. S. Hinds
Children's National Medical Center, Washington, DC, USA

P. S. Hinds
The George Washington University School of Medicine, Washington, DC, USA

R. D. Hays
David Geffen School of Medicine at UCLA, Los Angeles, CA, USA

R. Gershon · D. Cella · Z. Butt
Northwestern University Feinberg School of Medicine, Chicago, IL, USA

D. Feeny
University of Alberta, Alberta, Canada

P. M. Fayers
University of Aberdeen, Aberdeen, UK

P. M. Fayers
Norwegian University of Science and Technology (NTNU), Trondheim, Norway

M. Brundage
Queen's University, Kingston, ON, Canada

S. Ahmed
McGill University, Montreal, QC, Canada

N. K. Aaronson
The Netherlands Cancer Institute, Amsterdam, The Netherlands

N. K. Aaronson
University of Amsterdam, Amsterdam, The Netherlands

Methods

This paper is based on a study funded by the U.S. Patient-Centered Outcomes Research Institute (PCORI) [14]. The paper does not represent PCORI's Methodology Committee standards, issued separately by PCORI, though some of those standards were informed by this work [15]. An ISOQOL scientific advisory task force (SATF), consisting of the authors on this article, was set up to guide the drafting and final selection of recommended standards. We conducted a literature review that helped the SATF draft the recommendations that were subsequently reviewed by ISOQOL members in the formal survey. The literature review and the responses and feedback from ISOQOL members informed the final recommendations provided in this article.

Literature review

We conducted a systematic review of the published and unpublished literature to identify existing guidance documents related to PRO measures. The review identified current practices in selecting PRO measures in PCOR and CER, relevant scale attributes (e.g., reliability, validity, response burden, interpretability), and use of qualitative and quantitative methods to assess these properties. We focused on consensus statements, guidelines, and evidence-based papers, with an emphasis on articles or documents that described broadly generalizable principles. However, some papers that were population- or instrument-specific were included because of the rigor of the psychometric methods.

For the literature review, we adapted a published MEDLINE search strategy to identify measurement properties of PRO measures [16]. The published strategy was used as a foundation and adapted by using terms from MEDLINE thesaurus, Medical Subject Headings (MeSH), and the American Psychological Association's (APA) online Thesaurus of Psychological terms. We conducted parallel searches in several relevant electronic databases, including MEDLINE, PsycINFO, and Combined Index to Nursing and Allied Health Literature (CINAHL) (see database search terms in Appendix 1, ESM). There was no a priori restriction by publication date or age of sample. We also obtained relevant articles through a request to the ISOQOL membership email distribution list.

The titles and abstracts of identified articles and guidelines were reviewed by one of the authors (ZB). The full text of relevant articles was obtained and reviewed. The references cited in the selected articles were reviewed to identify additional relevant articles. ZB abstracted the necessary information for the study; two other authors (DC and RG) independently reviewed several of the articles to ensure coding consistency.

Based on PRO measurement standards gleaned from the literature review, the ISOQOL SATF drafted recommendations that were reviewed by ISOQOL members in a survey described below. Through an iterative series of SATF e-mails and conference calls, the potential standards identified by the systematic literature review were discussed and debated. Redundancies between potential standards were minimized, and similar items consolidated. Where there were differences in opinion among the members, different options were retained in the survey in order that the membership at large could rate and comment on each potential standard. The resultant survey consisted of 23 potential minimum standards to be rated by the ISOQOL membership.

Survey of ISOQOL membership

ISOQOL is dedicated to advancing the scientific study of HRQOL and other patient-centered outcomes to identify effective interventions, enhance the quality of healthcare and promote the health of populations [10]. Since 1993, ISOQOL has been an international collaborative network including researchers, clinicians, patient advocates, government scientists, industry representatives, and policy makers. Many ISOQOL members are PRO methodologists who focus on using state-of-the-art methods, both qualitative and quantitative, to improve the measurement and application of patient-reported data in research, healthcare delivery, and population surveillance. Many of the PRO measures widely used in research as well as the guidelines for developing and evaluating a PRO measure were developed by ISOQOL members. At the time of the survey, there were 506 ISOQOL members on the email distribution list.

In the web-based survey, we sought ISOQOL members' views on draft minimum standards, paying particular attention to areas where there did not appear to be consensus in the literature. For example, we asked ISOQOL members to rank the relative importance of various approaches for assessing reliability, including test–retest and internal consistency for multi-item PRO measures. In addition, we sought agreement on recommendations for six key attributes of PRO measures: (1) conceptual and measurement model, (2) reliability, (3) validity, (4) interpretability of scores, (5) translation, and (6) patient and investigator burden.

In the survey, it was deemed critical that respondents had a clear definition of a minimum standard. The second screen of the survey provided this guidance: "Please remember as you answer the questions in this survey that we are developing the minimum standards for the selection and design of a PRO measure for use in patient-centered outcomes research (PCOR). That is, we are saying a PRO measure that does not meet the *minimum standard* should not be considered appropriate for the research study." This statement was not intended to suggest that a PRO measure would not continue to be validated and strengthened as part of a maturation model of development. The survey directly mentioned PCOR, but the SATF believes these recommendations

are consistent for CER. For brevity, we use just "PCOR" in describing the results.

For each recommendation created by the SATF's synthesis of the literature review, the participant could select one of the following response options: required as a minimum standard, desirable but not required as a minimum standard, not required at all (not needed for a PRO measure), not sure, or no opinion. In analyzing the results, we used the general rule that if 50 % or more agreed that the recommendation was required as a minimum standard, then the recommendation was accepted. If less than 50 % of respondents were in agreement, then the recommendation was reviewed by the ISOQOL SATF to determine whether the recommendation may have been unclear or whether it would better be considered as a "best practice" (or "ideal standard") for PRO measures rather than a "minimum standard." Respondents were also encouraged to comment using a free text box that was provided after each recommendation. This text was extracted from the survey and helped inform the ISOQOL SATF's decisions and final recommendations.

The survey and a description of the survey methodology were submitted to the Institutional Review Board (IRB) at the University of North Carolina at Chapel Hill (UNC) for review and were determined to be exempt from IRB approval by the UNC Office of Human Research and Ethics. The online survey was designed and administered using the Qualtrics Software System under the UNC site license [17].

The survey link was sent out through the ISOQOL member email distribution list ($n = 506$) on 20 February , 2012. Survey instructions asked members to complete the survey within 9 days to meet deadlines for the PCORI contract. However, the response interval was extended to 20 March , 2012 (29 days), to accommodate more ISOQOL respondents. Information about the purpose of the voluntary survey, goals of the project, and funding source was included. All responses were anonymous, and no personal identifying information was collected. Two reminders were sent during the period the survey was available.

We did not expect responses from all ISOQOL members, because: (1) the survey was specifically aimed at those ISOQOL members who considered themselves to have the requisite expertise in the area of PRO measurement, and (2) we sought expert input in a short amount of time. Although we did not limit eligibility to those members who had such expertise, we did ask respondents to self-report their expertise level as part of the survey.

Results

Guidance identified through the literature review

A number of well-known guidance documents were identified, including guidance from the FDA [1, 18–20]; the 2002 Medical Outcomes Trust guidelines on attributes of a good HRQOL measure [2]; the extensive, international expert-driven recommendations from COSMIN (COnsensus-based Standards for the selection of health Measurement INstruments) [3, 4, 21–25]; the European Organization for Research and Treatment of Cancer (EORTC) guidelines for developing questionnaires [26]; the Functional Assessment of Chronic Illness Therapy (FACIT) approach [27]; the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) task force recommendation documents [28–31]; the American Psychological Association (APA) Standards for Educational and Psychological Testing [32]; and several others [33–38]. We also had access to the recent standards documents just completed by the National Institutes of Health's Patient-Reported Outcomes Measurement Information System® (PROMIS®) network, which we considered useful for informing the minimal standards for PRO measures. In addition, ISOQOL recently completed two guidance documents relevant for this landscape review on the use of PRO measures in comparative effectiveness research and on integrating PRO measures in healthcare delivery settings [5, 39].

ISOQOL members identified a total of 301 additional references relevant for our task. Our formal search of the MEDLINE database yielded 821 references, which were individually reviewed, resulting in 60 additional relevant articles. Review of the 172 potentially relevant PsycINFO results provided 22 additional relevant articles, and an additional four unique references were uncovered after review of 126 abstracts identified through CINAHL.

Table 1 describes 28 key guidance documents identified from the literature review that helped to inform the ISOQOL SATF's draft minimum guidelines to be evaluated in the ISOQOL survey. The documents selected for further review and discussion by our ISOQOL SATF represented exemplar description of guidelines and standards for the selection of PRO in PCOR. As part of our literature review, we identified many more relevant references; however, our focus was on existing guidance documents that had broad relevance. Multiple publications describing the same set of guidelines were not cited separately.

Characteristics of participants responding to the ISOQOL survey

Table 2 summarizes the characteristics of the 120 ISOQOL members (23.7 %) who responded to the survey. Approximately 64 % of the sample had a PhD (or similar doctoral degree) and 18 % had a MD. The sample included 68 % academic researchers, 21 % clinicians, 8 % industry representatives, 23 % industry consultants, and 6 % federal government employees. There was diverse geographic distribution with 48 % of respondents from North America

Table 1 Identified guidelines for patient-reported outcomes measures

Author, year	Guideline	Research design	Description
Acquadro et al. [48]	The Literature review of methods to translate health-related quality of life questionnaires for use in multinational clinical trials	Formal literature review	Call for more empirical research on translation methodology; reviews several existing guidelines; advocates multistep process for translations
Cella [27]	Manual for the Functional Assessment of Chronic Illness Therapy (FACIT)	Description of method	Provides summary of FACIT scale development and translation methodologies; presents basic psychometric info for existing measures
Coons et al. [28]	Recommendations on evidence needed to support measurement equivalence between electronic and paper-based patient-reported outcome measures	Expert opinion and literature review	Provides a general framework for decisions regarding evidence needed to support migration of paper PRO measures to electronic delivery
COSMIN group, 2010 [24]	COSMIN study: COnsensus-based Standards for the selection of health Measurement INstruments	Guidelines established via systematic literature review and iterative Delphi process	Consensus was reached on design requirements and preferred statistical methods for the assessment of internal consistency, reliability, measurement error, content validity, construct validity, criterion validity, responsiveness, and interpretability
Crosby et al. [49]	Defining clinically meaningful change in health-related quality of life	Literature review	Reviews current approaches to defining clinically meaningful change in health-related quality of life and provides guidelines for their use
Dewolf et al. [36]	Translation procedure	Expert opinion	Provides guidance on the methodology for translating EORTC Quality of Life Questionnaires (QLQ)
Erickson et al. [19]	A concept taxonomy and an instrument hierarchy: tools for establishing and evaluating the conceptual framework of a patient-reported outcome (PRO) instrument as applied to product labeling claims	Expert opinion	Proposes a PRO concept taxonomy and instrument hierarchy that may be useful for demonstration of PRO measure claim for drug development, although they have not been tested for such purpose
Frost et al. [50]	What is sufficient evidence for the reliability and validity of patient-reported outcome measures?	Literature review	Article provides specific guidance on necessary psychometric properties of a PRO measure, with special reference to the FDA guidance, using the literature as a guide for specific statistical thresholds
Hays et al. [51]	The concept of clinically meaningful change in health-related quality of life research: How meaningful is it?	Expert opinion	Argues against a single threshold to define the minimally clinical important difference
Johnson et al. [26]	Guidelines for developing questionnaire modules	Expert opinion	Provides detailed description of PRO measure module development per the EORTC methodology related to generation of issues, construction of item list, pre- and field-testing
Kemmler et al. [52]	A new approach to combining clinical relevance and statistical significance for evaluation of quality of life changes in the individual patient	Longitudinal data from a chemotherapy trial	Data from this trial were used to evaluate change for individual participants (vs. groups). Stressed the importance of evaluation on the basis of statistical and clinical significance
Kottner et al. [53]	Guidelines for reporting reliability and agreement studies (GRRAS) were proposed	Literature review and expert consensus	Proposes a set of guidelines for reporting inter-rater agreement, inter-rater reliability in healthcare and medicine
Magasi et al. [33]	Content validity of patient-reported outcome measures: Perspectives from a PROMIS meeting	Expert presentation and discussion	The paper describes findings from a PROMIS meeting focused on content validity. Several recommendations were outlined as a result, including the need for consensus driven guidelines (none were proposed)

Table 1 continued

Author, year	Guideline	Research design	Description
Norquist et al. [42]	Choice of recall period for patient-reported outcome measures: criteria for consideration	Literature review	Choice of recall period for a PRO measure depends on nature of the disease, stability of symptoms, and trajectory of symptoms over time
Revicki et al. [12]	Recommendations on health-related quality of life research to support labeling and promotional claims in the United States	Review	Outlines the importance of an evidentiary base for making claims with respect to medical labeling or promotional claims
Revicki et al. [7]	Documenting the rationale and psychometric characteristics of patient-reported outcomes for labeling and promotional claims: the PRO Evidence Dossier	Report	Describes the purpose and content of a PRO measure Evidence Dossier, as well as its potential role with respect to regulatory review
Revicki et al. [34]	Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes	Literature review and expert opinion	Makes concrete recommendations regarding estimation of minimally important differences (MID), which should be based on patient-based and clinical anchors and convergence across multiple approaches and methods
Rothman et al. [30]	Use of existing patient-reported outcome (PRO) instruments and their modification	Expert opinion	Discusses key issues regarding the assessment and documentation of content validity for an existing instrument; discusses potential threats to content validity and methods to ameliorate
Schmidt et al. [54]	Current issues in cross-cultural quality of life instrument development	Literature review	Provides an overview of cross-cultural adaptation of PRO measure and provides broad development guidelines, as well as a call for additional focus on international research
Schunemann et al. [8]	Interpreting the results of patient-reported outcome measures in clinical trials: The clinician's perspective	Report based on examples	The authors provided several examples to describe how to attach meaning to PROM score thresholds and/or score differences
Scientific Advisory Committee of Medical Outcomes Trust [2]	Assessing health status and quality of life instruments: attributes and review criteria	Expert opinion	Describes 8 key attributes of PRO measures, including conceptual and measurement model, reliability, validity, responsiveness, interpretability, respondent and administrative burden, alternate forms, and cultural and language adaptations
Sprangers et al. [55]	Assessing meaningful change in quality of life over time: a users' guide for clinicians	Literature review and expert opinion	Proposes a set of guidelines/questions to help guide clinicians as to how to use PRO data in the treatment decision process
Snyder et al. [5]	Implementing patient-reported outcomes assessment in clinical practice: a review of the options and considerations	Literature review	The ISOQOL group developed a series of options and considerations to help guide the use of PROs in clinical practice, along with strengths and weaknesses of alternate approaches
Turner et al. [56]	Patient-reported outcomes: Instrument development and selection issues	Literature review	Provides a broad summary of concepts and issues to consider in the development and selection of a PRO measure
United States Food and Drug Administration [1]	Guidance for Industry: Patient-reported outcome measures: use in medical product development to support drug labeling claims	Expert opinion	"This guidance describes how the Food and Drug Administration (FDA) reviews and evaluates existing, modified, or newly created *patient-reported outcome instruments* used to support *claims* in approved medical product labeling." It covers conceptual frameworks, content validity, reliability, validity, ability to detect change, modification of PRO, and use of PRO in special populations

Table 1 continued

Author, year	Guideline	Research design	Description
Wild et al. [29]	Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes measures	Literature review and expert opinion/ consensus	The ISPOR Task Force produced a critique of the strengths and weaknesses of various methods for translation and cultural adaptation of PROMS
Wild et al. [31]	Multinational trials-recommendations on the translations required, approaches to using the same language in different countries, and the approaches to support pooling the data	Expert opinion and literature review	Provides decision tools to decide on translation required for PRO measure; approach to use when same language is spoken in more than one country; and methods to gather evidence to support pooling of data across different language versions
Wyrwich et al. [38]	Methods for interpreting change over time in patient-reported outcome measures	Literature review	This article reviews the evolution of the methods and the terminology used to describe and aid in the communication of meaningful PRO change score thresholds

(86 % of these from the United States) and 33 % from Europe.

The participants reported being skilled in qualitative and quantitative methods and felt comfortable providing guidance for recommendations for PRO measurement standards. Approximately 81 % of the sample reported they had moderate to extensive training in quantitative methods and 53 % reported they had moderate to extensive training in qualitative methods. Overall, 89 % reported they felt competent or very competent providing guidance. As a sensitivity analysis, we examined the endorsement of recommendations excluding the 11 % who felt only somewhat or a little competent, but this resulted in no changes for our final recommendations. On average, the sample had 15 years of PRO measurement and research experience in the field.

Minimum standards for selecting a PRO measure for use in PCOR

Table 3 provides definitions of the properties of a PRO measure, and Table 4 provides an overview of the results from the ISOQOL survey on draft recommendations for minimal standards. Table 5 provides final recommendations based on these results and the feedback from ISOQOL members. A review of the findings from our literature review and survey is provided below.

Conceptual and measurement model

ISOQOL members were very supportive of the minimum standards described in Table 4 (#1) with 90 % of respondents endorsing the statement that a PRO measure should have documentation that defines the PRO construct and describes the intended application of the measure in the intended population. Also, 61 % of respondents agreed the documentation should describe how the measured concept(s) are operationalized in the measurement model.

Reliability of a PRO measure

A majority of ISOQOL respondents agreed that as a minimum standard a multi-item PRO measure should be assessed for internal consistency reliability, and a single-item PRO measure should be assessed by test–retest reliability (see Table 4, #2). However, they did not support as a minimum standard that a multi-item PRO measure should be required to have evidence of test–retest reliability. They noted practical concerns regarding test–retest reliability; primarily that some populations studied in PCOR are not stable and that their HRQOL can fluctuate. This phenomenon would reduce estimates of test–retest reliability, making the PRO measure look unreliable when it may be accurately detecting changes over time. In addition, memory effects will positively influence the test–retest reliability when the two survey points are scheduled close to each other.

Respondents endorsed the minimum level of reliability of 0.70 for group-level comparisons, which is commonly accepted in the field [2, 40, 41]. The standard error of measurement at this reliability level is approximately 0.55 of a standard deviation. However, there were concerns that establishing an absolute cut-off would be too prescriptive (e.g., a PRO measure with an estimated reliability coefficient of 0.69 would be deemed unreliable). Some respondents (36 %) supported the statement that "no minimum level of reliability should be stated; however, the reliability should be appropriately justified for the context of the proposed PRO measurement application."

Table 2 Participant-reported sample characteristics

Sample characteristic	% ($n = 120$)
Degrees[a]	
MD	18 %
PhD/Other Doctoral Degree (e.g., ScD)	64 %
RN/NP	5 %
Physical/Occupational Therapist	7 %
MA, MSc, MPH, or other Master's	43 %
Role[a]	
Academic Researcher	68 %
Clinician	21 %
Industry Representative	8 %
Industry Consultant/CRO Employee	23 %
Federal Government Employee	6 %
Patient Advocate	2 %
Other	8 %
Geographic location	
North America	48 %
United States	(86 %)
Europe	33 %
South America	5 %
Asia	10 %
Africa	1 %
Australia	3 %
Quantitative training in PRO measure design and evaluation	
Extensive training	37 %
Moderate amount of training	44 %
A little training	16 %
Not any training	3 %
Qualitative training in PRO measure design and evaluation	
Extensive training	18 %
Moderate amount of training	35 %
A little training	40 %
Not any training	7 %
Competency	
Very competent	50 %
Competent	39 %
Somewhat competent	8 %
A little competent	3 %
Average number of years in health-related quality (HRQOL) or patient-reported outcomes (PROs) field	
Mean years in HRQOL or PRO field	15 years; (range 1–40 years)

[a] More than one response was allowed for this characteristic

Validity of a PRO measure

The most common types of validity that were considered for minimum standards were content validity, construct validity, and responsiveness. Responsiveness is often regarded as an aspect of validity [4, 37]; however, it is often discussed separately given its importance to PRO measurement in longitudinal studies [4]. Criterion-related validity was not considered since there is generally no "gold standard" to which to compare a PRO measure. In the survey of ISOQOL members, only 7 and 10 % felt criterion-related validity was critical to have for a PRO measure in a cross-sectional or longitudinal study, respectively. It should be noted that the APA standards manual [32] suggests that validity is a unitary concept including all aspects of validity. However, the field of

Table 3 Definition of PRO measure properties

Conceptual and measurement model—The conceptual model provides a description and framework for the targeted construct(s) to be included in a PRO measure. The measurement model maps the individual items in the PRO measure to the construct
Reliability—The degree to which a PRO measure is free from measurement error [2, 4, 40, 41]
Internal consistency reliability—The degree of the interrelatedness among the items in a multi-item PRO measure [2, 4]
Test–retest reliability—A measure of the reproducibility of the scale, that is, the ability to provide consistent scores over time in a stable population [2]
Validity—The degree to which a PRO instrument measures the PRO concept it purports to measure [2, 4, 41]
Content validity—The extent to which the PRO measure includes the most relevant and important aspects of a concept in the context of a given measurement application [50]
Construct validity—The degree to which scores on the PRO measure relate to other measures (e.g., patient-reported or clinical indicators) in a manner that is consistent with theoretically derived a priori hypotheses concerning the concepts that are being measured [40]
Criterion validity—The degree to which the scores of a PRO measure are an adequate reflection of a "gold standard." [4]
Responsiveness—The extent to which a PRO measure can detect changes in the construct being measured over time [2, 37]
Interpretability of scores—The degree to which one can assign easily understood meaning to a PRO measure's scores [2, 4]
Minimal important difference (*MID*)—The smallest difference in score in the outcome of interest that informed patients or informed proxies perceive as important, either beneficial or harmful, and that would lead the patient or clinician to consider a change in the management [44, 57, 58]
Burden—The time, effort, and other demands placed on those to whom the instrument is administered (respondent burden) or on those who administer the instrument (investigator or administrative burden) [2]

outcomes research still distinguishes the above terms, probably because different methodologies are needed to address different forms of validity.

Content validity was rated as one of the most critical forms of validity to be assessed for a PRO measure with 58 and 61 % of ISOQOL members indicating a PRO measure must have evidence for content validity before using it in a cross-sectional or longitudinal study, respectively (data not shown in Table 4) [1]. Although the recommendations for minimum standards for content validity were endorsed by ISOQOL members (see Table 4, #3a), there was disagreement about the recall period, which is the period of time of reference (e.g., currently, past 24 h, past 7 days, past 4 weeks) for patients to describe their experiences with the measured PRO. Most (52 %) believed that a justification for the recall period was desirable but not required as a minimum standard for a PRO measure. In the final recommendation, we recommend that the reference period must be considered carefully in order for research participants to provide valid responses. However, we do not recommend a single recall period as it varies depending on the PRO domain being measured, the research context, and the population being studied [42].

Another aspect of content validity has to do with the provenance of items. One statement that was considered as a minimum standard but not supported by ISOQOL members was for the "documentation of sources from which items were derived, modified, and prioritized during the PRO measure development process." Because a majority of respondents felt this standard was important (46 % voted "required as minimum standard" and 46 % voted "desirable but not required"), we recommend this documentation be considered as a "best practice" but not a minimum standard for PRO measures.

Construct validity was also judged a critical component of validity. A majority of respondents (55 %) judged documentation of empirical findings supporting a priori hypotheses regarding expected associations among similar and dissimilar measures to be a minimal standard for a PRO measure (see Table 4, #3b). Another part of our original recommendation considered documented evidence for "known groups" validity, requiring empirical findings that support predefined hypotheses of the expected differences in scores between "known" groups. We considered this to be an important part of the evaluation of construct validity as it demonstrates the ability of a PRO measure to distinguish between one group and another where there is past empirical evidence of differences between the groups. However, the majority of ISOQOL members (57 %) rated it as a desirable but not required standard. Therefore, we considered this as a standard for "best practice" rather than a minimum standard.

Responsiveness, also referred to as longitudinal validity, is an aspect of construct validity [23, 37, 43]. A majority of ISOQOL respondents supported minimum standards of obtaining empirical evidence of changes in scores consistent with predefined hypotheses prior to using the PRO measure in longitudinal research (see Table 4, #3c). However, 65 % of respondents reported that they would use a PRO measure in a longitudinal study even if there was no prior study to support the responsiveness of the scale, but did have scientific evidence in a cross-sectional study of the reliability, content validity, and construct validity of the PRO measure.

Table 4 ISOQOL survey results on draft recommendations

	Draft recommendation for minimal standards	Survey results ($n = 120$)
1	*Conceptual and measurement model*	
	A PRO measure should have documentation defining and describing the concept(s) included and the intended population(s) for use	Required as a minimum standard—90 % Desirable but not required as a minimum standard—9 % Not required—0 % Not sure—1 % No opinion—0 %
	In addition, there should be documentation of how the concept(s) are organized into a measurement model, including evidence for the dimensionality of the measure, how items relate to each measured concept, and the relationship among concepts included in the PRO measure	Required as a minimum standard—61 % Desirable but not required—35 % Not required—3 % Not sure—1 % No opinion—0 %
2	*Reliability*	
	The reliability of a PRO measure should ideally be at or above 0.70 for group-level comparisons	Yes, it should be at or above 0.70—54 % No, it should be at or above _fill in blank_—8 % (responses ranged from 0.50 to 0.80) No minimum level of reliability should be appropriately justified for the context of the proposed application—36 % No opinion—2 %
	Reliability for a multi-item unidimensional scale should include an assessment of internal consistency	Required as a minimum standard—79 % Desirable but not required—14 % Not required—2 % Not sure—3 % No opinion—2 %
	Reliability for a multi-item unidimensional scale should include an assessment of test–retest reliability	Required as a minimum standard—43 % Desirable but not required—51 % Not required—3 % Not sure—3 % No opinion—0 %
	Reliability for a single-item measure should be assessed by test–retest reliability	Required as a minimum standard—60 % Desirable but not required—34 % Not required—2 % Not sure—3 % No opinion—1 %
3	*Validity*	
3a	*Content validity*	
	A PRO measure should have evidence supporting its content validity, including evidence that patients and/or experts consider the content of the PRO measure relevant and comprehensive for the concept, population, and aim of the measurement application	Required as a minimum standard—78 % Desirable but not required—19 % Not required—2 % Not sure—0 % No opinion—1 %
	Documentation of qualitative and/or quantitative methods used to solicit and confirm attributes (i.e., concepts measured by the items) of the PRO relevant to the measurement application	Required as a minimum standard—53 % Desirable but not required—44 % Not required—2 % Not sure—1 % No opinion—0 %

Table 4 continued

	Draft recommendation for minimal standards	Survey results ($n = 120$)
	Documentation of the characteristics of participants included in the evaluation (e.g., race/ethnicity, culture, age, socio-economic status, literacy)	Required as a minimum standard—52 % Desirable but not required—47 % Not required—0 % Not sure—0 % No opinion—1 %
	Documentation of sources from which items were derived, modified, and prioritized during the PRO measure development process	Required as a minimum standard—46 % Desirable but not required—46 % Not required—7 % Not sure—0 % No opinion—1 %
	Justification for the recall period for the measurement application	Required as a minimum standard—41 % Desirable but not required—52 % Not required—5 % Not sure—1 % No opinion—1 %
3b	*Construct validity*	
	A PRO measure should have evidence supporting its construct validity, including documentation of empirical findings that support predefined hypotheses on the expected associations among measures similar or dissimilar to the measured PRO	Required as a minimum standard—55 % Desirable but not required—44 % Not required—1 % Not sure—0 % No opinion—0 %
	A PRO measure should have evidence supporting its construct validity, including documentation of empirical findings that support predefined hypotheses of the expected differences in scores between "known" groups	Required as a minimum standard—41 % Desirable but not required—57 % Not required—2 % Not sure—0 % No opinion—0 %
3c	*Responsiveness*	
	A PRO measure for use in longitudinal research study should have evidence of responsiveness, including empirical evidence of changes in scores consistent with predefined hypotheses regarding changes in the target population for the research application	Required as a minimum standard—57 % Desirable but not required—42 % Not required—1 % Not sure—0 % No opinion—0 %
	If a PRO measure has cross-sectional data that provide sufficient evidence in regard to the reliability (internal consistency), content validity, and construct validity but has no data yet on responsiveness over time (i.e., ability of a PRO measure to detect changes in the construct being measured over time), would you accept use of the PRO measure to provide valid data over time in a longitudinal study if no other PRO measure was available?	Yes—65 % No, I would require evidence of responsiveness before accepting it—32 % No opinion—0 % Comments (fill in blank response)—22 %
4	*Interpretability of Scores*	
	A PRO measure should have documentation to support interpretation of scores, including what low and high scores represent for the measured concept	Required as a minimum standard—64 % Desirable but not required—35 % Not required—1 % Not sure—0 % No opinion—0 %
	A PRO measure should have documentation to support interpretation of scores, including representative mean(s) and standard deviation(s) in the reference population	Required as a minimum standard—39 % Desirable but not required—57 % Not required—4 % Not sure—0 % No opinion—0 %

Table 4 continued

	Draft recommendation for minimal standards	Survey results ($n = 120$)
	A PRO measure should have documentation to support interpretation of scores, including guidance on the minimally important difference in scores between groups and/or over time that can be considered meaningful from the patient and/or clinical perspective	Required as a minimum standard—23 % Desirable but not required—72 % Not required—5 % Not sure—0 % No opinion—0 %
5	*Translation of a PRO measure*	
	A PRO measure translated to one or more languages should have evidence of the equivalence of measurement properties for translated versions, allowing comparison or combination of data across language forms	Required as a minimum standard—47 % Desirable but not required—49 % Not required—4 % Not sure—0 % No opinion—0 %
	Documentation of background and experience of the persons involved in the translation	Required as a minimum standard—43 % Desirable but not required—49 % Not required—8 % Not sure—0 % No opinion—0 %
	Documentation of methods used to translate and evaluate the PRO measure in each language	Required as a minimum standard—81 % Desirable but not required—16 % Not required—3 % Not sure—0 % No opinion—0 %
	Documentation of extent of harmonization across different language versions	Required as a minimum standard—38 % Desirable but not required—53 % Not required—7 % Not sure—2 % No opinion—0 %
6	*Patient and investigator Burden*	
	The reading level of the PRO measure for research involving adult respondents from the general population should be at a minimum of...	4th grade education level—7 % 6th grade education level—23 % 8th grade education level—6 % Other grade level ____—8 % There should be no minimum requirement of the literacy level of the PRO measure; however, it should be appropriately justified for the context of it proposed application—43 % Not sure—9 % No opinion—4 %

Interpretability of scores

For a PRO measure to be well accepted for the use in PCOR, it must provide scores that are easily interpreted by different stakeholders including patients, clinicians, researchers, and policy makers [38]. The literature review revealed several ways to enhance interpretability of scores that may be considered for standard setting. End-users must be able to know what a high or low score represents. In addition, knowing what comprises a meaningful difference or change in the score from one group to another (or one time to another) would enhance understanding of the outcome being measured. Another way to enhance the interpretability of PRO measure scores would involve comparing scores from a study to known scores in a population (e.g., the general US population or a specific disease population). The availability of such benchmarks would enhance understanding of how the study group scored as compared to some reference or normative group.

A majority of respondents endorsed as a minimum standard that a PRO measure should have documentation to support the interpretation of scores including description of what low and high scores represent (see Table 4, #4). However, more useful metrics such as norm or reference

Table 5 Final recommendations for minimum standards for patient-reported outcome (PRO) measures used in patient-centered outcomes research or comparative effectiveness research

1	*Conceptual and measurement model*—A PRO measure should have documentation defining and describing the concept(s) included and the intended population(s) for use. In addition, there should be documentation of how the concept(s) are organized into a measurement model, including evidence for the dimensionality of the measure, how items relate to each measured concept, and the relationship among concepts included in the PRO measure
2	*Reliability*—The reliability of a PRO measure should preferably be at or above 0.70 for group-level comparisons, but may be lower if appropriately justified. Reliability can be estimated using a variety of methods including internal consistency reliability, test–retest reliability, or item response theory. Each method should be justified
3	*Validity*
3a	*Content validity*—A PRO measure should have evidence supporting its content validity, including evidence that patients and experts consider the content of the PRO measure relevant and comprehensive for the concept, population, and aim of the measurement application. This includes documentation of as follows: (1) qualitative and/or quantitative methods used to solicit and confirm attributes (i.e., concepts measured by the items) of the PRO relevant to the measurement application; (2) the characteristics of participants included in the evaluation (e.g., race/ethnicity, culture, age, gender, socio-economic status, literacy level) with an emphasis on similarities or differences with respect to the target population; and (3) justification for the recall period for the measurement application
3b	*Construct validity*—A PRO measure should have evidence supporting its construct validity, including documentation of empirical findings that support predefined hypotheses on the expected associations among measures similar or dissimilar to the measured PRO
3c	*Responsiveness*—A PRO measure for use in longitudinal research study should have evidence of responsiveness, including empirical evidence of changes in scores consistent with predefined hypotheses regarding changes in the measured PRO in the target population for the research application
4	*Interpretability of scores*—A PRO measure should have documentation to support interpretation of scores, including what low and high scores represent for the measured concept
5	*Translation of the PRO measure*—A PRO measure translated to one or more languages should have documentation of the methods used to translate and evaluate the PRO measure in each language. Studies should at least include evidence from qualitative methods (e.g., cognitive testing) to evaluate the translations
6	*Patient and investigator Burden*—A PRO measure must not be overly burdensome for patients or investigators. The length of the PRO measure should be considered in the context of other PRO measures included in the assessment, the frequency of PRO data collection, and the characteristics of the study population. The literacy demand of the items in the PRO measure should usually be at a 6th grade education level or lower (i.e., 12 year old or lower); however, it should be appropriately justified for the context of the proposed application

scores or minimally important difference (MID) estimates were not considered required, but were considered highly desirable [34, 44, 45].

Translation of a PRO measure

PCOR and CER are often carried out in multi-national or multi-cultural settings that require the PRO measure to be translated into different languages. To be able to compare or combine HRQOL results across those groups, it is critical that the measured HRQOL concept and the wording of the questionnaire used to measure it is interpreted in the same way across translations [29, 46].

Of the original draft recommendations reviewed in the survey (see Table 4, #5), ISOQOL members supported as a minimum standard the statement, "Documentation of methods used to translate and evaluate the PRO measure in each language." In response to follow-up questions (not summarized in Table 4), 41 % of respondents considered it necessary, while 40 % felt it was expected but not required, to employ qualitative methods (e.g., cognitive interviews) for reviewing the quality of translations before using a translated PRO measure. Only 24 % of respondents thought that quantitative methods should be required for reviewing the quality of the translations (e.g., differential item functioning testing) before using the PRO measure, and 42 % of respondents indicated that it was expected (but not absolutely necessary) to include quantitative evaluation before they would use a translated PRO measure. Based on these findings, the ISOQOL SATF recommended that qualitative evidence be included as a minimum standard for translated PRO measures (Table 5).

Patient and investigator Burden

The committee agreed that burden on patients and investigators must be considered when selecting PRO measures for a PCOR study. A PRO measure must not be overly burdensome for patients as they are often ill and should not be subjected to overly long questionnaires or too frequent data collection that disrupts their lives. Ninety-two percent of the survey respondents concurred, endorsing "respondent burden" as an important or very important consideration for selecting PRO measures for PCOR.

Similarly, 90 % of respondents endorsed literacy as an important or very important consideration in selecting PRO measures in PCOR. Data collected from PRO measures are

only valid if the participants in a study can understand what is being asked of them and can provide a response that accurately reflects their experiences or perspectives. It is critical that developers of PRO measures ensure the questions, and response options are clear and easy to understand. Qualitative testing of the PRO measure (e.g., cognitive interviewing) should include individuals with low literacy to evaluate the questions [47]. Twenty-three percent of respondents indicated that a PRO measure should be written at 6th grade education level (ages 11–12 years), while 43 % indicated that the literacy level should be appropriately justified for the given research application.

Discussion

Based on a literature review of existing guidelines and a survey of experts in PRO measurement and research, we, on behalf of the ISOQOL, put forth minimum standards for PRO measures to be used in patient-centered outcomes research and comparative effectiveness research. These recommendations include the documentation of the characteristics of the conceptual and measurement model, evidence for reliability, validity, and interpretability of scores, quality translations, and acceptable patient and investigator burden (summarized in Table 5). The extent to which a PRO measure adheres to the standards described in this report reflects the quality of the PRO measurement.

Good documentation of the evidence that a PRO measure meets and exceeds these measurement properties will result in greater acceptance of the PRO measure for use in PCOR and CER. This documentation could include a focused methodologically rigorous study of the measurement properties of the PRO measure or analysis of HRQOL data collected from the PRO measure within a PCOR or CER study. Such documentation should be made available in peer-reviewed literature as well as on publically accessible websites. To the extent that the evidence was obtained from populations similar to the target population in the study, the investigator(s) will have greater confidence in the PRO measure to capture patients' experiences and perspectives.

There are a number of considerations when applying these minimum standards in PCOR and CER. The populations participating in PCOR and CER will likely be more heterogeneous than those that are typically included in phase II or III clinical trials. This population heterogeneity should be reflected in the samples included in the evaluation of the measurement properties for the PRO measure. For example, both qualitative and quantitative studies may require quota sampling based on race/ethnicity, gender, or age groups that reflect the prevalence of the condition in the study target population.

Researchers must consider carefully the strength of evidence supporting the measurement properties of the PRO measure. There is no threshold for which an instrument is valid or not valid for all populations or applications. In addition, no single study can confirm all the measurement properties for all research contexts. Like all scientific disciplines, measurement science relies on the iterative accumulation of a body of evidence (maturation model), replicated in different settings. Thus, it is the weight of the evidence (i.e., the number and quality of the studies and consistency of findings) that informs the evaluation of the appropriateness of a PRO measure. Older PRO measures will sometimes have the benefit of having more evidence than newer measures, and this will be reflected in the standards.

A possible limitation of this study is the potential for the biases of individual members of the SATF to influence the survey content. The transparency of the process used, and the wide variety of expertise and perspectives among the members, mitigated against substantive bias being introduced. In addition, the response rate to the survey was modest, again indicating the potential for bias. We point out, however, that the demographic data collected on the survey indicated that the respondents were experienced ISOQOL members with a variety of professional perspectives, the vast majority of whom self-identified as being competent in providing ratings and responses for the survey items.

These minimum standards were created by ISOQOL to reflect when a PRO measure may be considered appropriate or inappropriate for a specific PCOR study; thus, the intent was to have a minimum standard by which PRO measures could be judged acceptable. These standards do not reflect "ideal standards" or "best practices," which will have more stringent criteria [2, 3, 40]. For example, established minimally important differences for a PRO measure will enhance the interpretability of scores to inform decision making. As another example, establishing measurement equivalence of the PRO across different modes of assessment (e.g., paper forms, computers, handheld devices, phone) may facilitate broader patient participation in PCOR. ISOQOL's recommendations for "best practices" for PRO measures in PCOR and CER will be a next step in the organization's strategic initiative to advance the science of HRQOL measurement.

The findings from this study were reviewed by the PCORI Methodology Committee as part of that Committee's review of relevant standards and guidelines pertinent to patient-centered outcomes research. The ISOQOL recommendations presented here focus on more specific information about PRO measurement properties than those found in the PCORI Methodology Committee standards [15].

The identification and selection of PRO measures meeting and exceeding these current ISOQOL recommended minimum standards will increase the likelihood that the evidence generated in PCOR and CER reliably and validly represents the patients' perspective on health-related outcomes. This PRO evidence, based on instruments with sound measurement properties, can then be used to inform clinical and health policy decision making about the benefits and risks associated with different health interventions or to monitor population health.

Acknowledgments This study was funded by the Patient-Centered Outcomes Research Institute (PCORI-SOL-RMWG-001: PIs: Zeeshan Butt, PhD, Northwestern University; Bryce Reeve, PhD, University of North Carolina at Chapel Hill). The views expressed in this article are those of the authors and do not necessarily reflect those of PCORI.

References

1. US Food and Drug Administration. (2009). Patient-reported outcome measures: Use in medical product development to support labeling claims. Guidance for industry. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM071975.pdf. Accessed November 26, 2011.
2. Scientific Advisory Committee of the Medical Outcomes Trust. (2002). Assessing health status and quality-of-life instruments: Attributes and review criteria. *Quality of Life Research, 11*(3), 193–205.
3. Mokkink, L. B., Terwee, C. B., Knol, D. L., Stratford, P. W., Alonso, J., Patrick, D. L., et al. (2006). Protocol of the COSMIN study: COnsensus-based standards for the selection of health measurement INstruments. *BMC Medical Research Methodology, 6*, 2. doi:10.1186/1471-2288-6-2.
4. Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., et al. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology, 63*(7), 737–745. doi:10.1016/j.jclinepi.2010.02.006.
5. Snyder, C. F., Aaronson, N. K., Choucair, A. K., Elliott, T. E., Greenhalgh, J., Halyard, M. Y., et al. (2011). Implementing patient-reported outcomes assessment in clinical practice: A review of the options and considerations. *Quality of Life Research,*. doi:10.1007/s11136-011-0054-x.
6. Basch, E. M., Reeve, B. B., Mitchell, S. A., Clauser, S. B., Minasian, L., Sit, L., et al. (2011). Electronic toxicity monitoring and patient-reported outcomes. *Cancer Journal, 17*(4), 231–234. doi:10.1097/PPO.0b013e31822c28b3.
7. Revicki, D. A., Gnanasakthy, A., & Weinfurt, K. (2007). Documenting the rationale and psychometric characteristics of patient reported outcomes for labeling and promotional claims: The PRO Evidence Dossier. *Quality of Life Research, 16*(4), 717–723. doi:10.1007/s11136-006-9153-5.
8. Schunemann, H. J., Akl, E. A., & Guyatt, G. H. (2006). Interpreting the results of patient reported outcome measures in clinical trials: The clinician's perspective. *Health and Quality of Life Outcomes, 4*, 62. doi:10.1186/1477-7525-4-62.
9. Deyo, R. A., & Patrick, D. L. (1989). Barriers to the use of health status measures in clinical investigation, patient care, and policy research. *Medical Care, 27*(3 Suppl), S254–S268.
10. International Society for Quality of Life Research. http://www.isoqol.org/. Accessed July 30, 2012.
11. Guyatt, G., & Schunemann, H. (2007). How can quality of life researchers make their work more useful to health workers and their patients? *Quality of Life Research, 16*(7), 1097–1105. doi:10.1007/s11136-007-9223-3.
12. Revicki, D. A., Osoba, D., Fairclough, D., Barofsky, I., Berzon, R., Leidy, N. K., et al. (2000). Recommendations on health-related quality of life research to support labeling and promotional claims in the United States. *Quality of Life Research, 9*(8), 887–900.
13. Lipscomb, J., Donaldson, M. S., Arora, N. K., Brown, M. L., Clauser, S. B., Potosky, A. L., et al. (2004). Cancer outcomes research. *Journal of the National Cancer Institute Monographs* (33), 178–197. doi:10.1093/jncimonographs/lgh039.
14. US Patient-Centered Outcomes Research Institute http://www.pcori.org. Accessed 26 November, 2011.
15. Methodology Committee of the Patient-Centered Outcomes Research, I. (2012). Methodological standards and patient-centeredness in comparative effectiveness research: The PCORI perspective. *Journal of the American Medical Association, 307*(15), 1636–1640. doi:10.1001/jama.2012.466.
16. Terwee, C. B., Jansma, E. P., Riphagen, I. I., & de Vet, H. C. (2009). Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. *Quality of Life Research, 18*(8), 1115–1123. doi:10.1007/s11136-009-9528-5.
17. Qualtrics Labs Inc. Why choose qualtrics survey software? https://www.qualtrics.com/why-survey-software. Accessed November 26, 2011.
18. US Food and Drug Administration. (2010). Qualification process for drug development tools. Draft Guidance for Industry. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM230597.pdf. Accessed November 26, 2011.
19. Erickson, P., Willke, R., & Burke, L. (2009). A concept taxonomy and an instrument hierarchy: Tools for establishing and evaluating the conceptual framework of a patient-reported outcome (PRO) instrument as applied to product labeling claims. *Value in Health, 12*(8), 1158–1167. doi:10.1111/j.1524-4733.2009.00609.x.
20. Patrick, D. L., Burke, L. B., Powers, J. H., Scott, J. A., Rock, E. P., Dawisha, S., et al. (2007). Patient-reported outcomes to support medical product labeling claims: FDA perspective. *Value in Health, 10*(Suppl 2), S125–S137. doi:10.1111/j.1524-4733.2007.00275.x.
21. Angst, F. (2011). The new COSMIN guidelines confront traditional concepts of responsiveness. *BMC Medical Research Methodology, 11*, 152; author reply 152. doi:10.1186/1471-2288-11-152.
22. Mokkink, L. B., Terwee, C. B., Gibbons, E., Stratford, P. W., Alonso, J., Patrick, D. L., et al. (2010). Inter-rater agreement and reliability of the COSMIN (COnsensus-based standards for the selection of health status measurement instruments) checklist. *BMC Medical Research Methodology, 10*, 82. doi:10.1186/1471-2288-10-82.
23. Mokkink, L. B., Terwee, C. B., Knol, D. L., Stratford, P. W., Alonso, J., Patrick, D. L., et al. (2010). The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. *BMC Medical Research Methodology, 10*, 22. doi:10.1186/1471-2288-10-22.
24. Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., et al. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. *Quality of Life Research, 19*(4), 539–549. doi:10.1007/s11136-010-9606-8.
25. Terwee, C. B., Mokkink, L. B., Knol, D. L., Ostelo, R. W., Bouter, L. M., & de Vet, H. C. (2012). Rating the methodological

quality in systematic reviews of studies on measurement properties: A scoring system for the COSMIN checklist. *Quality of Life Research, 21*(4), 651–657. doi:10.1007/s11136-011-9960-1.
26. Johnson, C., Aaronson, N., Blazeby, J. M., Bottomley, A., Fayers, P., Koller, M., et al. (2011). EORTC Quality of life group: Guidelines for developing questionnaire modules. http://groups.eortc.be/qol/sites/default/files/archives/guidelines_for_developing_questionnaire-_final.pdf. Accessed November 26, 2011.
27. Cella, D. (1997). *Manual of the functional assessment of chronic illness therapy (FACIT) measurement system*. Evanston, IL: Northwestern University.
28. Coons, S. J., Gwaltney, C. J., Hays, R. D., Lundy, J. J., Sloan, J. A., Revicki, D. A., et al. (2009). Recommendations on evidence needed to support measurement equivalence between electronic and paper-based patient-reported outcome (PRO) measures: ISPOR ePRO good research practices task force report. *Value in Health, 12*(4), 419–429. doi:10.1111/j.1524-4733.2008.00470.x.
29. Wild, D., Grove, A., Martin, M., Eremenco, S., McElroy, S., Verjee-Lorenz, A., et al. (2005). Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (PRO) measures: Report of the ISPOR task force for translation and cultural adaptation. *Value in Health, 8*(2), 94–104. doi:10.1111/j.1524-4733.2005.04054.x.
30. Rothman, M., Burke, L., Erickson, P., Leidy, N. K., Patrick, D. L., & Petrie, C. D. (2009). Use of existing patient-reported outcome (PRO) instruments and their modification: The ISPOR good research practices for evaluating and documenting content validity for the use of existing instruments and their modification PRO task force report. *Value in Health, 12*(8), 1075–1083. doi:10.1111/j.1524-4733.2009.00603.x.
31. Wild, D., Eremenco, S., Mear, I., Martin, M., Houchin, C., Gawlicki, M., et al. (2009). Multinational trials-recommendations on the translations required, approaches to using the same language in different countries, and the approaches to support pooling the data: the ISPOR patient-reported outcomes translation and linguistic validation good research practices task force report. *Value in Health, 12*(4), 430–440. doi:10.1111/j.1524-4733.2008.00471.x.
32. Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1998). Standards for educational and psychological testing Washington DC: American Psychological Association.
33. Magasi, S., Ryan, G., Revicki, D., Lenderking, W., Hays, R. D., Brod, M., et al. (2011). Content validity of patient-reported outcome measures: Perspectives from a PROMIS meeting. *Quality of Life Research,*. doi:10.1007/s11136-011-9990-8.
34. Revicki, D., Hays, R. D., Cella, D., & Sloan, J. (2008). Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of Clinical Epidemiology, 61*(2), 102–109. doi:10.1016/j.jclinepi.2007.03.012.
35. Valderas, J. M., Ferrer, M., Mendivil, J., Garin, O., Rajmil, L., Herdman, M., et al. (2008). Development of EMPRO: A tool for the standardized assessment of patient-reported outcome measures. *Value in Health, 11*(4), 700–708. doi:10.1111/j.1524-4733.2007.00309.x.
36. Dewolf, L., Koller, M., Velikova, G., Johnson, C., Scott, N., & Bottomley, A. (2009). EORTC quality of life group: Translation procedure. http://groups.eortc.be/qol/sites/default/files/archives/translation_manual_2009.pdf. Accessed November 26, 2011.
37. Hays, R. D., & Hadorn, D. (1992). Responsiveness to change: An aspect of validity, not a separate dimension. *Quality of Life Research, 1*(1), 73–75.
38. Wyrwich, K. W., Norquist, J. M., Lenderking, W. R., Acaster, S., & the Industry Advisory Committee of International Society for Quality of Life, R. (2012). Methods for interpreting change over time in patient-reported outcome measures. *Quality of Life Research, 2012 Apr 17*. [Epub ahead of print.]. doi:10.1007/s11136-012-0175-x.
39. Ahmed, S., Berzon, R. A., Revicki, D., Lenderking, W., Moinpour, C. M., Basch, E., et al. (2012). The use of patient-reported outcomes (PRO) within comparative effectivenss research: Implications for clinical practice and healthcare policy. *Medical Care, 50*(12), 1060–1070.
40. Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., et al. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology, 60*(1), 34–42. doi:10.1016/j.jclinepi.2006.03.012.
41. Nunnally, J. C., & Berstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
42. Norquist, J. M., Girman, C., Fehnel, S., Demuro-Mercon, C., & Santanello, N. (2011). Choice of recall period for patient-reported outcome (PRO) measures: Criteria for consideration. *Quality of Life Research,*. doi:10.1007/s11136-011-0003-8.
43. Revicki, D. A., Cella, D., Hays, R. D., Sloan, J. A., Lenderking, W. R., & Aaronson, N. K. (2006). Responsiveness and minimal important differences for patient reported outcomes. *Health and Quality of Life Outcomes, 4*, 70. doi:10.1186/1477-7525-4-70.
44. Brozek, J. L., Guyatt, G. H., & Schunemann, H. J. (2006). How a well-grounded minimal important difference can enhance transparency of labelling claims and improve interpretation of a patient reported outcome measure. *Health and Quality of Life Outcomes, 4*, 69. doi:10.1186/1477-7525-4-69.
45. Norman, G. R., Sridhar, F. G., Guyatt, G. H., & Walter, S. D. (2001). Relation of distribution- and anchor-based approaches in interpretation of changes in health-related quality of life. *Medical Care, 39*(10), 1039–1047.
46. Koller, M., Kantzer, V., Mear, I., Zarzar, K., Martin, M., Greimel, E., et al. (2012). The process of reconciliation: Evaluation of guidelines for translating quality-of-life questionnaires. *Expert Review of Pharmacoeconomics & Outcomes Research, 12*(2), 189–197. doi:10.1586/erp.11.102.
47. Jordan, J. E., Osborne, R. H., & Buchbinder, R. (2011). Critical appraisal of health literacy indices revealed variable underlying constructs, narrow content and psychometric weaknesses. *Journal of Clinical Epidemiology, 64*(4), 366–379. doi:10.1016/j.jclinepi.2010.04.005.
48. Acquadro, C., Conway, K., Hareendran, A., & Aaronson, N. (2008). Literature review of methods to translate health-related quality of life questionnaires for use in multinational clinical trials. *Value in Health, 11*(3), 509–521. doi:10.1111/j.1524-4733.2007.00292.x.
49. Crosby, R. D., Kolotkin, R. L., & Williams, G. R. (2003). Defining clinically meaningful change in health-related quality of life. *Journal of Clinical Epidemiology, 56*(5), 395–407.
50. Frost, M. H., Reeve, B. B., Liepa, A. M., Stauffer, J. W., Hays, R. D., & Mayo, F. D. A. P.-R. O. C. M. G. (2007). What is sufficient evidence for the reliability and validity of patient-reported outcome measures? *Value in Health, 10*(Suppl 2), S94–S105. doi:10.1111/j.1524-4733.2007.00272.x.
51. Hays, R. D., & Woolley, J. M. (2000). The concept of clinically meaningful difference in health-related quality-of-life research. How meaningful is it? *PharmacoEconomics, 18*(5), 419–423.
52. Kemmler, G., Zabernigg, A., Gattringer, K., Rumpold, G., Giesinger, J., Sperner-Unterweger, B., et al. (2010). A new approach to combining clinical relevance and statistical significance for evaluation of quality of life changes in the individual patient. *Journal of Clinical Epidemiology, 63*(2), 171–179. doi:10.1016/j.jclinepi.2009.03.016.
53. Kottner, J., Audige, L., Brorson, S., Donner, A., Gajewski, B. J., Hrobjartsson, A., et al. (2011). Guidelines for reporting reliability

and agreement studies (GRRAS) were proposed. *Journal of Clinical Epidemiology, 64*(1), 96–106. doi:10.1016/j.jclinepi.2010.03.002.

54. Schmidt, S., & Bullinger, M. (2003). Current issues in cross-cultural quality of life instrument development. *Archives of Physical Medicine and Rehabilitation, 84*(4 Suppl 2), S29–S34. doi:10.1053/apmr.2003.50244.
55. Sprangers, M. A., Moinpour, C. M., Moynihan, T. J., Patrick, D. L., Revicki, D. A., & Clinical Significance Consensus Meeting Group. (2002). Assessing meaningful change in quality of life over time: A users' guide for clinicians. *Mayo Clinic Proceedings, 77*(6), 561–571. doi:10.4065/77.6.561.
56. Turner, R. R., Quittner, A. L., Parasuraman, B. M., Kallich, J. D., Cleeland, C. S., & Mayo, F. D. A. P.-R. O. C. M. G. (2007). Patient-reported outcomes: Instrument development and selection issues. *Value in Health, 10*(Suppl 2), S86–S93. doi:10.1111/j.1524-4733.2007.00271.x.
57. Schunemann, H. J., & Guyatt, G. H. (2005). Commentary–goodbye M(C)ID! Hello MID, where do you come from? *Health Services Research, 40*(2), 593–597. doi:10.1111/j.1475-6773.2005.00374.x.
58. Schunemann, H. J., Puhan, M., Goldstein, R., Jaeschke, R., & Guyatt, G. H. (2005). Measurement properties and interpretability of the Chronic respiratory disease questionnaire (CRQ). *Copd, 2*(1), 81–89.