DEVELOPMENT OF A METHODOLOGICAL STANDARDS REPORT: TOPIC # 3

THE DESIGN AND SELECTION OF PATIENT-REPORTED OUTCOMES MEASURES (PROMS) FOR USE IN PATIENT CENTERED OUTCOMES RESEARCH

Patient Centered Outcomes Research Institute (PCORI)

22 March 2012

Oxford Outcomes

CONFIDENTIAL

Authors:
Sarah Acaster
Tricia Cimms
Andrew Lloyd

Oxford Outcomes
188 The Embarcadero
Suite 200
San Francisco, CA
94105

Tel: (415) 371-2114
Fax: (415) 856-0840

Sarah.Acaster@oxfordoutcomes.com

www.oxfordoutcomes.com

Oxford – Bethesda – San Francisco – Morristown - Vancouver - Toronto

Table of Contents

1 Introduction

A growing body of research over many years has explored the potential role of understanding patient preferences and values when making health care decisions. Alongside this body of work has been research on the development of health related quality of life assessment and more latterly patient reported outcome measures (PROMs). These survey tools are used to assess the personal impact of disease or treatment on many different areas of a patient's life including physical, psychological and social functioning and general wellbeing. The data from PROMs are used around the world to support decision making at many different levels. National regulators and other agencies use PROM data to make decisions regarding the benefit of treatments and ultimately patient access to those treatments. At a more local level physicians are now much more experienced in reviewing and considering subjective measures of outcomes in addition to the more traditional clinical measures.

The Patient Centered Outcomes Research Institute (PCORI) has developed a remit to significantly advance this field of research. Patient Centered Outcomes Research (PCOR) is an initiative to further support patient decision making through the application of outcomes that patients value. PCOR is designed to support decision making by helping people to understand what will happen to them if they choose a treatment; what the potential benefits and harms of such treatment are; and by focusing on outcomes that patients value. This represents a coming together of different methodologies and experience. One of the first initiatives that PCORI have kicked off is the development of a set of draft minimum standards for the development, selection and use of patient reported outcomes data in PCOR studies. This report describes work that was undertaken to support the development of these standards.

2 Methods

2.1 Collation of Relevant Guidance Documents

Guidance documents relating to PROM development, selection and use were gathered from internal resources, an internet search, and a brief targeted literature review. A flow diagram showing the process used in presented in Figure 1 (page 6).

Initially, an email message was sent to all PRO, health economics, and epidemiology team members within Oxford Outcomes requesting copies of guidance documents and literature based on their area(s) of expertise. The email requested that team members identify and send:

- Existing published and draft (unpublished) guidance documents for designing or selecting a PROM
- Articles that provide guidelines or recommendations for appropriate steps to design a PROM, e.g. qualitative and/or quantitative approaches, interpretation of change, special populations, etc.
- Articles that provide recommendations for selecting a PROM, particularly for use in CER, e.g. guidelines for validity, reliability, responsiveness, translations, multiple modes of administration, etc.

Concurrently, a search strategy was developed to identify PRO guidance in published literature that might have been missed in the internal search. The search was conducted using EMBASE and Medline databases to ensure no key documents were missed. The search strategy was limited to English language studies, humans only. The keywords used were: patient reported outcome(s); PRO; observational studies; guidance; guidelines; standards; real world design; real world study; comparative effectiveness; patient centered.

Finally, a targeted internet search was conducted to search for relevant guidance documents that may not be listed in publication databases. The following websites were included:

- Food and Drug Administration (FDA)
- European Medicines Agency (EMA)
- International Society for Quality Of Life Research (ISOQOL)

- International Society of Pharmacoeconomics and Outcomes research (ISPOR)
- Agency for Healthcare Research and Quality (AHRQ)
- Initiative on Methods, Measurement and Pain Assessment in Clinical Trials (IMMPACT)
- NHS National Institutes for Health and Clinical Excellence (NICE)
- World Health Organization (WHO)
- Center for Disease Control (CDC)
- Consensus-based Standards for the selection of Health Measurement Instruments (COSMIN)
- Center for Medical Technology Policy (CMTP)

2.2 Establishing Concepts for Minimum Standards

After all key guidance documents were reviewed, internal teleconferences were held with the Oxford Outcomes PRO research teams to identify concepts and themes as either potential 'Minimum Standards' or 'Issues for Consideration'. The draft list of concepts was reviewed by internal experts and shared with the PCORI patient centered working group (PCWG) to ensure that a reasonable range of concepts had been covered. After the internal PRO team review and discussion with PCORI, the draft list was sent to external experts for further review.

Consultation was requested from experts in Health Economics, Epidemiology, CER, Biostatistics and ePRO: Adrian Levy, Rachael Fleurence, Andrew Briggs, Mark Sculpher, John Brazier, Ray Fitzpatrick and the Oxford university PROMs group, Helen Doll, Wilhelm Muehlhausen, and Diane Wild. The list of concepts and other considerations was finalized by the PRO research team; no further concepts or issues for consideration were raised during the review process.

2.3 Draft Minimum Standards and Issues for Consideration

When the list of concepts was agreed upon, Oxford Outcomes PRO team members and external experts were identified to draft the 'Minimum Standards' and 'Issues for Consideration' based on their expertise. At that time, any additional targeted primary literature needed to draft the standards was identified and sourced by the lead author of each standard. The minimum standard template and examples provided by PCORI were used to draft the 'Minimum Standards'; details of the guidance documents and primary literature used for each standard are included as appropriate.

Internal and external experts (listed above) reviewed the draft 'Minimum Standards' and 'Issues for Consideration' and their feedback was incorporated by the core PRO research team. The minimum standards and issues for consideration were also shared with the PCWG throughout the development process in order to incorporate their feedback and ensure the content was in line with expectations.

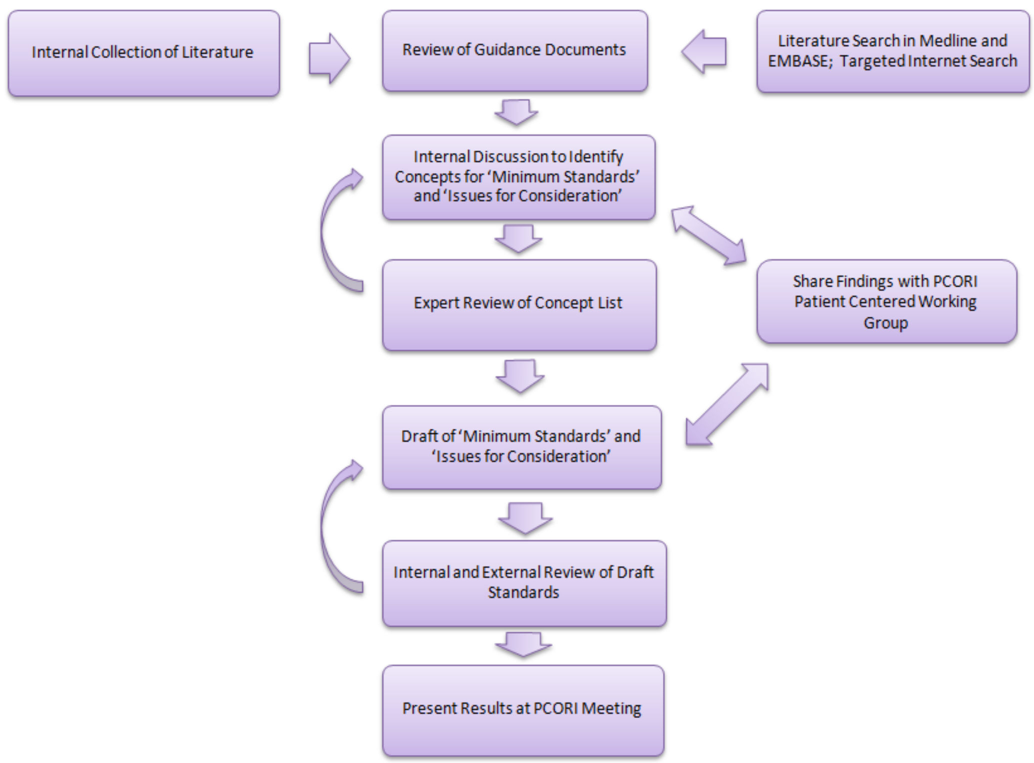

Figure 1 Flow Diagram of Methods

3 Results

3.1 Identification of Guidance Documents

The internal search resulted in a total of 136 relevant documents identified in response to the email. Seven key sources of published and draft guidance documents were identified as relevant:

- US Food and Drug Administration (FDA) Final PRO Label Guidance (FDA, 2009)
- European Medicines Agency (EMA) Reflection Paper (EMA, 2005)
- Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT) Papers (Turk et al., 2006; Dworkin et al., 2005; 2008)
- International Society for Quality of Life Research (ISOQOL): Guide to Implementing PRO Assessment in Clinical Practice (Aaronson et al., 2011)
- Center for Medical Technology Policy (CMTP) Draft Guidance: Recommendations for Incorporating PRO into the Design of Clinical Trials in Adult Oncology (Basch et al., 2011)
- Consensus-based Standards for the Selection of Health Measurement Instruments (COSMIN; (Mokkink et al., 2010)
- Getting the Most Out of PROMS (Devlin & Appleby, 2010; The King's Fund and the Office of Health Economics)

The remaining documents were specific to certain aspects of designing or selecting a PROM, e.g. psychometrics, interpretation of change, ePROs and others were good research practice guidelines, e.g. ISPOR task force papers. Many of these additional sources were used as primary literature in the development of the minimum standards.

Fifty eight papers were identified in the concurrent database search, however no further guidance documents were identified beyond those that had been identified with the internal search.

3.2 Minimum Standards

A total of twelve concepts were identified as appropriate for the development of a minimum standard for PCOR. Each of the twelve standards is summarized below with a brief description and reference to the relevance of the standard to patient centeredness and potential implementation issues. Detailed descriptions of each standard in accordance with the PCORI minimum standard template are provided in Appendix A.

3.2.1 Consideration of Patient Burden

Patient burden refers to the time, effort, and emotional strain associated with completing a PROM. Patient burden should be carefully considered when selecting, developing, or using a PROM, and every effort should be made to minimize the burden on the patient. Some of the factors affecting burden that need to be considered include: the PROM, study design and patient population.

- **Patient Centeredness**: Prioritizes the patient in study design. Potentially attenuates missing / erroneous data, and improves reliability and the standard of information communicated to patients.
- **Implementation Issues**: Difficult to monitor compliance.
- **Other Considerations**: Applying this standard may involve trading off on assessments of all concepts of interest. PROM development within clinical trial settings is narrow and single concept focused; multiple concepts are generally synonymous with multiple PROMs. PCOR is more likely to focus broadly on the overall impact of a disease and treatment, addressing several concepts, e.g. symptoms, functioning, side effects, and treatment satisfaction. Until more multi-dimensional PROMs are developed (based on qualitative input from patients regarding their key concerns), patients could be involved in the consideration of which PRO concepts to include in PCOR in order to maximize patient relevance and minimize burden.

3.2.2 Estimating and Reporting Reliability of a Patient Reported Outcome Measure (PROM)

Internal consistency and test retest reliability should be estimated and reported for each domain when appropriate (based on the nature of the PROM and the patient population). The widely reported 0.7 – 0.9 range for α coefficients and >0.70 for ICC coefficients should be considered as threshold guidelines. In order to evaluate the reliability of a PROM, detailed reporting of all results in studies using the PROM is required; this also assists meta-analyses.

- **Patient Centeredness**: Applies to all PROM research settings.
- **Implementation Issues**: No issues for assessment. Consistent standards of detailed reporting may be difficult to implement.
- **Other Considerations**: Internal consistency and test-re test reliability are not always appropriate, for example when assessing single item concepts or highly variable / transient symptoms, respectively. The appropriateness of the reliability coefficients reported should be considered in context of the PROM and the population. This standard should be considered in conjunction with the other psychometric standards.

3.2.3 Training Research Staff on the Administration of Patient Reported Outcome Measures (PROMs)

Training should be provided for site staff on the purpose of PROM data collection, administration procedures, completing checking processes, and the handling / storage of data. Staff should also be trained on patient communication (e.g. patients being given an opportunity to ask questions, and being informed participation is voluntary and responses are confidential). All staff training should be facilitated through use of a PROM user manual and the study protocol.

- **Patient Centeredness**: Trained not just on use and administration of the PROM but on communicating with patients.
- **Implementation Issues**: Minimal if incorporated into required staff training.
- **Other Considerations**: While this standard focuses on staff training in a study/trial environment, training on administration of PROMs for use in regular clinical practice should also be considered. Clear policies should be put in place regarding the confidential treatment and secure storage of PROM data at participating sites.

3.2.4 *Choosing an Appropriate Recall Period for Patient Reported Outcome Measures (PROMs)*

It is important to assess if the recall period is appropriate for the purpose/intended use, the patient population, the disease/condition, the treatment/device, and the study design. Patient understanding of the recall period should be assessed during cognitive debriefing.

- **Desired Standard**: Patient contributes to discussion of recall period during concept elicitation.
- **Patient Centeredness**: Patient contribution and assessment of patient understanding.
- **Implementation Issues**: Considering patient input to existing PROM may change recall and require additional validation work.
- **Other Considerations**: Specific consideration should be given to study design when selecting an appropriate recall period. The frequency of clinic/office visits suggested by study/trial designs could influence the choice for shorter or longer recall periods. In general, shorter recall periods (e.g., 24 hours, or 1 week at most) can be preferable to longer recall periods mainly because with longer recall data can be heavily biased by current health and any significant events.

3.2.5 *PROM Selection*

The selection of a PROM should be concept driven to include the concepts that are most meaningful and important to patients. PROMs should be selected on the basis of their content validity, measurement properties, interpretability, relation to the concept (s) of interest, and consideration of patient burden. Selection criteria and approach to addressing gaps in the evidence should be clearly documented.

- **Patient Centeredness**: Focus on concepts most important and meaningful to patients.
- **Implementation Issues**: Monitoring and reporting issues. For selection to be based on patient driven concepts patient involvement during study design may be needed until more multi-dimensional PROMs have been developed for PCOR.
- **Other Considerations**: All of Minimum Standards should be considered

when selecting a PROM.

3.2.6 *Interpretation of Meaningful Change on a Patient Reported Outcome Measure (PROM)*

Patient reported anchor based methods should be the primary source of data. Distribution based methods should be used to confirm that change identified by patient report is not likely to occur due to chance alone. Interpretation of meaningful change should be established for each domain of a PROM, with domain specific anchors.

- **Desired Standard**: Patient as an active participant.
- **Patient Centeredness**: Patient input to definition. Domain level interpretation provides patients with a clearer overall profile. Improved communication of information.
- **Implementation Issues**: No issues. *The desirable standard would require additional qualitative work and more research.*
- **Other Considerations**: The level of meaningful change is context dependent and may vary by patient group, severity level or other socio-demographic factors; as such any PRO instrument may be associated with numerous definitions of meaningful change. Thus, data should be continually gathered and consolidated over time; interpretability can then develop as the body of evidence accumulates.

3.2.7 *Establishing / Assessing Content Validity for Patient Reported Outcome Measures (PROMs)*

During the development phase of a PROM, concept elicitation interviews should be carried out to explore all concepts deemed relevant and important based on patient input. Cognitive debriefing interviews can then be completed to assess patient understanding of the measure. When evaluating a PROM for selection, concept elicitation and cognitive debriefing interviews should be carried out to ensure content relevant, understandable, and complete.

- **Desired Standard**: Include documentation of concept elicitation and cognitive debriefing.
- **Patient Centeredness**: Concepts included are patient driven and broad /

multidimensional.

- **Implementation Issues**: Broader populations mean greater time and costs. For documentation and monitoring a central repository for PROM would be required.
- **Other Considerations**: Content validity can also be supported (established/assessed) by speaking with experts. Content validity must be re-established if a PROM is modified (e.g. translated).

3.2.8 *Sampling in PROM Development / Selection / Validation*

Sampling methods differ for quantitative (i.e. validation) and qualitative (i.e. development and selection) PCOR studies, though both are essentially descriptive. Qualitative research is concerned with sampling diversity on sample characteristics, while quantitative research is focused on being representative.

- **Patient Centeredness**: Recruitment of diverse, fully representative samples will support interpretation of data in different groups.
- **Implementation Issues**: These studies can often be incorporated into planned or existing CER studies via secondary analysis.
- **Other Considerations**: Patients with comorbidities often have difficulty attributing specific symptoms/impacts to a specific condition. Including a subset of patients in concept elicitation who have no comorbidities could be considered.

3.2.9 *Estimating and Reporting Construct Validity of a Patient Reported Outcome Measure (PROM)*

Construct validity is a broad concept, encompassing criterion validity, known groups validity, and predictive validity. Construct validity should be estimated and clearly reported for all domains of any PROM developed or selected. No specific thresholds for establishing validity are detailed; the commonly reported 0.3 – 0.5 for criterion validity is recommended as a guideline.

- **Patient Centeredness**: Not unique to PCOR. However, accumulating validity data relating to predicting outcomes will be of more value to PCOR.
- **Implementation Issues**: No issues for assessment. Consistent standards of detailed reporting may be difficult to implement.

- **Other Considerations**: Known groups and predictive validity are likely to be highly desirable in PCOR and efforts to gather good scientific evidence of these properties should be undertaken.

3.2.10 *Estimating and Reporting Ability to Detect Change in a Patient Reported Outcome Measure (PROM)*

The ability of a PROM to detect change in terms of stability, improvement, and deterioration, should be assessed and clearly reported for all PROM domains. Reporting ability to detect change should include a clear statement about how change is assessed or determined, the target population, statistical test used and effect sizes.

- **Desired Standard**: Ability to detect change considered meaningful to patients.
- **Patient Centeredness**: Not unique to PCOR unless desirable standard applied
- **Implementation Issues**: Ability to detect change can be difficult in non-interventional settings. In longitudinal settings response shift needs to be considered.
- **Other Considerations**: Implementing this standard in predominantly non interventional studies where change within a relatively short period cannot be anticipated.

3.2.11 *Modification of an Existing Patient Reported Outcome Measure (PROM)*

The modification of the content of a PROM requires cognitive debrief interviews. All modifications, excluding those to instructions that do not impact the recall period or concept being measured, also require documentation of the new psychometric properties. The addition or removal of concepts from a PROM also requires qualitative evidence from concept elicitation interviews.

- **Patient Centeredness**: Patient involvement in all modifications to ensure the PROM remains relevant, clear and that no important concepts from the patient perspective are deleted or overlooked.
- **Implementation Issues**: Modifications are infrequently reported.

Monitoring and documenting compliance with this standard may require a central repository for PROMs used in PCOR.

- **Other Considerations**: This standard should be considered in conjunction with all other Minimum Standards.

3.2.12 Establishing Multi-Mode Equivalence for Patient Reported Outcome Measures (PROMs)

The transfer of any PROM from its original format to another mode of administration requires patient involvement in cognitive debrief and usability interviews. The need for quantitative assessment of equivalence should be determined based on the recommendations of the ISPOR ePRO Task Force (Coons et al., 2009). A moderate change requires equivalence testing and a substantial change also requires psychometric assessment.

- **Patient Centeredness**: Choice of mode of administration could make participation more convenient. May also make patients feel more of an active participant. *The views of study participants regarding modes of data collection could be usefully recorded and considered.*
- **Implementation Issues**: Will require researchers to be able to review the robustness of any ePROM and the extent to which additional equivalence testing may be required.
- **Other Considerations**: The choice of the mode should be formally assessed against other options and determined as preferable on the basis of suitability in this specific patient group, measurement equivalence, and cost. Studies that use patient diaries where patient reported data are collected repeatedly over many days only electronic data collection should be permitted.

3.3 Issues for Consideration

During the process of developing the draft minimum standards summarized above, several issues relevant to the development, selection and use of PROMs data in PCOR that would not necessarily fit within the framework of a minimum standard were identified. These issues relate to such matters as how PROMs data have been incorporated in health care decision making in other countries, specific sub populations included in PCOR, the need for ongoing evidence generation, and the role of patients in PCOR that extends beyond their contribution to the content of a PROM. These 'other considerations' are summarized below and more detailed versions are provided in Appendix B.

3.3.1 Lessons from the Health Technology Assessment (HTA) process

The HTA process aims to understand the real world value of a medical technology. Some countries (Australia, UK) establish the value of a medicine in terms of a metric which combines quality of life and length of life (quality adjusted life year or QALY). This could be useful for PCORI as it places patients' HRWL and values at the center of decision making. The use of the QALY has forced the HTA community to use single index measures, and to consider the value of health rather than just measuring it. This has potential application for PCORI as a way of addressing comparative effectiveness questions.

Understanding value: Data regarding HRQL reflect the value that people place on a health state. A preference based measure reflects the strength of preference which in turn reflects value with respect to how much life one is willing to give up to achieve a health state.

A single metric: To estimate QALYs HRQL is expressed as a single value (ranging between full health and dead).

Standardizing outcomes measurement: Standardizing HRQL measurement through the use of one measure allows easier comparison and meta-analysis.

Patient preferences: Formal methods for understanding patient preferences for interventions are quite often used in the HTA process as a supplement to cost effectiveness analyses. These methods may be useful for PCORI's aim of understanding what patients value.

Meta-analysis: Meta-analysis is widely used in HTA, but much less work on PROs has been reported. This would be a useful issue to support the work of PCORI.

3.3.2 Interpreting profile measures

Most PROs are profile measures with summary scores for different dimensions. This raises problems for interpretation:

- If patients improve a little on two dimensions but get worse on another it is difficult to understand the net change in health.

- Different dimensions of a questionnaire may not all be equally important for patients.
- The multiple domains of profile measures can present problems with hypothesis testing because of the need to test α multiple times. Hierarchical testing or other ways of handling multiplicity could solve this.

Profile measures have many benefits over single index measures. They allow us to understand the impact of an intervention on different aspects of functioning, symptoms and wellbeing. The use of profile measures is dependent on the research question. An example question relevant for comparative effectiveness may be "Should treatment A or treatment B used to treat X".

3.3.3 Confirmation of measurement properties

Determining that the measurement properties are appropriate for PCOR type research and fit to answer the research questions and for the samples included in studies is crucial.

1. Establishing validity involves re-confirming the measurement properties in new settings with new patients. There may be a greater emphasis on known groups and predictive validity in PCOR studies. Forecasting changes in health based upon PROMs could be an important aim of PCOR.
2. The studies that are undertaken to support PCOR and comparative effectiveness will provide useful datasets to further explore the psychometric performance of instruments.
3. Validation studies may have certain limitations, such as samples being restricted by race or other background variable.
4. Where a PRO may lack content validity in a sub-group of patients, a small number of cognitive debriefing interviews could be conducted. Studies PCORI are involved with could be used to confirm PRO properties through assessing differential item functioning, followed by traditional psychometric analyses of these data.

3.3.4 Response shift

Participants of longitudinal studies may experience an adjustment in their internal standards (i.e. recalibration), values (i.e. reprioritization) or meaning (i.e. reconceptualization) concerning the target construct (Schwartz & Sprangers, 1999). For example, a study about participation of disabled individuals found that individuals may develop a disability associated with a severity of fatigue that was previously unknown and consequently recalibrate what severe fatigue means to them, complicating predisability and postdisability assessments, or they may reconceptualize participation to focus on domains where they continue to have control (Schwartz, 2010). Such response shifts are to be expected with subjective rather than strictly objective assessments.

Response shift presents a challenge to basic assumptions concerning the use of standardized questionnaires and psychometric properties. Statistical techniques for the detection of response shift include the use of structural equation modeling or latent trajectory analysis. Research has also examined the use of individualized measures to examine stability in domains (O'Boyle et al, 1992; Ring et al, 2005).

Further research is required to examine issues such as clinical meaningfulness of response shift or predictors of response shift magnitude and direction. Currently there is no consensus on the terminology and theoretical models, which is of particular concern when seeking to address this issue in longitudinal studies. Much focus is currently on statistical methods that can be employed to address the issue. PCORI could provide an opportunity to gather qualitative and quantitative information to advance understanding of response shift, and further investigate how best to communicate any findings and associated implications to patients and clinicians.

3.3.5 Developing short forms of existing PROMs

Short form versions of existing instruments aim to increase acceptability and reduce administration burden. A number of well-recognized instruments have short forms that are widely used.

Current techniques used to create short forms include regression or alpha reliabilities. There is little evidence that these techniques address which items on a measure are most relevant to the underlying construct. Advanced psychometric methods such as item response theory (IRT) and Rasch modeling are being used in the instrument development process to evaluate individual items (Turk et al., 2006). IRT is now also used in the development of computer adaptive testing (CAT) approaches which seek to minimize patient burden through selective administration of test items. Consideration of CAT approaches could be of significant benefit to PCOR when examining potential data collection strategies.

Currently there is little agreement on what constitutes a minimum standard for the development of short form instruments. The recent availability of statistical packages with more advanced techniques for assessing item functioning is slowly revolutionizing instrument development. PCOR could benefit from maintaining an awareness of developments in this area and perhaps even shape the research agenda.

3.3.6 Proxy and caregiver measures

For patients unable to self-report (e.g. cognitively impaired patients or infants) proxy measures can be used. Proxy measures should be limited to observable events or behaviors, as it may be difficult for a proxy reported to know how the patient is feeling and to avoid bias (FDA, 2009; ISOQOL, 2009). The development and selection of proxy measures for PCOR may require some unique minimum standards, and should be addressed in PCORI's future work.

For patients receiving informal care, the impact of their disease and/or treatment on their caregiver may have an impact on their decision making. The impact of a patient's disease or treatment on a caregiver's health-related quality of life is recognized by institutions such as the National Institute for Health and Clinical Excellence (NICE, 2008) in their health technology appraisals. The caregiver voice could be considered in PCOR through the use of caregiver outcome measures, which could be developed with the same qualitative and quantitative standards. Further research is required to establish how the caregiver voice should be incorporated in PCOR.

3.3.7 Patient involvement beyond the development of a PROM

Representing the patient's voice in PCOR can be achieved as follows:

- When the study design does not employ the use of PRO instruments (e.g. retrospective chart review), investigators may consider reviewing a sample of original CER data for reference to patients' subjective experience (preceded by appropriate ethical review). Alternatively, patients could be involved in the decision making process around which clinical outcomes should be included.
- When PROM data are available in retrospective analysis, patients could be involved in selecting the outcomes most relevant to their experience of disease or treatment.
- In prospective studies, investigators could include lay representatives (patient or care-giver) in study steering groups or consult patient associations for guiding research questions. This approach will enhance validity and relevance of research.

3.3.8 Communication of PRO research to patients

In the past, healthcare decision making relied mostly on the physician's clinical experience and data from medical tests. More recently, healthcare has been moving toward a "patient-centered model" emphasizing patients' active participation, which can lead to better quality data.

PROMs can facilitate clinician-patient interactions because they can illuminate the needs and concerns of patients (Lohr and Zebrack, 2009). PROM results need to be meaningful to clinicians and patients in order to contribute to their decision-making processes. Incorporating the use of PROMs into clinical training could positively influence clinicians' understanding of PROMs. Continued research should ensure PROM data are meaningfully interpreted and clearly presented to maximize use and patient involvement. Further research is required to increase patient involvement in the interpretation of PROMs. Ideally, easily interpretable and clearly presented PROM data relevant to diseases/treatments would be readily available to patients to facilitate patients' decision making, which in turn could improve patients' self-efficacy and satisfaction with care.

3.3.9 Feasibility of use in low literacy and/or non-English speaking patients

PROMs also need to be comprehensible across patient populations, including patients with low literacy levels or poor English language fluency. Many PROMS are not suitable for use in such populations, creating issues around validity and missing data.

There is no easy way to address the issue of low literacy. Patients with lower levels of education and various linguistic backgrounds should be included in cognitive debriefing samples to identify areas of difficulty. In addition, areas of difficulties identified by cognitive debriefing with low literacy patients may not easily be addressed through traditional instrument development options. A related challenge of low participation among these populations may arise.

Some modes of PROM administration, such as the telephone based interactive voice response (IVR) platform or ePRO, have potential but they may not be suitable for all types of questions and response options.

4 Conclusions / Next Steps

The development of these minimum standards and other considerations for PCOR has been undertaken with several objectives in mind. The content has drawn on existing standards and guidelines that have largely been developed in relation to clinical trial research; adapting these existing standards through discussion and debate to make them suitable for use to support PCOR and comparative effectiveness research.

The minimum standards have also been designed to reflect how PROs have been used in different contexts and for different purposes. Health Technology Assessment, individual decision making and clinical audits all impose different constraints on the way that outcomes measures are used, analyzed and interpreted because of their different aims. This has provided very useful insights that the field of comparative effectiveness can learn from. We have attempted to incorporate these perspectives and insights into our minimum standards.

In developing the standards we were conscious that it is important that they do not become a barrier to PCOR. If too many standards are developed and are all set too high then no PRO will meet that criteria and the field will be held back. In addition comparative effectiveness research itself provides an opportunity through the data that are collected to address shortcomings in our evidence base. Researchers may be able to proceed with PROMs that don't meet the standards if the research studies provide an opportunity to address these issues.

It should also be noted that the minimum standards come with certain limitations or caveats. They were developed through expert consensus and we believe in large part are generally not controversial. However the standards have not been applied to actual studies. The standards lean on clinical trial methodologies and so may not suit all applications in comparative effectiveness and PCOR. The standards may not apply equally well in all clinical scenarios or therapeutic areas. And as more PCOR studies are initiated it may become apparent that more minimum standards are required to address topics that we haven't covered. As such this is working document which needs revisiting and updating over time. PCOR is an emerging area which is building on the work of different fields of research. At this early stage we should continue to appraise procedures and standards to determine they are fit for purpose.

References

Anastasi A. (1998). Psychological Testing. 6th ed. New York, NY: Macmillan.

Barclay-Goddard, R., Epstein, J. D., & Mayo, N. E. (2009). Response shift: a brief overview and proposed research priorities. *Quality of Life Research*, 18, 335-346.

Basch E, Abernethy AP. (2011). Supporting clinical decisions with real-time patient-reported outcomes. *Journal Clinical Oncology*; 29(8):954-56.

Basch, E., Abernethy, AP, Mullins, DC, Spencer, M. (2011). *Recommendations for incorporating patient-reported outcomes into the design of clinical trials in adult oncology.* Center for Medical Technology and Policy.

Basch E, Abernethy AP, Reeve BB. (2011). Assuring the patient centredness of patient-reported oucomes: content validity in medical product development and comparative effectiveness research. *Value in Health.* 14: 965-966.

Bjorner JB, Ware JE, Jr. Using Modern Psychometric Methods to Measure Health Outcomes. Medical Outcomes Trust Monitor 1998; 3(2):11-6. [SF-36]

Bowden A, Fox-Rushby JA. (2003). A systematic and critical review of the process of translation and adaptation of generic health-related wuality of life measures in Africa, Asia, Eastern Europe, the Middle East, South America. *Social Science & Medicine*. 57:1289-1306.

Bridges, JFP, Hauber, AB, Marshall, DA, Lloyd AJ et al. (2011). Conjoint Analysis Applications in Health—A Checklist: A Report of the ISPOR Good Research Practices for Conjoint Analysis Task Force Value Health. 14:403-13.

Brod M, Tesler LE, Christensen TL. (2009). Qualitative research and content validity: developing best practices based on science and experience. *Quality of Life Research*; 18:1263-78.

Burke LB, Kennedy DL, Miskala PH, et al. (2008). The use of patient-reported outcome measures in the evaluation of medical products for regulatory approval. *Clin Pharmacol Ther*. 84(2):281-3.

Burke LB and Trenacosti AM. (2010, May). Interpretation of PRO trial results to support FDA labelling claims: the regulator perspective. Presented at the International Society for Pharmacoecomomics and Outcomes Research 15th Annual International Meeting, Atlanta, USA.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155-159.

Coons SJ, Gwaltney CJ, Hays RD, Lundy JJ, Sloan JA, PhD, Revicki DA, Lenderking WR, Cella D, Basch E (2009), on behalf of the ISPOR ePRO Task Force. 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.

Cortina, J. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78 (1), 98-104.

Dawson J, Boller I, Doll H, Lavis G, Sharp R, Cooke P, Jenkinson C. The MOXFQ patient-reported questionnaire: assessment of data quality, reliability and validity in relation to foot and ankle surgery. Foot (Edinb). 2011 Jun;21(2):92-102. Epub 2011 May 23.

Devlin and Appleby (2010). *Getting the most out of PROMs: Putting health outcomes at the heart of NHS decision making.* London: The King's Fund.

DeWalt DA, Rothrock N, Yount S, Stone A, (2007) on behalf of the PROMIS Cooperative Group. Evaluation of item candidates – the PROMIS qualitative item review. Med Care; 45(Suppl.): S12-21.

Doward LC, Gnanasakthy A, Baker MG. (2010) Patient reported outcomes: looking beyond the label claim. *Health and Quality of Life Outcomes*. 2(82).

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. (2005) Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. *Pain*;113:9-19

Dworkin RH, Turk DC, Wyrwich KW et al., (2008) Interpreting the clinical importance of of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. *J Pain;* 9(2): 105 – 121.

Eremenco S, Paty J & Lloyd AJ. (2011) Study Designs to Evaluate Multiple Modes of Administration and Data Capture ISPOR European Conference Madrid.

European Medicines Agency (2005). *Reflection paper on the regulatory guidance for use of health-related quality of life (HRQOL) measures in the evaluation of medicinal products.*

Expanding Patient-Centered Care To Empower Patients and Assist Providers. *Research in Action*, Issue 5. AHRQ Publication No. 02-0024. May 2002. Agency for Healthcare Research and Quality, Rockville, MD. http://www.ahrq.gov/qual/ptcareria.htm

Fitzpatrick, R., Davey, C., Buxton, M. & Jones, D. (1998). Evaluating patient-based outcome measures for use in clinical trials. *Health Technology Assessment,* 2(14).

Gwaltney CJ, Shields AL, Shiffman S. (2008) Equivalence of electronic and paper-and-pencil administration of patient reported outcome measures: a meta-analytic review. Value Health; 11:322–33.

Hay J, Atkinson TM, Mendoza TR et al. (2010) Refinement of the patient-reported outcomes version of the common terminology criteria for adverse events (PRO-CTCAE) via cognitive interviewing. Journal of Clinical Oncology; 28(Supl.):15s; abstr 9060.

Herdman, M, Gudex, C, Lloyd, A, Janssen, MF, Kind,P, Parkin, D, Bonsel, G, Badia, X. (2011). Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Quality of Life Research,* 20: 1727-1736

International Society for Quality of Life Research. (November 11, 2011) (prepared by Aaronson N, Choucair A, Elliott T, Greenhalgh J, Halyard M, Hess R, Miller D, Reeve B, Santana M, Snyder C). User's Guide to Implementing Patient-Reported Outcomes Assessment in Clinical Practice.

Johnson FR, Ozdemir S, Manjunath R, Hauber AB, Burch SP, Thompson TR. (2007). Factors that affect adherence to bipolar disorder treatments: a stated-preference approach.

Keller, SD, Bayliss, MS, Ware, JE, Hsu, MA, Damiano, AM, Goss, TF (1996). Comparison of responses to SF-36 health survey questions with one-week and four-week recall periods. *Health Services Research,* 32(3): 367-384

Kerr C, Nixon A, Wild D. (2010) Assessing and demonstrating data saturation in qualitative inquiry supporting patient-reported outcomes research. *Expert Rev Pharmacoeconomics Outcomes Res*. 10(3):269-81.

King, MT. (2011) A point of minimal important difference (MID): a critique of terminology and methods. *Expert Rev. Pharmacoeconomics Outcomes Res*. 11(2): 1171-84.

Lasch KE, Marquis P, Vigneux M, et al. (2010) PRO development: rigorous qualitative research as the crucial foundation. *Quality of Life Research*.

Lloyd AJ & EuroQol Group Digital Task Force Draft guidelines for the adoption of ePRO methods, EuroQol Group, September 2008

Lohr KN, Zebrack BJ. (2009) Using patient-reported outcomes in clinical practice: challenges and opportunities. *Quality of Life Research*; 18:99-107.

McNair AGK, Brookes ST, Davis CR, Argyropoulos M, Blazeby JM. (2010) Communicating the results of randomized clinical trials: do patients understand multidimensional patient-reported outcomes? *Journal of Clinical Oncology*; 28:738-743.

Mokkink, LB, Terwee, CB, Patrick, DL, Alonso, J, Stratford, PW, Knol, DL, Bouter, LM, de Vet, HCW. (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*, 19:539–549.

National Center for Education Statistics. Adult literacy in America: A first look at the findings of the National Adult Literacy Survey. (April 2002). Available at http://nces.ed.gov/pubs93/93275.pdf Accessed February 14, 2012.

National Institute for Health and Clinical Excellence (2008). Guide to the methods of technology appraisal.

Norman, G., Stratford, P., & Regehr, G. (1997). Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. *Journal of Clinical Epidemiology,* 50 (8), 869-879.

Norquist JM, Girman C, Fehnel S, et al. (2011) Choice of recall period for patient-reported outcome (PRO) measures: criteria for consideration. *Quality of Life Research*.

Nunnally JC (1978). Psychometric Theory. 2nd Ed. McGraw Hill: New York.

Obosa, D, Zee, B, Pater, J, et al. (1994). Psychometric properties and responsiveness of the EORTC Quality of Life Questionnaire in patients with breast, ovarian and lung cancer. *Quality of Life Research,* 3: 353-364

Obosa, D, Aaronson, N, Zee, B, et al. (1997). Modification of the EORTC QLQ-30 (version 2.0) based on content validity and reliability testing in large samples or patients with cancer. *Quality of Life Research* 6: 103-108

O'Boyle CA, McGee H, Browne JP. Measuring response shift using the Schedule for Evaluation of Individual Quality of Life. 129 (2000) in: Schwarz CE, Spranger MAG, editors. Adaptation to changing health: response shift in quality-of-life research. Washington (DC): American Psychological Association; P 123-36.

Osborne R, Hawkins M, Sprangers M. (2006) Change of perspective: A measureable and desired outcome of chronic disease self-management intervention programs that violates the premise of preintervention/postintervention assessment. Arthritis and Rheumatism, 55,458-465.

Patrick DL, Burke LB, Gwaltney CJ, et al. (2011) Content Validity – Establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO good research practices task force report: Part I – Eliciting concepts for a new PRO instrument. *Value in Health*. In Press.

Patrick DL, Burke LB, Gwaltney CJ, et al. (2011) Content Validity – Establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO good research practices task force report: Part II – Assessing respondent understanding. *Value in Health*. In Press.

Patrick DL, Burke LB, Powers JH, et al. (2007) Patient-reported outcomes to support medical product labeling claims: FDA prespective. *Value in Health*. 10(Suppl 2):S125-37.

Putman and Rothbart (2006). Development of short and very short forms of the Children's Behavioural Questionnaire. JOURNAL OF PERSONALITY ASSESSMENT, 87(1), 103–113

Randolph, C., McCrea, M. & Barr, W.B. (2005) Is neuropsychological testing useful in the management of sport-related concussion? *Journal of Athletic Training,* 40, 139–154.

Revicki D, Osoba D, Fairclough D. et al. (2000) Recommendations on health-related quality of life research to support labelling and promotional claims in the United States. *Qual Life Res*. 9:887-900.

Revicki, D., Cella, D., Hays, R., Sloan, J., Lenderking, W. & Aaronson, N. (2006). Responsiveness and minimal important differences for patient reported outcomes. *Health and Quality of Life Outcomes,* 4 (70).

Revicki, D., Gnanasakthy, A. & Weinfurt, K. (2007). Documenting the rationale and psychometric characteristics of patient reported outcomes for labelling and promotional claims: the PRO Evidence Dossier. *Quality of Life Research*, 16, 717-723.

Ring L, Höfer S, Heuston F, Harris D, O'Boyle CA. (2005) Response shift masks the treatment impact on patient reported outcomes (PROs): the example of individual quality of life in edentulous patient. *Health and Quality of Life Outcome*; 3:55.

Rothman M, Burke L, Erickson P, et al. (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):1075-81.

Rust J and Golombok S (2009). Modern Psychometrics: The Science of Psychological Assessment. 3rd Ed. Routledge: New York.

Scholtes VA, Terwee CB, Poolman RW. (2011) What makes a measurement instrument valid and reliable? *Injury Int J Care Injured*. 42:236-40.

Schwartz, C. E. Applications of response shift theory and methods to participation measurement: A brief history of a young field (2010). *Archives of Physical Medicine and Rehabilitation*, 91 (9 SUPPL.) S38-S43.

Schwartz CE, Sprangers MA. (1999) Methodological approaches for assessing response shift in longitudinal health-related quality of life research. *Social Science and Medicine, 48, 1531-1548.*

Skevington, S.M., Lotfy, M. & O'Connell, K.A. (2004). The World Health organization's WHOQOL-BREF quality of life assessment: psychometric properties and results of the international field trial. *Quality of Life Research*, 13, 299-310.

Snyder, CF, Watson, ME, Jackson, JD, Cella, D, Halyard, MY, the Mayo/FDA Patient-Reported Outcomes Consensus Meeting Group (2007). Patient-Reported Outcome Instrument

Selection: Designing a Measurement Strategy. *Value in Health, 10 (Suppl 2): S76-S85*

Stone AA, Shiffman S, Schwartz JE, et al. (2002) Patient noncompliance with paper diaries. *BMJ*;324:1193–4.

Stratford and Riddle (2005). Assessing sensitivity to change: choosing the appropriate change coefficient. *Health and Quality of Life Outcomes*, 3 (23).

Streiner DL, Norman GR (2003). *Health Measurement Scales. A practical guide to their development and use*. 3rd edition. Oxford (NY): Oxford University Press.

Stull DE, Leidy NK, Parasuraman B, Chassany O. (2009) Optimal recall periods for patient-reported outcomes: challenges and potential solutions. *Curr Med Res Opin*. 25(4):929-42.

Swinburn P & Lloyd AJ. (2010) PDA Draft Design Specification Document for the EuroQol Group Digital Task Force, EuroQol Group.

Szende, A., Schramm, W., Flood, E., Larson, P., Gorina, E., Rentz, A. & Snyder, L. (2003). Health-related quality of life assessment in adult haemophilia patients: a systematic review and evaluation of instruments. *Haemophilia*, 9, 678-687.

Testa MA, Anderson RB, Nackley JF, et al. (1993). Quality of life and antihypertensive therapy in men: a comparison of captopril and enalapril. *N Engl J Med*; 328:907-13.

Turk DC, Dworkin RH, Burke LB, Gershon R, Rothman M, Scott J, Allen RR, Atkinson JH, Chandler J, Cleeland C, Cowan P, Dimitrova R, Dionne R, Farrar JT, Haythornthwaite JA, Hertz S, Jadad AR, Jensen MP, Kellstein D, Kerns RD, Manning DC, Martin S, Max MB, McDermott MP, McGrath P, Moulin DE, Nurmikko T, Quessy S, Raja S, Rappaport BA, Rauschkolb C, Robinson JP, Royal MA, Simon L, Stauffer JW, Stucki G, Tollett J, von Stein T, Wallace MS, Wernicke J, White RE, Williams AC, Witter J, Wyrwich KW. (2006) Developing outcome measures for pain clinical trials: IMMPACT recommendations. *Pain*;125:208-215

Turner, RR, Quittner, AL, Parasuraman, BM, Kallich, JD, Cleeland, CS (2007). Patient-reported outcomes: Instrument development and selection issues. *Value in Health*, 10(Suppl2): S86-S93

U.S. Census Bureau, 2010 American Community Survey. Available at http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_10_1YR_S1602&prodType=table Accessed February 07, 2012.

US Food and Drug Administration (2009). *Guidance for Industry. Patient-reported outcome measures: Use in medical product development of support labelling claims.*

US Food and Drug Administration (2009). *Patient-reported measures: announcing FDA's final PRO guidance.*

Ward, W., Hahn, E., Mo, F., Hernandez, L., Tulsky, D. & Cella, D. (1999). Reliability and validity of the Functional Assessment of Cancer Therapy-Colorectal (FACT-C) quality of life instrument. *Quality of Life Research*, 8, 181-195.

Ware, J.E. Jr, Kosinski, M. & Keller, S. (1996). A 12-item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. *Medical Care,* 34 (3), 220-223.

Weiss BD, Blanchard JS, McGee DL, et al. (1994). Illiteracy among Medicaid recipients and its relationship to health care costs. *J Health Care Poor Underserved*;5:99–111.

WHOQOL Group. (1995) The World Health Organization Quality of Life assessment

(WHOQOL): position paper from the World Health Organization. *Soc Sci Med*. 41(10):1403-9.

Wild D, Grove A, Martin M, Eremenco S, McElroy S, Verjee-Lorenz A, Erikson, P. (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.

APPENDIX A: Minimum Standards

<table>
<tr><th>Name of standard</th><th>Consideration of Patient Burden</th></tr>
<tr><td>Description of standard</td><td>Patient burden refers to the time, effort and emotional strain associated with completing a PROM. Patient burden should be carefully considered when selecting, developing or using a PROM, and every effort should be made to minimize the burden on the patient. Factors affecting burden that need to be considered include:

The PROM
Number of items
Complexity
Format
Mode of administration
Recall period
Perceived relevance
The study design
Timing and frequency of assessments
Number of PROMs included
Conceptual overlap between administered measures
The patient
Disease characteristics (severity, stage of disease, symptoms, treatment toxicity, etc.)
Other characteristics (age, cognitive ability, visual acuity, motivation, etc.)

Burden on proxy reporters (i.e., caregivers, parents, etc.) should also be minimized.</td></tr>
<tr><td>Current Practice and Examples</td><td>The minimum standard is consistent with many advances in current PROM research, such as the increased use of ePROMs. However the extent to which patient burden is actually considered in current practice is likely to vary.

Advances in the use of technology have been implemented to help ease patient burden and increase reliability of the data. For example electronic PRO formats can be more convenient for patients to complete, can include reminders and allow questionnaires to be tailored to the patient so that non-applicable items can be skipped.

The National Institutes of Health (NIH) funded the Patient-</td></tr>
</table>

	Reported Outcomes Measurement Information System (PROMIS; www.nihpromis.org/) with the goal of developing calibrated item banks for standardized measures of generic health domains. Specific questions can be selected from the PROMIS item banks to form a targeted short-form measure with the fewest number of items for the given patient population. By using computer adaptive testing, item banks are tailored to any patient and to the precise health status of the patient. Such an approach represents an efficient assessment method, thereby potentially addressing patient burden. Increased focus on content validity (FDA PRO Guidance 2009), has also highlighted the importance of cognitive debriefing which assesses patients understanding and interpretation of a PROMs items, instructions, format, usability, recall period, and item redundancy etc.
Published Guidance	The minimum standard described above is consistent with the FDA PRO Final Guidance (2009), as well as the Center for Medical Technology Policy (CMTP) Recommendations for Incorporating Patient-Reported Outcomes into the Design of Clinical Trials in Adult Oncology.
Contribution to Patient Centeredness	By considering patient burden this standard recognizes the value and importance of patients' contributions to research, and aims to maximize data collection and quality by minimizing burden. This standard also protects the interests of patients participating in PCOR. Whereas patient burden should be considered in all research settings, PCOR should ensure the concepts of interest to the patient are prioritized when making any burden versus PROMs inclusion decisions, as discussed in the 'other considerations' section and elsewhere in the report.
Contribution to Transparency	Minimizing patient burden will reduce the risk of missing or erroneous data and improve reliability.
Empirical evidence and theoretical basis	As mentioned in the literature (Turner et al. 2007; CMPT PRO guidance 2011), arguments for the current minimum standard are to: 1) Minimize missing data 2) Minimize erroneous data 3) Reduce non-compliance
Degree of Implementation	The minimum standard suggested here is consistent with current FDA guidance (2009) and CMTP Recommendations. While this could be adopted without significant implementation issues, a system for monitoring / documenting compliance may

Issues	be required.
Other Consideration	There is a current trend for PROMs to focus on single narrow concepts. Therefore, if a more general assessment is required it may necessitate the use of multiple measures. PCOR is more likely to focus broadly on the overall impact of a disease and treatment, addressing several concepts, such as symptoms, functioning, side effects and treatment satisfaction. Until more PCOR focused multi-dimensional PROMs are developed (based on qualitative input from patients regarding their key concerns), patients could be involved in the consideration of which PRO concepts to include in PCOR; maximizing patient relevance and minimizing burden.

Name of standard	Estimating and reporting reliability of a patient reported outcome measure (PROM)
Description of standard	Psychometric reliability (the extent to which a PROM reflects true rather than random variability) should be estimated and clearly reported for any PROM developed or selected for use in patient centered outcomes research (PCOR). There are two separate aspects of reliability – internal consistency and test-retest reliability. Internal consistency is most commonly estimated using Cronbach's α. Test-retest reliability is assessed using the intra-class correlation coefficient (ICC), but other analyses such as the t-test can be used to identify systematic error. Test-retest reliability should be assessed in patients for whom there is evidence that they are stable over the test period. Stability (with respect to the concept measured in the PROM) can be assessed based on self-report, use of other PROMs or sometimes through the use of clinical indices. The ICC coefficient reflects the level of agreement and disagreement between administrations of a PROM, but a high ICC could still mask a systematic error. Therefore in addition to the ICC a test of difference should be estimated such as the paired t-test. The widely reported 0.7 – 0.9 range for α coefficients and >0.70 for ICC coefficients (Nunnally 1978; Fitzpatrick et al., 1998; Streiner and Norman, 2003) should be considered as threshold guidelines. However, since all influence the reliability estimate, the specific population used, sample size, characteristics of the PROM (e.g. number of items, domains and recall period), type of reliability assessed, methods of data collection (e.g. interval(s) between administrations) and the specific test used should be reported. In particular, when developing or selecting a measure with multiple domains, reliability needs to be estimated and clearly reported for each domain. While the appropriate technique(s) for estimating reliability (internal consistency or test retest reliability)will vary dependent upon the nature of the PROM, the population and the concept being assessed, all appropriate techniques should be utilized and reported. *This minimum standard does not detail an absolute value as a threshold of reliability; however the widely reported 0.7 – 0.9 range for α coefficients and >0.70 for ICC coefficients (Nunnally 1978; Streiner and Norman, 2003) are suggested as guidelines. Given the impact of sample size and number of items on the estimation of reliability it is believed the minimum standard detailed above will permit a meaningful assessment of the reliability of a PRO in a specific context. Furthermore this standard will facilitate the potential pooling or meta analyses of reliability data. Similarly specific types of reliability, specific tests, and specific methods, have not been detailed as these will vary dependent upon the nature of the PROM, the population*

	and the concept being assessed; hence the final clause of the standard, ensuring all that could be done is clearly reported.
Current Practice and Examples	In current practice the generally accepted standard for evidence of internal consistency reliability is an α coefficient between 0.70 and 0.90 and of test-retest reliability of >0.70 (Nunnally, 1978; Fitzpatrick et al., 1998; Streiner and Norman, 2003). However, all sources recognize the impact of population, sample size, type of reliability test, and the purpose of the measure as contributing to the reliability estimate, and the definition of a required standard. As such, acceptable reliability coefficients of above 0.60 (Anastasi, 1998), between 0.60 and 0.70 (Ward et al., 1999) and above 0.90 (Szende et al., 2003), have been reported in the PRO literature. Furthermore, higher thresholds (>0.90) have been reported in the context of making individual treatment decisions (Randolph et al, 2005).
Published Guidance	The FDA PRO Final Guidance (2009) requires good internal consistency and recommends that, when appropriate, test-retest reliability is carried out over a period that minimizes memory effects. Revicki et al., (2007) suggest that reliability should be assessed through internal consistency and test-retest reliability where possible and all available information summarized. Both documents support the proposed minimum standard as no specific coefficients are cited, the conduct of all possible tests is encouraged and all available information is requested. In contrast, it has been suggested (e.g., Nunnally, 1978; Fitzpatrick et al., 1998; Randolph et al, 2005) that a test-retest reliability of 0.90 should be required for PROMs used in individual (as opposed to group) clinical decision making.
Contribution to Patient Centeredness	This standard is not entirely unique to PCOR; it applies equally to all PROM research settings. However, any assessment of the psychometric properties of an instrument contributes to the voice of the patient being captured more accurately. In addition, using reliable measures allows clinicians to feel more confident in using these measures to collect data and communicate the findings with their patients. *The consideration of the application of PROMs to individual decision making however, may require specific reliability standards (see discussion in other considerations).*
Contribution to Scientific Rigor	The assessment of reliability ensures the PROM is sufficiently stable across items and over time. By requiring all possible reliability estimates are explored and reported in detail, the long term accumulation of the psychometric properties of a PROM are more readily achievable, for example through data pooling or meta-analysis. Furthermore, detailed reporting guides scientific judgment of the reliability coefficients presented.
Contribution to Transparency	This standard will contribute to ensuring consistency of scores on a measure. Furthermore, the detailed reporting of all

	reliability estimates will permit users to assess the strengths and weaknesses of the measures' reliability.
Empirical evidence and theoretical basis	The assessment of reliability is essential to establish the extent to which a PRO is measuring anything at all (Streiner and Norman, 2003; Rust and Golombok, 2009). However, most sources recognize the impact of the population, the sample size, the properties of the measure, and the types of test used, and thus warn against relying on a single coefficient value or a single data source (e.g. Cortina, 1993). Thus, as it has been said that 'the most important element in the use of reliability coefficients is human judgment' (Rust and Golombok, 2009), detailed reporting of all available information appears the most appropriate basis for a minimum standard. Different types of reliability assessment (e.g. internal consistency and test-retest reliability) will produce different estimates of reliability; likewise different estimates can be produced by different tests of the same type of reliability (e.g. different types of intra-class correlation coefficients assessing test-retest reliability).
Degree of Implementation Issues	Assessing reliability of measures has long been adopted and is in line with current standards. This is a very standard aspect of PROM development and validation. As such, this standard is likely to be adopted without implementation issues. However, implementing the detailed reporting may be problematic; journal editors would need to be aware of the standard, unless as suggested elsewhere a central repository for PROMs developed or selected for PCOR is established. The development of such a system, monitoring materials to ensure compliance with the appropriate standards for PCOR, would have implementation issues, largely due to time and cost.
Other Considerations	Internal consistency and test-re test reliability are not always appropriate, for example when assessing single item concepts or highly variable / transient symptoms, respectively. The appropriateness of the reliability coefficients reported should be considered in context of the PROM and the population. Where test-re test and internal consistency are both appropriate techniques it is possible to see mixed results (e.g. high internal consistency with low test – retest reliability). In such instances it is crucial to have the maximum amount of information available in order to identify where the problems may be (and rectify them if possible) and make a judgment on the PROMs reliability. Relatedly, while poor reliability might indicate that validity is also poor, good reliability, although essential, is meaningless if the PROM is not valid. Good reliability does also not mean that the measure is in any way valid. Therefore, this minimum standard needs to be considered in conjunction with all other standards.

	As detailed elsewhere the psychometric properties of a PROM will become established overtime as more information is gathered (e.g. different populations, or sample sizes). However, stipulating the amount of information required may be overly restrictive at this stage. Reliability estimates should be critically reviewed and where possible verified in new samples and with different analyses as data become available When PCORI begin to consider the development and use of PROMs in daily clinical practice, the standards presented here would need to be modified to reflect individual level data, such as the 0.90 test retest reliability expectations associated with clinical decision making (e.g., Nunnally, 1978; Fitzpatrick et al., 1998; Randolph et al, 2005).

Name of standard	**Training research staff on the administration of patient reported outcome measures (PROMs)**
Description of standard	Research staff involved in the administration of patient reported outcome measures (PROMs) should receive instruction on a number of aspects related to specific patient data collection efforts. Training should be provided on the purpose and intent of PROM data collection, PROM administration procedures (including the timing and frequency of PROM administration), any completion checking processes to be undertaken and the handling and storage of resulting PROM data. Staff training should be facilitated through use of the PROM user manual, provided by the developer, and the study/trial protocol. Staff training related to communicating with patients regarding PROMs should include the following specific staff responsibilities: when invited to take part in a study, patients should be given an opportunity to ask questions of the research staff. A brief verbal instruction about PROM completion should be given, in addition to written instructions on the PROM itself. Furthermore, it is important to reinforce that participation is voluntary and responses are confidential. Where a PROM deals with concerns of a potentially sensitive nature, staff should take adequate precautions to ensure that administration is conducted in an appropriate fashion (the patient is in a suitably private location, will be not interrupted etc.). For PROM administration patients should have access to any necessary reading aids and it is essential that any reading or writing assistance given by research staff does not interpret the PROM questions for the patient. The use of devices for electronic data capture will also need to be fully explained. *The desired standard would also involve the incorporation of detailed instructions to research staff for PROM administration and communication with patients in a well-developed user's manual. That is, in addition to general information about the PROM and the scoring instructions (which can typically be found in current PROM user manuals provided by developers), research staff would also see specific, detailed instructions on how the PROM should be administered to patients (i.e., the user manual would be used as a training tool for research staff) This would ensure that the PROM is always being administered accurately and consistently within and across research centers.*
Current Practice and Examples	Research staff training on the administration of PROMs is likely to vary based on the research setting and study design. There is likely to be a wide range of current practices regarding staff training ranging from no specific PROM training to dedicated investigators meetings with specific portions of the meeting dedicated to PRO data collection. While some studies may

	include staff training on PROMs within the study protocol, this is unlikely to be implemented consistently. In addition, some studies incorporate a site reference binder that is used for training purposes. Specific aspects of the binder can be dedicated to PRO data collection. Staff training on the binder is implemented via in-person site visits or via web or teleconference.
Published Guidance	The FDA PRO Final Guidance (2009) suggests that study quality can be enhanced through standardized instruction for investigators, and increased training for clinicians/ research staff and patients will be critical to ensure that PROMs are used and applied effectively. The user manual provided by the developer for the PROM should specify how to incorporate the PROM into a study/trial in a way that minimizes patient and administrator burden, and missing and poor quality data. This manual should also clearly explain the critical principles of PROM administration. Furthermore, the FDA recommends including clear instructions for clinical investigators in the study/trial protocol regarding patient supervision, timing and order of questionnaire administration during or outside the office visit, processes and rules for questionnaire review for completeness, and documentation of how and when data are filed, stored, and transmitted to or from the clinical site. The Center for Medical Technology and Policy (CMTP, 2011) also specifies that the protocol should include a plan for consistently, and systematically training research staff to confirm their understanding of the content and importance of PRO data collection within the study/trial design. Regular contact with research staff throughout the study/trial is encouraged. CMTP also suggests that the process for PROM data collection be centrally coordinated through a lead data manager. This person should monitor real-time patient adherence and communicate directly with research staff when non-adherence is apparent to avoid missing PROM data, to the extent possible. All research staff involved in the study/trial should understand the overall study objectives and specific procedures relating to their role in the study. It is the responsibility of the investigator to ensure that research staff members have an appropriate level of training and to review the study procedures and instruments prior to data collection (UK Department of Health, 2005). All guidance documents support the suggested minimum standard.
Contribution to Patient Centeredness	This standard will help ensure that PROM data receives the same amount of attention, careful consideration, and follow up (as appropriate) as other study/trial outcomes outlined in the protocol. Training research staff on the importance of PROM data collection enables the patient voice to be captured in a standardized and consistent fashion, and contribute towards

	ensuring both the integrity and quality of PROM data. While this minimum standard could apply to all research settings, it particularly contributes to Patient Centered research in its focus on training staff not just in the use and administration of the PROM, but on communication with the patient.
Contribution to Scientific Rigor	According to the FDA (2009), "results obtained using a PRO instrument can vary according to the instructions given to patients or the training given to the interviewer or persons supervising PRO data collection during a clinical trial." Systematic and consistent training of research staff on the justification for and administration of PROMs will help to minimize bias since all PROMs would be administered in a standardized way across all research centers. If all research staff are trained on the identification and handling of missing data, this will help ensure more complete reporting of PROM data. In addition, the collection of more complete and accurate PROM data will allow for more valuable communication to patients regarding specific aspects of their disease/treatment.
Contribution to Transparency	This standard addresses explicit methods for and consistent implementation of PROMs training for research staff. Incorporating this level of systematic and consistent training for PROMs across participating research centers would strengthen the design of a study overall. Stressing the importance and value of PROM data helps ensure that the patient voice is adequately and appropriately represented in PCOR.
Empirical evidence and theoretical basis	Theoretical: The main arguments for the minimum standard are: Inconsistent administration of PROMs across research sites and within research sites can affect the quality of the PROMs data that is captured. Research staff should be systematically trained on the administration of PROMs to ensure that PROM data is captured in a standardized way across all research centers. Staff training will help minimize missing PROM data, and will ensure that the review of completed measures and missing data are handled consistently across all research centers.
Degree of Implementation Issues	Research staff should already receive training on study/trial protocols as a requirement for their participation. This standard requests that more attention be paid to PROM administration and data handling. This increased level of staff training on PROMs should be adoptable with minimal cost or other implementation issues when conducted concurrently with other training for the study. This standard is in line with current guidelines, and is likely being incorporated in current study/trial protocols.

Other Considerations	• Some research centers might be better equipped than others to administer PROMs in certain formats (e.g., paper-based versus web-based administration). Consideration should be given to the feasibility of PROM administration at all centers participating in a study/trial. A feasibility questionnaire could be provided to potential research centers prior to enrollment to ensure that the PROM format planned for a given study/trial can actually be implemented in practice. If the PROM is completed at the research center/clinic, it should be administered to the patient prior to any procedures taking place and before the patient is seen by their treating clinician. Patients should be given a quiet place to complete the PROMs-free of distractions, interruptions, etc. If PROMs will be completed by patients at home, detailed instructions should be given to them by the research staff during their baseline study visit, and specific contact information should be provided if the patient has any questions. Reminders should be sent to patients via mail and/or email or ePROM regarding PROM completion throughout the duration of the study. • Clear policies need to be in place relating to confidential treatment and secure storage of raw PROM data at participating research centers. • It would be very difficult to monitor the adequate adoption of this standard across all PCOR studies, research centers and staff. Random on-site visits at the beginning and throughout the duration of a study may be required to ensure that research staff are adequately trained and are communicating consistently with patients. • While this standard focuses on staff training in a study/trial environment, training on administration of PROMs for use in regular clinical practice (to help inform clinician/patient decision making) should also be considered. In terms of implementation within the clinical care setting, clinic staff would need to understand the benefits of the PROM and the desired outcome(s) from its implementation in order to adopt it into every day practice. Staff burden would have to be assessed and minimized. Costs and feasibility could be more challenging in this environment if the funding must come from the clinical practice as opposed to an external sponsor for a specific study protocol.

Name of standard	**Choosing an appropriate recall period for patient reported outcome measures (PROMs)**
Description of standard	The rationale and appropriateness of the recall period should be evaluated when selecting or developing a PROM. The following should be assessed individually: is the recall period appropriate for the purpose/intended use of the PROM, the patient population, the disease/condition, the treatment/device, and the study design? Patient ability to accurately recall the information being asked in the PROM and patient burden should be considered, and patient understanding of the recall period should be assessed during cognitive debriefing. This will help ensure the quality and completeness of the PROM data collected. *A more desirable standard would involve patient contributions during the concept elicitation interviews regarding potential recall periods, as well as fully exploring the issues of frequency and duration of symptoms/impacts, variability of experience and change over time. This would ensure both patient centeredness and clinical context are considered when selecting an appropriate recall period.* *The clinical context as well as the patient voice needs to be considered here due to the variable nature of some conditions (e.g. Relapsing Multiple Sclerosis, or Irritable Bowel Syndrome) and the potential impact of treatment (e.g. receiving chemotherapy).*
Current Practice and Examples	As described by Norquist et al. (2011), the general consensus among PRO experts is that the choice of appropriate recall period depends on many factors, and there is no perfect recall period that would be suitable for all diseases/conditions, patient populations, and study designs. The FDA (PRO Final Guidance, 2009) suggests that requiring patients to rely on memory- recall over a long period of time, compare current state to an earlier state, or average responses over time- could negatively affect the content validity of the PRO. As a result, the FDA recommends including short recall periods that ask about the patient's current or recent state, their best/worst experience over the recall period, or the use of a diary. Other guidelines suggest that longer recall periods could be appropriate in certain situations: Norquist et al., (2011) suggests that PROs requesting patient feedback at the current time or in short-term intervals might require frequent administration that could impact patient compliance, burden, and the quality of the data collected. Specific considerations when selecting the length of the recall period along with examples of appropriate recall periods for different situations/disease areas/conditions are provided by Norquist et al. (2011), Stull et al. (2009), and CMTP (2011).

Published Guidance	The minimum standard described above took into consideration the following published guidance/recommendations: the FDA PRO Final Guidance (2009), Norquist et al., (2011), Stull et al., (2009), and the Effectiveness Guidance Document: Recommendations for Incorporating Patient-Reported Outcomes into the Design of Clinical Trials in Adult Oncology Initiative on Methods by the Center for Medical Technology and Policy (CMTP, 2011).
Contribution to Patient Centeredness	Developing or selecting a recall period that is specific to the patient population and the characteristics of the diseases / conditions of interest helps ensure patient centeredness. Patients should be involved in the PRO development process (through interviews and/or focus groups), and this includes obtaining patient feedback on the appropriateness and understandability of the recall period. In addition, the burden placed on a patient by the recall period and frequency of PROM assessments within a study design are considered as part of this standard. While this standard contributes to patient centeredness, it would equally apply to all research settings involving PROMs.
Contribution to Scientific Rigor	An appropriate recall period should make it easier for patients to respond accurately, which is likely to result in better quality (reduced error) and more complete data (lower rates of missing data where patients felt they couldn't respond for the time period specified). Considering patient burden, due to frequency of assessment and what is being requested by the recall period itself (i.e., the extent of memory required to recall an event), when selecting an appropriate recall period can also help ensure better quality and more complete data. Recall bias or error can be reduced by keeping the recall task straightforward and considering all of the factors that can influence recall.
Contribution to Transparency	Implementation of this standard would allow users to assess whether the recall period is appropriate for the PROM within a particular study design based on documentation of the evidence in support of the recall period. It would also allow for an assessment of the strengths and weaknesses of the PROM based on the factors to be assessed as part of the standard. *A further assessment of the strengths/weaknesses of the PROM could be made based on documentation of the incorporation of patient feedback on the recall period, and the relevancy of recall for specific diseases/conditions and the concept(s) being measured during concept elicitation.*
Empirical evidence and theoretical basis	Theoretical: The main arguments for the minimum standard are: There is not one standard recall period that can be applied to all concepts, diseases, or study designs, and many factors must be

	considered when selecting an appropriate recall period for a PROM. Choosing an inappropriate recall period introduces measurement error, which could make it harder to detect an effect (Stull et al., 2009). Appropriate recall period should make it easier for patients to respond accurately, which is likely to result in better quality (reduced error) and more complete data (lower rates of missing data where patients felt they couldn't respond for the time period specified). Patient burden, in terms of frequency of assessment and what is being requested by the recall period itself (i.e., the extent of memory required to recall an event), is an important consideration for recall period, and placing too much burden on patients can lead to inaccurate or incomplete PROM data. *Theoretical: The main arguments for the desired standard are:* 1) *Involving patient contributions to the selection of recall period during concept elicitation can ensure it is appropriate for the characteristics associated with specific diseases/ conditions.*
Degree of Implementation Issues	Selecting an appropriate recall period is considered part of the PROM development process, and it contributes to the internal validity of the PROM. Patient input on the recall period can be obtained during interviews and focus groups conducted during the concept elicitation and cognitive debriefing stages of PROM development. Thus, this standard should not introduce significant implementation issues if minimum standards for content validity are also met. The more desirable standard including patient discussion regarding recall at concept elicitation is likely to require additional qualitative work when selecting existing PROMs and could lead to additional validation work if the recall period is changed.
Other Considerations	Specific consideration should be given to study design, including observational and comparative effectiveness studies, when selecting an appropriate recall period. The frequency of clinic/office visits suggested by study/trial designs could influence the choice for shorter or longer recall periods. In general, shorter recall periods (e.g., 24 hours, or 1 week at most) can be preferable to longer recall periods mainly because with longer recall data can be heavily biased by current health and any significant events, such as hospitalizations. As noted in this standard, considering the clinical context of the disease/condition is really critical when determining the most appropriate recall period. The recall period that is most appropriate for acute conditions might be completely inappropriate for chronic conditions and vice versa. In CER we want to be able to understand how effective a treatment will be if it is introduced among patients, considering

	everything else they have going on. In some cases, CER will involve large samples/datasets, which would mean less of a need for a PROM recall period to incorporate a lot of time (e.g., 2 weeks or more). The data collection itself and the large number of patients can effectively fill in those missing data points.

Name of standard	**PROM Selection**
Description of standard	PROM selection should be concept driven, with a focus on concepts most meaningful or important to patients, rather than driven by issues such as convenience or historical use. PROMs should be selected on the basis of their content validity, measurement properties and interpretability, related to the concept (s) of interest, with consideration of patient burden and the diversity of samples in which content validity and other measurement properties were established. *Minimum and more desirable standards for each of these attributes are detailed elsewhere in this report.* Where PROMs do not meet the minimum standards detailed in this report, modification will be required in order to bring them in line with the minimum standards using the described appropriate methodologies. The selection criteria and approach to addressing identified gaps in the evidence should be clearly stated in study reports and publications. *A more desirable standard would be to develop and apply explicit criteria for evaluating the PROM with respect to its content validity, psychometric properties (validity, reliability, sensitivity to change), patient burden, and language availability (see for example Scientific Advisory Committee of the Medical Outcomes Trust 2002, FDA 2009, Fitzpatrick et al 1998), with consideration of the diversity of the samples in which these were established (FDA, 2009)*
Current Practice and Examples	Different approaches have been used in the selection of PROMs: (1) PROMs are selected for inclusion in studies because convention has developed around a particular PROM in the disease area. Often in these circumstances PROM selection is guided by clinical experts involved in the study design based on their knowledge of prior use of PROMs in similar studies; (2) PROMs are selected by review of the available PROMs with explicit selection criteria (see for example Turner et al, 2007). The rationale for selection has historically been focused on psychometric properties of PROMs with less focus on the content validity of instruments and their relevance for the patient population; and (3) PROMs are selected because they are 'expected to be affected by experimental therapy' (Snyder et al, 2007) rather than because they reflect content of importance to patients. In response to updated regulatory requirements (FDA, 2009) there has been increasing levels of research to address gaps in PRO instruments evaluated against explicit criteria, although this research is rarely published.
Published Guidance	The Scientific Advisory Committee of the Medical Outcomes Trust (2002) refer to eight attributes for reviewing PROMs for which they have established review criteria based on existing

	standards and evolving practices in the behavioral science and health outcomes fields: conceptual and measurement model, reliability, validity, responsiveness, interpretability, burden, alternative modes of administration, cultural and language adaptations or translations. The Food and Drug Administration (2009) list 14 attributes of PROMs: concepts being measured, number of items, conceptual framework of the instrument, medical condition for intended use, population for intended use, data collection method, administration mode, response options, recall period, scoring, weighting of items or domains, format, respondent burden, translation or cultural adaptation availability. Fitzpatrick et al (1998) endorse a similar set of standards to those outlined above: appropriateness, reliability, validity, responsiveness, precision, interpretability, acceptability, feasibility.
Contribution to Patient Centeredness	This minimum standard will help in the selection of PROMs that represent the fuller patient voice, focusing on concepts/endpoints of interest, relevance and importance to patients rather than concepts of interest specific to a drug or device development program.
Contribution to Scientific Rigor	The minimum standard is likely to result in inclusion of PROMs with a broader conceptual basis (e.g. HRQL, treatment satisfaction, adherence, patient preference), which will provide a more comprehensive understanding of the impact of treatment interventions. It is hoped that the inclusion of PROMs of relevance to patients will lead to more complete reporting as a result of greater patient buy-in to providing complete PRO data. *Selecting PROMs on the basis of explicit criteria and addressing gaps in the evidence through appropriate methodologies will result in the selection of the most valid and reliable measures available.*
Contribution to Transparency	Documenting the PROM selection process in key study documents including related publications increases transparency of the process and permits a more detailed evaluation of the patient centeredness of the research.
Empirical evidence and theoretical basis	Theoretical: The main arguments for the minimum and desired standards are: 1) PROM selection tends to be either ad-hoc or too focused on the intervention rather than taking a true patient-centered approach to PROM selection and this needs to be corrected to ensure that selected PROMs reflect the patient perspective.

	2) PROMs that have been designed to capture data from patients that relate only to issues of specific relevance to a drug development program may be focused on a narrow range of concepts. These PROs are unlikely to have a broad enough content base for real world or non-experimental settings. *3) In most circumstances there are no documented, explicit criteria for PROM selection associated with clinical studies which limits opportunity for objectivity, transparency and scientific rigor.*
Degree of Implementation Issues	The minimum standard outlined here is largely in line with current guidance and should therefore be able to be adopted without significant implementation issues. One ongoing issue will be the process for deciding how to address any gaps in the evidence for selected PROMs, and the time and financial resources required to address those gaps. Linked to this, detailed information regarding the procedures associated with content validity are rarely adequately reported and thus additional evidence is likely to be required in order to ensure the content validity of a selected PROM. This would have time and costs implications for PCOR. Furthermore, while researchers may adopt this standard, monitoring compliance would be problematic unless reporting PROM selection for PCOR was required to be documented. This would have implementation issues related to time and cost for PCORI (or those responsible for the administration of that documentation process) and the study sponsor. However, for PROM selection to be concept driven based on issues most important or meaningful to patients, patient involvement beyond their contribution to the development of a PROM is required, for example as part of a steering group *(this issue is mentioned below and also addressed in more detail elsewhere in the report).* This would be associated with time, cost and logistical issues, as well as decisions around the appropriate level of patient input and the potential differences that may be expressed between patients and clinicians.
Other Considerations	Involvement of patient representatives in the selection process should also be considered. There are no known standards in PROMs research around the involvement of patient representatives in PRO concept selection, although there is a precedent for involving the public/patients in public health research (e.g. INVOLVE, 2003) and decision making (e.g. patient advocates in HTA settings). Minimum standards relating to content validity require more detailed reporting of qualitative approaches to instrument development. These details are frequently unreported and can be overlooked or minimized by scientific journals in favor of

	reporting psychometric data. To prevent duplication of efforts in compiling content validity evidence a central repository for each PROM could be established. In PROM selection a balance of generic and disease-specific PRO instruments to facilitate cross-study comparisons but allow for sensitivity to treatment effect (Revicki et al 2000, Devlin and Appleby 2010) may be beneficial for PCOR. Cross reference the following Minimum Standards: • Content Validity • Recall Period • Sampling • Validity • Reliability • Sensitivity to Change • Interpretation of change • Patient Burden • Multi-mode equivalence • Modification of an Existing PROM

Name of standard	**Interpretation of meaningful change on a patient reported outcome measure (PROM)**
Description of standard	Patient input in the form of patient reported anchor based methods should be the primary source of data in the interpretation of meaningful change on a PROM. Distribution based methods should be used to ensure the level of change identified by patient report is not likely to occur due to chance alone. The interpretation of meaningful change should be established for each domain of a PROM, with domain specific anchors. *While the minimum standard only states the use of patient reported anchor based methods, the desired goal would be for methods involving patient ratings of concept (e.g. How would you rate your fatigue?) at 2 time points rather than patient ratings of concept change over time (e.g. how would you rate your fatigue now compared to your last visit?). The latter more commonly used patient transition questions, rating change over time, are associated with recall bias and response shift, and are inappropriate in some instances. Additionally a more rigorous standard would include consideration of the direction of meaningful change (improved or worsened).* *For PCOR more research is needed to best ensure that the patients' interpretation of meaningful change is captured. Current methods involve patients' assessments of their health state but do not involve their voice in the interpretation of what would be considered a meaningful change to them. A more desirable assessment of meaningful change would involve the patient as an active rather than passive participant.*
Current Practice and Examples	Patient input in the form of global transition questions, assessing patient perception of change over time on the given domain or concept, is one of the most commonly used anchors for defining meaningful change on PRO instruments. This typically involves defining change as minimal, moderate or large based on the global transition response options. The use of patient ratings of concept at two time points, where appropriate, has been requested by the FDA (Burke and Trentacosti, 2010) The general consensus is that patient reported and clinical anchor based methods should provide the primary evidence for meaningful change in PRO scores over time, and that distribution based methods, such as the one Standard Error of Measurement (SEM) and ½ a Standard Deviation (SD) approach should be considered as supportive evidence. Numerous examples of the applications of these methods, along with a detailed review of the evolution of the terminology and methods associated with the interpretation of meaningful change were recently described by King (2011).
Published Guidance	The minimum standard described above is in line with the FDA PRO Final Guidance (2009), which is in turn aligned with

	recommendations by the Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT; Dworkin et al., 2008), and Revicki et al., (2008). *The desired standard using the patient ratings of concept methodology is supported by the FDA (Burke and Trentacosti, 2010).*
Contribution to Patient Centeredness	Involving the patient in the interpretation of meaningful change ensures that the patient's voice is part of the interpretation process. By assessing meaningful change for each domain, multi-dimensional measures, assessing various concepts important to patients (such as symptoms, function, side effects, treatment satisfaction), can be interpreted by and for patients to give a complete profile of a treatments impact, instead of losing vital information for treatment decision making to an overall score. Involving the patient in the definition process also offers the potential to improve the communication of information to patients having already worked with them to identify descriptors of change most understandable or relevant to them when making treatment decisions.
Contribution to Scientific Rigor	By including the use of distribution based methods as well as patient reported anchor based methods in the interpretation of meaningful change the researcher can evaluate whether the level of change reported by patients is not a change that may occur by chance alone, i.e., one that would be expected from the random variability of the measure. This also encourages the development of PRO measures with sound measurement properties in order to increase the likelihood of being able to detect a level of change that is meaningful to patients. The use of domain level interpretations of meaningful change also minimizes loss of information to an overall score. *The adoption of the more rigorous standards mentioned would also reduce error due to recall bias (Norman et al., 1997; Walters and Brazier 2005).*
Contribution to Transparency	This standard includes two contributing methods that will ensure a degree of consistency in the interpretation of meaningful change that is currently missing. The fact that there are different types of patient based anchor methods (e.g. transition questions and current assessment of concept) and distribution based method (e.g. SEM and ½ SD) will permit users to assess the strengths and weaknesses of the meaningful change of the measure.
Empirical evidence and theoretical basis	Theoretical: The main arguments for the minimum standard are: 1) In many circumstances no clinical anchor exists to assess meaningful change (e.g. pain, depression etc.), or that can be reliably associated with the concept of interest 2) In many disease areas there are a priori reasons to believe that clinical change and PROM change will not

	necessarily align. 3) For a change to be described as meaningful to patients the source of the data used as anchor needs to be meaningful 4) The use of distribution based methods ensures that the estimate of meaningful change is not one that would be expected by chance alone 5) Assessing change for each domain where appropriate ensures a patient can be provided information on the effect of a treatment across multiple domains, and make a more informed treatment choice based on all data available (e.g. efficacy and side effects) rather than a composite or aggregate score, across distinct domains, which is potentially less interpretable.
Degree of Implementation Issues	Interpreting meaningful change for PRO domains and overall scores is a process that has long been adopted, though the terminology and methods have varied. Furthermore, the minimum standard suggested here is largely in line with current guidance. Thus this minimum standard is likely to be adopted without significant implementation issues. One consideration in the observational or clinical practice setting would be the length of time between assessments if the patient reported transition question approach was utilized. Equally, any assessment of meaningful change is susceptible to response shift (discussed elsewhere) and therefore consideration of timing is always necessary. *To address the more desirable standard of elevating the patient voice in a definition of meaningful change would involve more qualitative research rather than just collecting quantitative assessment data from transition or global ratings of concept items. This could potentially be included as part of the cognitive debrief interview process, but would need to be established for all populations as appropriate.*
Other Considerations	The level of meaningful change is context dependent (King, 2011; FDA 2009), and may vary by patient group, severity level or other socio-demographic factors; as such any PRO instrument may be associated with numerous definitions of meaningful change. Thus, data should be continually gathered and consolidated over time; interpretability can then develop as the body of evidence accumulates. Equally, as stated as a desired standard, the direction of meaningful change should also be considered; evidence suggests asymmetry in magnitude of meaningful change associated with improvement and worsening (Revicki et al., 2008; prospect theory, Kahneman and Tversky, 1979). Just as the data collected should be monitored and updated over time, so should the methods. This minimum standard

	advocates any patient reported anchor based method could be deemed appropriate, however ideally this would be refined to ratings of concept rather than transition over time. This less prescriptive approach at this stage is to reflect standards that could feasibly be met by some existing PROMs. Equally, further research is needed to extend the role of the patient in these definitions; for example asking the patient more directly about what would be deemed a meaningful change, and exploring the fact that most PRO and global transition or concept items are measured on scales without interval properties. The difference between moderate and severe may not be equal to the difference between mild and moderate, and these distinctions (or baseline states) are rarely considered when defining meaningful change. Additional research could also explore the best way to interpret change for patients, for example a move toward the calibration approach described by Testa et al. (1993), linking change on a PRO score to the impact of life events. The development / evolution of these new methods could result in changing or additional standards as this field evolves to be more patient centered.

Name of standard	**Establishing/ assessing content validity for patient reported outcome measures (PROMs)**
Description of standard	Content validity (i.e., the extent to which the PROM measures the concept of interest) should be established/ assessed and clearly reported for any PROM developed or selected for use in patient centered outcomes research (PCOR). PROMs should reflect what is important to patients, including issues and concerns that are relevant to their disease/condition. When developing a new PROM, qualitative concept elicitation interviews and/or focus groups should be conducted with a sample representing the target patient population to ensure that all relevant and important concepts are included in the PROM based on direct patient input. Qualitative cognitive debriefing interviews should be conducted with a diverse sample of patients prior to use of a PROM to check clarity, relevance, comprehensiveness and ease of completion. When assessing content validity for an existing measure, it is important to determine whether the content of the measure is relevant to the target patient population. In addition, it is important to determine if there are additional concept areas that are not covered by the existing measure that should be added. This can be established through qualitative research- concept elicitation and cognitive debriefing-similar to the development of a new measure, with a focus on the relevancy and comprehensiveness of existing content. *A more desirable standard would be to implement and document all of the following to establish content validity for a PROM; it should be noted that this process is iterative:* 1) *The sample selected for the qualitative research should reflect the target patient population with respect to clinical and sociodemographic characteristics (see Sampling Standard). Well-trained interviewers, focus group facilitators, and analysts should be used for qualitative assessment.* ***Concept Elicitation**: Clear, open-ended questions should be used to allow patients to talk broadly about their disease/ condition to help inform PROM content. Interviews/focus groups should be recorded and transcribed verbatim, and the qualitative data should be analyzed to identify important and relevant concepts described by patients. Development of PROM items should incorporate patient-friendly language.* ***Cognitive Debriefing**: A semi-structured interview guide should be developed to conduct the cognitive debriefing interviews. Any issues that arise with the PROM during the cognitive debriefing process that result in a modification of the measure would require additional cognitive interviews to confirm adequacy of the revisions. The cognitive interviews should be recorded and transcribed verbatim. Results from the interviews should be summarized.* *PROM development can also be supported by reviewing relevant*

	literature and existing measures and speaking with experts. 2) *An Item Tracking Matrix should be developed and include information on the initial development of each item through to the finalization of each item. Rationale should also be provided for all concepts that were included and excluded. (Patrick et al, 2011; Basch et al, 2011)* 3) *A hypothesized conceptual framework can be developed that shows relationships between items, domains, and the concept(s)/ outcome(s) of interest. The framework should be updated throughout the PROM development process.*
Current Practice and Examples	Historically, researchers have paid less attention to establishing content validity than to other measurement properties. (Patrick et al, 2007) In addition, clinicians/researchers have tended to develop content for PROMs based on their expertise, and may not have pilot tested the PROM in the target patient population. Since the finalization of the FDA PRO Guidance (2009), content validity is likely being established for PROMs that are incorporated into clinical trials for the purpose of a PRO labeling claim. However, detailed procedures involved in establishing content validity for a PROM, especially detailed qualitative methods, are rarely published, and current practice varies based on study design and research setting. The EORTC QLQ-C30 and certain FACT measures, such as the FACT&GOG-Ntx (for neurotoxicity), included direct patient input in their development. Another specific example involving extensive reporting for content validity and qualitative methods is the development of the WHOQOL (although in a different setting- cross-cultural content validity for a HRQL measure). Over 4000 respondents in 15 cultural settings were involved in the development and pilot testing of this measure, and the developers made a great effort to include patient consultation on the WHOQOL in the form of focus groups. (WHO, 1995; Bowden and Fox-Rushby, 2003)
Published Guidance	The minimum standard is aligned with the Final FDA PRO Guidance (2009), The Center for medical Technology Policy's Recommendations for Incorporating PROMs in Clinical Trials in Adult Oncology (2011), and recommendations provided in primary literature: Basch et al, 2011; Bowden and Fox-Rushby, 2003; Brod et al, 2009; Doward et al, 2010; Kerr et al, 2010; Lasch et al, 2010; Patrick et al, 2007; Patrick et al, 2011, parts 1 and 2; Scholtes et al, 2011; Turner et al, 2007. Further details for establishing and reporting content validity, including qualitative methods, can be found in these papers.
Contribution to Patient Centeredness	Establishing content validity ensures that the patient voice is incorporated in the PROM development process. Patients contribute directly to the development of the PROM (this can include PROM items, recall period, response options, instructions, and the concept(s) to be evaluated). Direct patient feedback helps ensure that PROM content is completely relevant to their disease/condition. Patients are also involved in discussions about the clarity of the PROM, and their ability to fully

	understand what is being asked of them when they complete the measure. The importance of the patient experience is considered throughout the PROM development process. While this minimum standard could apply to all research settings, it particularly contributes to PCOR in its focus on the potential for the patient to openly discuss the impact of their disease / treatment. This allows for the development of multi-dimensional PROMs based on concepts that are relevant to the patient rather than limiting the assessment to the exploration of concepts of interest to a study sponsor or investigator.
Contribution to Scientific Rigor	Establishing content validity is of primary importance for PROMs used in real-world or non-experimental settings, as it ensures that the content (i.e., items, domains) of the PROM adequately reflects the construct to be measured. It also ensures that the PROM will be clear and understandable to patients, and that the content will reflect what is most relevant and meaningful from the patient's perspective. PROMs that are highly relevant to the patient population of interest will maximize the quality of the data collected. Irrelevant PROM content can alienate respondents, and unnecessarily increase patient burden which can lead to missing data (Doward et al, 2010). *Researcher, interviewer, or interview guide bias can be avoided through the development of a clear, open-ended interview/ discussion guide for use in patient interviews and focus groups (Patrick et al, 2011).*
Contribution to Transparency	Clear documentation of the methods used to support/establish the content validity of a PROM will allow users to determine the strengths/weaknesses of the PROM, and subsequently the studies that incorporate the PROM. *Transparency related to the qualitative methods used to develop PROMs and to establish content validity can be provided through documentation of the conceptual framework from the initial stages of concept elicitation and item development through to the finalized PROM.*
Empirical evidence and theoretical basis	Theoretical: The main arguments for the minimum standard are: 4) PROMs must possess content validity because they are designed to collect data on concepts related to individuals' health experiences-how they feel or function in relation to their disease, condition or treatment. (Patrick et al, 2011) 5) PROMs should be developed based on direct patient input to ensure their relevance to the targeted patient population. PROMs that are highly relevant to a specific patient population of interest will maximize the quality of the data collected. (Doward et al, 2010) 6) Qualitative data are necessary for establishing content validity. While quantitative data (factor analysis, Rasch analysis, item response theory) can be supportive, they are insufficient without qualitative data. (Patrick et al, 2011) *7) Qualitative research that is rigorous and well-documented provides evidence that items, domains, and concepts in a PROM are appropriate,*

	comprehensible and interpretable. (Lasch et al, 2010)
Degree of Implementation Issues	The extent to which this standard is currently adopted varies based on study design and research setting. Establishing content validity for newly developed PROMs is a time consuming and costly process. It requires a great deal of planning and resources. The details of procedures used to establish content validity of a PROM are rarely adequately reported, therefore, assessing the content validity of an existing PROM could be difficult, and additional evidence would likely be needed to ensure content validity. Similar to newly developed PROMs, this could have significant implications for costs and timelines, especially related to PCOR. Content validity standards require more detailed reporting of qualitative approaches to instrument development. These details are frequently unreported and can be overlooked/ minimized by scientific journals in favor of reporting psychometric data. For example, item tracking matrices include a significant amount of data that would never be published. To prevent duplication of efforts in compiling content validity evidence a central repository for each PROM could be established. While a central repository would require substantial costs and resources to develop, it could save costs eventually by helping to avoid duplication of efforts.
Other Considerations	• Content validity can also be supported (established/assessed) by speaking with experts. Experts can assist with the development of the discussion guides, PROM items, response options and recall period (based on disease and treatment characteristics), and the finalization of the PROM content. Involving both clinicians and patients in the development of a PROM can raise the issue of which perspective is most important. For research that is truly meant to be patient-centered, it might make sense to prioritize the patient perspective. However, for PROMs meant to be used in clinical practice the clinician perspective would be important to incorporate. • In addition to qualitative methods, quantitative methods such as factor analysis, classic test theory, item response theory or Rasch analysis can be used to establish content validity. Combining both qualitative and quantitative methods can add value when establishing good content validity for a PROM. • Translatability and linguistic validation are important to keep in mind when establishing content validity during PROM development. Assessing translatability ensures that the PROM can be easily translated assuming there are no cultural reasons to change the content of the measure. For example, items should avoid culture-specific language (e.g., shoveling snow, buttoning buttons, gardening, etc.), and items should be developed in a way that would not cause grammatical problems in certain countries (e.g., avoid the item format of having a single question stem as the start of a list of multiple items). • Change in mode of administration can affect content validity. Unique features of electronic PROMs should be considered when establishing content validity. These unique features and overall usability can be

	addressed in cognitive debriefing interviews. Paper mock-ups/ screen shots can be used to assess patient understanding prior to finalization of electronic programming, however, cognitive testing on the electronic device itself is recommended. (Patrick et al, 2011)

Name of standard	Sampling in PROM Development / Selection / Validation
Description of standard	The following steps should be followed to identify an appropriate sample for PROM development, selection and psychometric validation. <u>Minimum Standard</u> 1. **Define the Study Population**. The study population from which the sample will be drawn should be clearly defined. The study population for PCOR may be broad to include anyone with the disease/condition of interest or may be narrow to include only those meeting certain criteria, such as a specific stage of disease. The sample should appropriately reflect the study population with respect to clinical and socio-demographic characteristics. a. Qualitative Research: Given the relatively small samples involved in qualitative research, it may not be possible to fully match the target population with respect to all characteristics. Review of literature and expert consultation should be conducted to identify the characteristics that would be most important to target with respect to diversity. b. Quantitative Research (Psychometric Validation). The sample involved in the psychometric validation should be representative of the target population to the extent possible. Review of published epidemiologic data and expert consultation should be conducted to determine requirements for a representative sample 2. **Identify the Sampling Frame and Method**. The source and method of participant recruitment should be clearly stated. Selection of the recruitment source and method should consider potential for sampling bias with the aim of maximizing representativeness and generalizability. **3. Determine Sample Size.** a. Qualitative Research (Concept elicitation interviews/focus groups and cognitive debriefing): Sample size cannot be determined *a priori* in qualitative research. Rather, sample size should be based on information saturation, or the point at which interviews or focus groups are no longer yielding new, relevant information. Once saturation is reached, the sample size is deemed sufficient.

	b. Quantitative Research (Psychometric validation): Sample sizes for psychometric validation should be driven by the statistics being evaluated and the desired level of measurement precision or standard error (Frost et al., 2007). _More Desirable Standard_ 4. ***Qualitative Research: Consider Important Subgroups.*** *There may be important subgroups that should be sampled specifically and sufficiently (ie, to the point of saturation). For example, if substantial differences in disease or treatment impacts are expected for males vs. females, it would be important to sample sufficient numbers of males and females to fully understand the impacts for each subgroup. Additionally, for cognitive debriefing, which is used to ensure the clarity and understandability of the PROM, a diverse sample with respect to education level / literacy should be included. (See the 'other considerations' discussion of literacy elsewhere in the report.)*
Current Practice and Examples	Minimum Standard 1. **Study Population**: Current practice varies, though since the finalization of the FDA PRO Guidance (2009), for PROMs that are incorporated into clinical trials for the purpose of a PRO labeling claim, it has become more common to recruit a sample that closely matches the characteristics of the target clinical trial population in which the PROM will be used and for qualitative research to aim to recruit a diverse sample with respect to age, gender, race/ethnicity, and education level, though the extent to which this is achieved varies. 2. **Sampling Frame and Method**: Current practice is to recruit patients through clinical sites, recruitment agencies, patient advocacy groups and/or print or on-line advertising using a non-probabilistic purposive or convenience sampling method. 3. **Sample Size**: Current practice for qualitative research is to determine sample size based on information saturation, though the extent to which this is actually confirmed varies. Concept elicitation typically involves a minimum of 20 patients, but it is not uncommon to involve 40 patients or more for more complex concepts or more heterogeneous populations. Cognitive debriefing typically involves 10 to 20 patients. For psychometric validation studies, sample sizes often range from 100 to 300 patients, and depending on the analyses, smaller or larger samples may be used. Some guidelines suggest that a validation study should include

	a sample of at least 200 (Frost et al., 2007). *More Desirable Standard* 4. ***Consider Important Subgroups:*** *Current practice varies. Some researchers consider whether subgroups need to be explored, though it is not common practice for saturation to be assessed by subgroup.*
Published Guidance	Minimum Standard The minimum standard described above is in accordance with the FDA PRO Final Guidance (2009), the ISPOR PRO Good Research Practices Task Force Report (Patrick et al. 2011) and the Mayo/FDA PRO Consensus Meeting Group Paper (Frost et al, 2007). *More Desirable Standard* *There is no published guidance specifically related to recruiting subgroups.*
Contribution to Patient Centeredness	Including a representative sample will ensure that results will be generalizable to the target population and that the PROM will capture the important concepts relevant to patients from the target population. Consideration of differences with respect to clinical and sociodemographic characteristics will support the validity of subsequent research for the general population.
Contribution to Scientific Rigor	Content validity of the PROM is based on direct input from a diverse sample reflecting the target population. Including a representative and diverse sample will ensure that the PROM developed measures the intended concept(s) with accounting for variations in patient characteristics and experiences within the target population. Including a representative sample in psychometric validation will yield evidence of reliability and validity across the entire patient population in which the PROM will be used.
Contribution to Transparency	N/A
Empirical evidence and theoretical basis	Minimum Standard Sampling is concerned with the selection of a representative sample of individuals from a target population to evaluate characteristics of the entire population. The main arguments for the minimum standard are to ensure that the sample is representative of the target study population and that the results are generalizable. *More Desirable Standard* *The main argument for the more desirable standard is to ensure*

	that sufficient sample is included to account for subgroups that may have different disease- or treatment- related experiences or impacts.
Degree of Implementation Issues	Minimum Standard It is expected that there will not be major issues with implementation, given that the minimum standard is consistent with published guidelines. One challenge related to implementation, however, is finding a diverse sample with respect to certain sociodemographic characteristics, such as education level and race/ethnicity, particularly for rare conditions. Similarly, finding a representative sample for psychometric validation can be challenging and costly. *More Desirable Standard* *The more desirable standard has the potential to drive up the required sample size, which would make recruitment more challenging, especially for rare conditions.*
Other Considerations	Patients with comorbidities often have difficulty attributing specific symptoms/impacts to a specific condition. Including a subset of patients in concept elicitation who have no comorbidities could be considered. This will ensure that the key symptoms/impacts attributable to the disease are captured in the PROM. Though, including a subset of "pure" patients may be challenging with respect to recruitment, particularly for conditions that are likely to co-occur with others or populations that are likely to have more than one health condition. Recruiting representative samples for psychometric validation studies can be challenging, and CER and registry studies offer an opportunity to re-confirm measurement properties of PROMs. Cross Reference: • Content Validity • PROM Selection • Psychometric Validation (Reliability, Validity, Sensitivity to Change)

Name of standard	Estimating and reporting construct validity of a patient reported outcome measure (PROM)
Description of standard	Construct validity, evidence that a measure assesses the construct / concept it purports to measure, should be estimated and clearly reported for any PROM developed or selected for use in patient centered outcomes research (PCOR). Construct validity is a broad concept, encompassing criterion validity, known groups validity, and predictive validity. Reporting should include a detailed description of the PROM (e.g. number of items, domains and recall period), the population in which validity is being established, the sample size, and the external measure with which it is compared (e.g. valid PROM, clinical outcome measure, objective measure, proxy or caregiver measure). Justification for selection of the external measure should also be provided. When developing or selecting a PROM with multiple domains, construct validity must be established for each domain. If no construct validity data are presented justification should be given as to why this assessment of validity was inappropriate. Construct validity with a continuous external measure is typically assessed using an appropriate correlation coefficient, following an *a priori* hypothesis on the direction and magnitude of association. Additionally, construct validity of the PROM may be assessed by determining whether it can discriminate between categories (known groups validity). When suitable data are available, known groups validity should be assessed to demonstrate that the PROM is able to differentiate among distinct groups (e.g. levels of disease severity). When reporting known groups validity an *a priori* hypothesis should be given to predict the difference between groups, as well as the method used to define group membership, the statistical test used, and effect size. If a PROM is being used to predict an alternative measure such as symptom severity (predictive validity), either considered categorically in terms of group membership (e.g. a PROM threshold to indicate a particular symptom severity) or continuously in terms of the full score range (the PROM score indicating the level of symptom severity) then the appropriate statistical approach taken to assess predictive validity should be clearly reported. For categorical group membership the statistical approach might include Receiver Operating Characteristic (ROC) analysis (with estimation of best cut point and area under the curve), and calculation of diagnostic measures such as sensitivity, specificity, positive and negative predictive values. Where the variable to be predicted is continuous, then an appropriate correlation coefficient should be reported. 95% CI should be presented around all estimates. *A more desirable standard for construct validity would include reporting of convergent and discriminant validity hypotheses a*

	priori. This minimum standard does not detail an absolute value as a threshold of construct validity, due to expected variability based on the PROM concept and the external criterion, but recommends the general practice standard of at least .30 to .50 is used a guideline. Similarly, specific external criteria have not been detailed as these will vary dependent upon the nature of the PROM and the construct being assessed. *Predictive validity can be demonstrated by using an instrument to establish group membership or for forecasting (e.g. treatment success). This standard focuses on predictive validity in the form of predicting or establishing group membership. Forecasting requires an accumulation of data overtime and is briefly discussed elsewhere in the report.* *Known groups' validity and the ability of a PROM to predict group membership or forecast such outcomes as treatment benefit or disease progression, are likely to be highly desirable in PCOR. This information will need to be accumulated over time and standards of contributing research evaluated. How to centralize such a process adequately will need to be explored.* *More specific and desirable standards could be developed as more is understood about PRO concepts and the external criteria with which they are compared or used to predict.*
Current Practice and Examples	In current practice the generally accepted standard for evidence of construct validity by assessing correlation with an external measure is a correlation coefficient of at least .30 to .50 (Streiner and Norman, 2003). The magnitude of the relationship is hypothesis dependent, but if the value of coefficient is too high (e.g. >0.80) this could indicate that the measure is superfluous. In terms of known groups and predictive validity, measures of both statistical and clinical significance are typically presented, through the size of the p-value and the effect size (e.g., Dawson et al, 2011). The effect size, which provides information about the degree of differences between two groups, a t-test being used to calculate statistical significance, is typically used to indicate either a small (.20 and below), medium (.30 to .70), or large (.80 and above) difference (Cohen, 1992). For ANOVA, when three or more groups are compared, the partial eta-squared value provides an estimate of the variation in scores that can be explained by group membership. With very large samples, it is important to distinguish statistical significance, which is more easily detected, from a more practical significance, which is informed by effect size. PROMs are currently being used in clinical practice to identify thresholds for: severity of disease; treatment; and defining treatment success (Devlin and Appleby, 2010).
Published Guidance	The FDA PRO Final Guidance (2009), the EMA Reflection Paper

	for the use of HRQL measures (2005), and the CMTP Effectiveness Guidance (2011) all state validated instruments should be used in research. The FDA guidance refers to *"Evidence that relationships among items, domains, and concepts conform to a priori hypotheses concerning logical relationships that should exist with measures of related concepts or scores produced in similar or diverse patient groups".* This is in line with the current minimum standard when all appropriate data exists. Furthermore, these documents provide no specific guidance regarding correlation coefficients or effect sizes that would be classified as appropriate evidence; again in line with the suggested minimum standard. Fitzpatrick et al. (1998) state that it is necessary to assess construct validity, and that a body of validation evidence can be built up over time as an instrument continues to be used. They further suggest that a correlation coefficient of .60 with established measures may be considered strong evidence of construct validity, but do not specify a minimum coefficient that would be an acceptable indication of validity, instead suggesting that *a priori* minimum levels are set for each study so that validity can be determined in relation to this standard. Published guidance also includes Cohen's d effect size standards (<0.20 = small; 0.30 to 0.70=moderate; >0.80=large) (Cohen, 1992).
Contribution to Patient Centeredness	The contribution of this standard is not unique to PCOR, it is also relevant to any form of outcomes research setting. However, any assessment of the psychometric properties of an instrument contributes to the voice of the patient being captured more accurately. Using valid measures is also essential for clinicians to feel confident in communicating PROM findings with their patients. *The accumulation of predictive validity data, based on patient experience, such as disease progression, remission or treatment response, would be of more value to patient centeredness. This information could be valuable to communicate to patients and inform decision making.*
Contribution to Scientific Rigor	The assessment of construct validity ensures the PROM is measuring what it purports to measure. Furthermore, detailed reporting enables scientific judgment regarding the validity of a PROM. As validity data accumulates more desirable standards can be established.
Contribution to Transparency	This standard will contribute to ensuring that a PROM is measuring what it purports to measure. Furthermore, detailed reporting of validity data will permit users to assess the strengths and weaknesses of the validity of a specific PROM.
Empirical evidence	The assessment of validity is essential to establish the extent to

and theoretical basis	which a PROM is measuring what it purports to measure (Streiner and Norman, 2003). However, establishing validity can be an ongoing process of accumulating evidence to show how well an instrument measures the construct it is intending to assess (Fitzpatrick et al., 1998; Streiner and Norman, 2003). Given that no specific thresholds regarding construct validity can be imposed given variability in constructs assessed and the external criterion selected, detailed reporting of all available information appears the most appropriate basis for a minimum standard. This can be used to help build a body of validity evidence that can be easily assessed during instrument selection.
Degree of Implementation Issues	Assessing validity of measures has long been adopted and is in line with current standards. This standard is likely to be adopted without implementation issues.
Other Considerations	As discussed elsewhere in the report, the psychometric properties of a PROM will become established overtime as more information is gathered (e.g. different clinical populations or ethnic groups). Known groups and predictive validity are likely to be highly desirable in PCOR and efforts to gather good scientific evidence of these properties should be undertaken. Further consideration / research would be required to establish standards related to the role of PROMS in forecasting outcomes, and communicating such information to patients. While validity is essential it is also meaningless if the PROM is not reliable. This minimum standard alone is not sufficient to determine whether or not a PROM is suitable for inclusion in PCOR, and needs to be considered in conjunction with other standards, such as: • Reliability • Sensitivity to change • Interpretation of meaningful change • Content validity

Name of standard	Estimating and reporting ability to detect change in a patient reported outcome measure (PROM)
Description of standard	A PROMs ability to detect change, both improvement and deterioration, should be assessed and clearly reported for any PROM developed or selected for use in patient centered outcomes research (PCOR). As well as demonstrating the ability to detect change, evidence should also be presented to demonstrate stability of a PROM in a clinically stable group. Instruments with multiple domains should establish ability to detect change, and stability in stable patients, for each domain. Reporting of ability to detect change should include a clear statement about how change is assessed or determined (e.g. patient anchor based assessment or clinical outcomes), the target population, statistical test used and effect sizes. *When patient based anchors are used to determine change, assessment of change should not rely on patient recall to assess change over time, rather patient should be asked to complete a patient global impression of concept at different time points in order to determine if there has been change. See the 'Interpretation of Meaningful Change' minimum standard for further discussion of this point.* *A desirable standard would be for a measure to detect more granular level of improvement or worsening, as determined meaningful by patients.*
Current Practice and Examples	Revicki et al., (2006) state that longitudinal studies using external criteria (e.g. patient-reported change, laboratory measures, physiological measures, clinician ratings) are needed to conclude that a measure is sensitive to change across time. The most common assessment of sensitivity to change used in these designs is asking patients to retrospectively rate their global change (Stratford and Riddle, 2005). However, Norman at al., (1997) argue that retrospective global rating can artificially inflate the relationship between ratings, and show change in stable patients.
Published Guidance	The minimum standard described above is in line with the FDA PRO Final Guidance (2009), the EMA Reflection Paper on Regulatory Guidance (2005), and the CMTP Effectiveness Guidance (2011). All of these documents state that ability to detect change should be established in the target population prior to using a PROM, and that it is necessary to show change in both directions.
Contribution to Patient Centeredness	The contribution of this minimum standard is not unique to PCOR, it is also relevant to any form of outcomes research setting. However, the more desirable standard, ensuring the PROM can detect change that is considered meaningful to patients, would be more patient centered and links to the

	'Interpretation' standard. While not unique to PCOR this standard does enhance the reliability and validity of the interpretation of PROMs data when communicating findings to patients.
Contribution to Scientific Rigor	Using a PROM with established ability to detect change enhances the reliability and validity of the interpretation of data collected. The more desirable standard that eliminates any recall bias would also increase the reliability and validity of data collected.
Contribution to Transparency	Demonstrating a PROM is sensitive to clinical or patient reported improvement, deterioration or stability, and clearly reporting the definition of each group permits an evaluation of the PROMs sensitivity in the population studied.
Theoretical basis	The main arguments for the minimum standard are: 8) Assessing ability to detect change ensures a patient can be provided information on the effect of a treatment and make a more informed treatment choice based on all data available (e.g. efficacy over time). 9) If there is clear evidence that patient experience relative to the concept has changed, but the PRO scores do not change, then either the ability to detect change is inadequate or the PRO instrument's validity should be questioned. 10) If a PROM is used to assess the impact of treatment over time, that measure is only useful if it can detect change in response to treatment over time.
Degree of Implementation Issues	Assessing ability to detect change of PROMs can have implementation issues, as this is seldom an attribute that can be assessed in standalone nonintervention studies. Observational longitudinal studies can be appropriate, but consideration needs to be given to response shift and other potential confounding factors, dependent upon the length of the study.
Other Considerations	Response shift Implementing this standard in predominantly non interventional studies where change within a relatively short period cannot be anticipated.

Name of standard	Modification of an existing patient reported outcome measure (PROM)
Description of standard	Modification of an existing PROM may alter a patient's subsequent response (FDA, 2009) and resultant score on a PROM (Patrick et al. 2007), and may jeopardize content validity (Rothman et al. 2009). This minimum standard sets out the requirements for re-validation to demonstrate that any modifications made to an existing PROM will result in a modified PROM that retains robust content validity and psychometric properties. Modifications that involve changing any content of the PROM, such as instructions, item wording, response options, or recall period, require cognitive debrief interviews with patients to evaluate the comprehension and appropriateness of the changes. All modifications, excluding those to instructions that do not impact the recall period or specific concept being measured, also require additional assessment of the modified PROMs psychometric properties (such as, reliability, validity and ability to detect change). Modifications which include the addition or removal of items/domains/concepts require qualitative evidence for the modification based on concept elicitation interviews with patients, as well as the cognitive debrief and psychometric validation evidence mentioned above. *This standard assumes that the original PROM being modified satisfied the minimum standards detailed elsewhere. If this is not the case the qualitative / quantitative evidence that is lacking can be addressed in the modification process, conducting additional concept elicitation interviews where necessary.* *This standard does not address the creation of short forms of existing PROMs, this issue is addressed in the 'other considerations' section of the report. Similarly, this standard does not include modifications related to mode of administration (e.g. ePRO or IVRS); this is addressed in a separate specific minimum standard. Finally this standard does not address translation / linguistic validation; the standards associated with translation of PROMs are considered beyond the scope of this report.*
Current Practice and Examples	There is a lack of published evidence to enable a statement on current practice for modification of existing PROM. From the literature that was available most modifications are tested purely through statistical analyses. However, some authors have reported using mixed methods, using both quantitative and qualitative methodologies. *Examples:* The European Organization for Research and Treatment for

	Cancer (EORTC) modified the 36-item Quality of Life Questionnaire (QLQ) by removing six items to produce QLQ C-30. However, research demonstrated poor internal consistency (Osoba et al. 1994). Following these psychometric results, two of the items were re-worded and the response format was modified to a four category rather than a dichotomous response system. These changes facilitated an increase in internal consistency to an acceptable level (Osoba et al. 1997). However, the authors do not report the use of any of the qualitative methods addressed in this current minimum standard. Keller et al. 1996 reported on modifying the recall period of the SF-36 from four weeks to one week. They hypothesized that the shorter recall period would be more sensitive than the original longer recall period to recent changes in health status in a sample of patients with asthma. Again, using a purely statistical approach the authors report a randomized trial where one group of participants completed the original SF-36 and the other group completed the 'acute' shorter-recall SF-36. The authors concluded that the results provide evidence for the use of the shorter-recall period to be used in further research. No qualitative methods were employed in this study. The EQ-5D has been modified using a qualitative methodology similar to the proposed minimum standards. In a two phase study the authors conducted face-to-face interviews with lay people to determine new labels for each of the five dimensions in the EQ-5D. The second phase of the study then utilized cognitive debrief strategies (ease of use, comprehension, interpretation and acceptability) in focus groups with both healthy individuals and patients (Herdman et al. 2011). The conclusion of this research was the EQ-5D-5L.
Published Guidance	The above minimum standard is in line with the following published guidelines and recommendations: FDA PRO Final Guidance (2009), Rothman et al., (2009), Patrick et al., (2007) and EMEA (2005). Although these documents are largely concerned with the use of PROMs in clinical trials they also provide guidance for modification of existing PROMs. Specifically, the FDA PRO Guidance suggests that qualitative work is undertaken when changes are made which affect the order of items, item wording, response options, deleting sections of PROMS, or when changing instructions to PROMs. Rothman et al. (2009) incorporate guidance from a number of sources including Snyder et al, (2007) and Burke et al, (2008) stating that it is acceptable to modify existing PROMs if additional qualitative research has been conducted demonstrating content validity of the modified PROM. Turner et al. 2007 also provide some guidance, although principally the development of PROMs can apply to modification of existing PROMS. They state that cognitive debrief is a "critical method" in PRO development. Furthermore, they argue that this

	qualitative method should be applied when modifying an original PRO which did not include substantial patient input in its original development. The current minimum standard reflects the guidance given in these documents.
Contribution to Patient Centeredness	While this minimum standard could apply to all research settings, it particularly contributes to Patient Centered research in its focus on patient involvement through qualitative research methods including concept elicitation and cognitive debriefing. This minimum standard ensures that any modifications made to an existing PROM are relevant to the patient population and are correctly interpreted and understood. Furthermore, involving patients when making decisions to delete or add concepts to a measure ensures that PROM captures relevant information and does not delete important information from the patient perspective.
Contribution to Scientific Rigor	Appropriate re-validation procedures set out in this minimum standard ensure that the content validity and psychometric properties of a PROM are not jeopardized by modification.
Contribution to Transparency	This minimum standard explicitly addresses recommendations for retaining content validity and psychometric properties following modification of an existing PROM. Implementation of this minimum standard would allow investigators to assess whether modifications made to an existing PROM are appropriate within a particular population and / or study design based on documentation of the evidence in support of the modifications.
Empirical evidence and theoretical basis	Theoretical: The main arguments for this minimum standard are: 1) Maintains content validity of modified PROM (Rothman et al. 2009). 2) Standardized procedures for modifying existing PROM. 3) Ensures the same level of scientific and theoretical rigor is applied to a modified PROM as a newly developed PROM.
Degree of Implementation Issues	The minimum standard outlined here is largely in line with current guidance and should therefore be able to be adopted without significant implementation issues. However, the evidence of this standard being used in current practice is lacking. This may be due to the time and financial resources required to carry out re-validation of modified instruments; however, the lack of a need / interest to publish is most likely. Therefore, in order to implement this minimum standard and avoid similar or the same modifications being conducted by

	different groups (e.g. changes to recall period or number of response options) a central repository to store information related to each PROM would be a hugely beneficial resource. Such a repository would require substantial resources. *Implementing this standard on an existing measure that lacks content validity would require greater resources for the research process to accommodate having to establish content validity for the whole PROM while also assessing the modifications.*
Other Considerations	Cross reference the following Minimum Standards: • Multi-mode equivalence • PROM Selection • Content Validity • Recall Period • Sampling • Validity • Reliability • Ability to detect change • Interpretation of PROMs

Name of standard	**Establishing multi-mode equivalence for patient reported outcome measures (PROMs)**
Description of standard	This standard describes methodological issues related to the assessment of measurement equivalence between different formats for presenting PROMs. Different modes of administering PROMs are available which offer many advantages, but also present methodological challenges. The equivalence of any PROM on different modes should be documented before initiating a study. The following factors need to be taken into consideration: The usability of the PROM on any electronic mode should be formally assessed in line with the data collection procedures in the study protocol (i.e. completion at home for a set period of time; testing with appropriate sample etc.)? The transfer of any PROM from its original format to another mode of administration (e.g. pen and paper to an electronic device) requires patients involvement in cognitive debrief and usability interviews. The cognitive debrief interviews should be semi structured to assess whether the patients' understanding and interpretation, of the content of the PROM is equivalent across modes. Usability assessment is designed to determine whether the ePROM in the context of the protocol leads to usability problems. The need for more quantitative assessment of equivalence (such as equivalence testing and psychometric analyses) should be determined based on the recommendations of the ISPOR ePRO Task Force (Coons et al., 2009). A moderate change requires equivalence testing and a substantial change also requires psychometric assessment. Equivalence studies should be based on a randomized design which exposes people to two administrations of the PROM. The administration of the PROM should be separated by a distraction task, and if possible participants should not be informed that they will complete the survey twice. Equivalence should be statistically assessed using the intra-class correlation coefficient (ICC) to assess the level of agreement between data points (see Coons et al. 2009). *It would be preferable to have the second administration on a later day (unless the patient's health status has changed).* Modes of presentation / administration of PROMs should not be mixed (e.g. paper and web within the same study) without robust evidence that the different modes have high levels of equivalence (this could be defined as an ICC that is at least as good as the test retest reliability ICC for the measure). Researchers should further control for mode of administration in such a study in the analyses of treatment effect (e.g. modality as a covariate).

	A more desirable standard would involve establishing the measurement equivalence across modes even when only minor changes have been made. This may be especially important if the PROM is to be used in samples who may have less experience of some electronic devices such as some older people, recent immigrants, children or any patients who indicate that they are not familiar with the use of computers etc. *Test retest reliability over short time periods is not always available and so a more desirable standard of equivalence testing would include within mode arms which assess the degree of variation in data that would arise by chance (e.g. P&P on first and second administration). This is the benchmark then for assessing cross modal equivalence.* *New instruments should be simultaneously developed on multiple modes. Original concept elicitation interviews with patients (required to establish content validity) could then include discussions of the most appropriate / convenient modes of administration.*
Current Practice and Examples	The most important guidance to emerge in this area is that from the ISPOR ePRO Task Force (Coons et al. 2009). This describes standards for determining what level of equivalence testing is required and issues to consider in the design of such studies. From this guidance a minor change (i.e. no change in content or meaning) requires usability testing and cognitive debrief. Moderate change (such as switching to aural administration) also requires equivalence testing. Substantial change (such as changes to wording or response options) also requires psychometric revalidation. More recent guidance from an ISPOR Mixed Modes Task Force recommends that studies should not mix PROM data collection between different modes (Eremenco et al., 2011). In addition studies that rely on patient diaries should only use electronic diaries because of the potential problems with inaccurate paper diary completion (Stone et al., 2002). The EuroQol Group has developed their own guidance regarding methods for establishing the measurement equivalence of different forms of EQ-5D (Lloyd, 2009; Swinburn & Lloyd, 2010). This explores issues such as changing the physical size and mode of completion of visual analogue scales (VAS). It advocates making changes to instructions or formats where evidence suggests it may improve completion of the EQ-5D – as was actually done for the EQ-5D web version. The psychometric equivalence of this was established in a large study which demonstrated very high rates of identical responses and very high intra-class correlation coefficients on the VAS. The FDA's PRO Final Guidance, (2009) recommends that measurement equivalence of electronic formats should be established before trial data can be pooled. Gwaltney et al.

	(2008) report a body of evidence that the majority of studies that report the migration of PROMs to electronic formats indicate high levels of equivalence.
Published Guidance	This minimum standard is largely supported by the following guidance/recommendations: FDA PRO Final Guidance (2009); ISPOR ePRO Good Research Practices Task Force Report, Coons et al. (2009); ISPOR Mixed Methods Task Force, Eremenco et al. (2011); EuroQol Group Digital Task Force Guidance reports (Lloyd, 2009; Swinburn & Lloyd, 2010).
Contribution to Patient Centeredness	The collection of PROM data can be achieved using different formats (web, telephone, tablet, P&P) which may better suit the needs of patients. It may be possible to allow patients to choose their mode of administration, and switch modes throughout a study, if high levels of equivalence can be demonstrated. This could make participating in PCOR more convenient for patients. By providing patients with a choice of mode of administration, they may also feel more of an active participant in the research and that their role is more valued. All of these benefits for the patient may also lead to improved compliance and data quality. However the wrong choice regarding the adoption of a technology may disadvantage certain patient groups if they feel alienated by it which underlines the importance of thorough usability testing. *The views of study participants regarding modes of data collection could be usefully recorded and considered so that a better understanding of what methods are most suited to different types of people with different conditions.*
Contribution to Scientific Rigor	Implementation of this minimum standard will help to minimize or eliminate measurement error arising from the mode of presentation/administration of the PROM. Without meeting this standard there is a risk that switching to an ePROM format could lead to misinterpretation of study findings. In addition the use of ePROMs can provide additional useful information such as an exact date stamp for completion, potentially improved adherence to data collection, and easier transmission of data to central servers. More complex data collection patterns are also made easier. The electronic tool used for data collection could also support collection of other data (such as physician reports; medical history and safety data).
Contribution to Transparency	This minimum standard provides recommendations for determining equivalence between different modes of presenting PROMs. This minimum standard should allow investigators to assess whether changes to the mode are appropriate within a particular population and / or study design and whether the

	data support claims for equivalence. *A further assessment of the strengths/weaknesses of the PROM could be made based on documentation of the incorporation of patient feedback on the technologies and the appropriateness of the technology for specific diseases/conditions.*
Empirical evidence and theoretical basis	Gwaltney et al. (2008) have presented their meta-analysis of ePRO equivalence studies which informed and supported the development of the ePRO guidance from ISPOR. Much of the other subsequent work in this area has been substantially influenced by the recommendations in Coons et al. (2009). This field has its base in the field of psychometrics or measurement theory. 1. Different ePROM formats provide considerable flexibility and choice to researchers and study participants. They are each suited to different types of data collection. 2. The use of non-equivalent ePROM versions introduces measurement error, which will affect the interpretation of study data. 3. Assessing the usability of technologies in a study context is essential. 4. The ISPOR TF paper (Coons et al. 2009) provides a framework for determining the level of equivalence assessment that is required when PROMs are migrated to electronic formats.
Degree of Implementation Issues	Electronic data capture invokes technical problems and is governed by Federal regulations. The technical problems include data security, integrity, access and processing. However in reality these issues also exist for P&P data collection. The implementation of this minimum standard will require researchers to be able to review the robustness of any ePROM and the extent to which additional equivalence testing may be required. If further studies are required then this standard and the other guidance documents that are referenced provide details of how such studies should be designed and conducted.
Other Considerations	The choice of the mode should be formally assessed against other options and determined as preferable on the basis of suitability in this specific patient group, measurement equivalence, and cost. There may be a substantial difference in the cost of different alternatives (e.g. web applications versus iPads) but the resultant data quality may not differ so this should always be considered. Studies that use patient diaries where patient reported data are collected repeatedly over many days only electronic data collection should be permitted. Paper diaries have been shown to be subject to serious flaws in data reliability caused by

	participants back filling and forward filling the diary (Stone et al., 2006).

APPENDIX B: Other Considerations

Lessons from the HTA process

The Health Technology Assessment process includes a range of different methods to understand the real world value of a medical technology. Some countries (Australia, UK) establish the value of a medicine in terms of a metric which combines quality of life and length of life (the so called quality adjusted life year – or QALY). In cost effectiveness analysis the benefit of a treatment is expressed in QALYs. The QALY has some notable features which may be usefully considered by PCORI. Fundamentally the use of the QALY places patients' HRQL and values at the center of decision making. This contrasts with clinical trial arena where PROMs are usually supportive secondary endpoints and interpreted alongside the key primary endpoint. The use of the QALY has forced the HTA community to use single index (rather than profile) measures, and to consider ways of valuing health rather than just measuring it. This has potential application for PCORI as a way of addressing comparative effectiveness questions.

Understanding value: Data regarding health related quality of life (HRQL) reflect the value that people place on a health state. The HRQL measures that are used are not simply numerically scored. Instead they reflect they reflect strength of preference which in turn reflects value with respect to how much life you may be willing to give up to achieve the state. This preference based scoring is very different to psychometric instruments. It does though provide a potential mechanism for meeting PCORI's objectives of understanding outcomes that patients value.

A single metric: To estimate QALYs HRQL is expressed as a single value (ranging between full health and dead). See discussion on Profile measures.

Standardizing outcomes measurement: NICE has stated a preferred option of standardizing the measurement of HRQL through the use of one measure. This approach also allows for much easier comparisons between different trials and different treatments in an indication. For NICE this is a generic measure - but within a disease area there is no reason in principle why it couldn't be a disease specific measure. Standardizing outcomes measurement to single measures (within an indication) may be a very useful approach for PCORI to allow for easier comparison and meta-analysis.

Patient preferences: Formal methods for understanding patient preferences for interventions are quite often used in the HTA process as a supplement to the cost effectiveness analyses. Issues such as convenience, nuisance, dosing, mode of administration, bothersome but mild side effects will usually not impact upon QALYs but they may be important aspects of the value of treatment for patients. Methods include conjoint analysis where best practice guidelines are available (Bridges et al., 2011). The impact of these features on adherence has also been explored (Johnson et al. 2007). These methodologies may be useful for assisting with PCORI's aims of understanding what patients value.

Meta-analysis: This methodology for synthesizing literature on treatment effects is widely used in HTA, but there much less work on the meta-analysis of PRO outcomes has been reported. This would be a useful issue to explore to support the work of PCORI and comparative effectiveness.

Interpreting Profile Measures

The vast majority of PROs are profile measures. They provide summary scores for different dimensions – either in terms of functioning, symptoms or other concepts. This raises a problem for interpretation. How do we know if a treatment is beneficial for patients if two dimensions show improved scores and one shows a deterioration. If the question being addressed is of the form – should treatment A or treatment B be used to treat X – then this is a quite important issue. There are several aspects to this problem.

- If patients improve a little on two dimensions but get worse on another it is difficult to understand the net change in health.
- Different dimensions of a questionnaire may not all be equally important for patients. So if a patient with overactive bladder reported improvements in symptoms of urgency this may be more than offset by any increase in incontinence episodes. Few PRO measures include any attempt to provide relative weightings of different dimensions for decision making.
- The multiple domains of profile measures can also present problems with hypothesis testing because of the need to test α multiple times. One solution to this is through the use of hierarchical testing. However this approach may be less suited to PCOR because there is a need to understand the impact of a treatment on all relevant endpoints. Other ways of handling multiplicity may be more appropriate.

Profile measures do also have many benefits over single index measures. Profile measures allow us to understand the impact of an intervention on different aspects of functioning, symptoms and wellbeing separately. This means it is possible to disaggregate the effects of a drug and to perhaps more fully understand side effects versus effectiveness.

Lastly the relevance of this issue regarding the value of profile measures versus single indices is dependent on the question being addressed by the research. One form of question relevant for comparative effectiveness is highlighted above. However other questions may be framed as "What outcomes are most important for you as you make treatment decisions". The profile measure here has far greater relevance.

Response Shift

Individuals who participate in longitudinal studies may experience an adjustment in their internal standards (i.e. recalibration), values (i.e. reprioritization) or meaning (i.e. reconceptualization) concerning the target construct they are being asked to self-report (Schwartz & Sprangers, 1999). A clear example of these phenomena is presented in Schwartz (2010) which discusses the concept of participation as it relates to the capacity of individuals with disabilities to engage in meaningful and pleasurable activities. Individuals may develop a disability which is associated with a severity of fatigue that was previously unknown and consequently recalibrates what severe fatigue means to them as it relates to participation. This development significantly complicates the process of comparing predisability and postdisability assessments of fatigue. Individuals may also experience a reprioritization of life domains such as sense of community and interpersonal intimacy which become more salient to their sense of participation than perhaps career success or material gains. Lastly, individuals may reconceptualize participation to focus on domains where they continue to have control and be effective when assessing their participation. Such response shifts are to be expected with subjective rather than strictly objective assessment approaches.

Response shift presents a significant challenge to basic assumptions concerning the use of standardized questionnaires (e.g. measurement invariance) and psychometric properties such as reliability, validity and responsiveness. Statistical techniques for the detection of response shift are still evolving but include the use of structural equation modeling, latent trajectory analysis and regression tree analysis. Research has also examined the use of individualized measures to examine stability in domains and internal metrics over time and across groups (O'Boyle et al, 1992; Ring et al, 2005). Some questionnaires have incorporated items which attempt to directly identify response shift such as The Health Education Impact Questionnaire (HEI-Q) (Osborne et al, 2006) but this has been a rare consideration for those involved in PRO instrument development.

Substantial research is still required in examination of the response shift phenomena. Barclay-Goddard (2009) highlights a number of important questions that remain unanswered; is response shift clinically meaningful when identified? What factors may predict response shift magnitude and direction? What is the "best" way to measure response shift? At present there is a lack of consensus on even the terminology and theoretical models pertaining to response shift and this is of particular concern when seeking to understand how best to address such an issue within longitudinal studies. While response shift may be a concern for any research setting assessing PRO concepts over time, much focus is currently on statistical methods that can be employed to address the issue. PCORI, on the other hand, could provide an opportunity to gather qualitative and quantitative information that may help to advance understanding of this phenomena, and further investigate how best to communicate any findings and associated implications to patients and clinicians utilizing PROM information in their decision making.

Developing Short Forms of Existing PROMs

The development of short form versions of existing instruments has previously been undertaken as part of efforts to increase their acceptability to patients and reduce administration burden. A number of well-recognized instruments have short form versions that are widely used in both clinical and trial contexts.

A variety of psychometric methods have been used previously in the development of abbreviated instruments; The SF-36 was shortened using regression techniques to select and score the 12 items that make up the SF-12 (Ware et al., 1996). Alpha reliabilities were used to extract items from the WHOQOL-100 to create the WHOQOL-BREF (Skevington et al., 2004); and similarly the short and very short forms of the Children's Behavior Questionnaire used alpha reliabilities combined with factor analysis to determine which items to include from the original measure (Putman and Rothbart, 2006).

Criticism has been leveled at the use of some traditional techniques in the development of short form instruments. There is little evidence to suggest that these techniques address which items on a measure are most relevant to the underlying construct and it is becoming increasingly common to use more advanced psychometric methods such as item response theory (IRT) and Rasch modeling in the instrument development process. The IMMPACT guidance recommends that IRT be used to evaluate individual items and the equivalence of using the revised instrument in different settings or with different populations (Turk et al., 2006). Members of the SF-36 group have also explicitly advocated the use of such advanced psychometric methods to measure health outcomes (Bjorner et al., 1998). In addition, IRT is now increasingly finding use in the development of computer adaptive testing (CAT) approaches which seek to minimize patient burden through selective administration of test items. Consideration of CAT approaches could be of significant benefit to PCOR when examining potential data collection strategies.

At present there is little agreement on what constitutes a minimum standard for the development of short form instruments. The recent availability of statistical packages which permit the use of more advanced techniques for assessing item functioning is slowly revolutionizing instrument development efforts. As more research evidence becomes available and high profile projects such as PROMIS report on their progress, it is likely that increased adoption of such techniques will occur. PCOR could benefit from maintaining an awareness of developments in this field and perhaps helping to shape the research agenda.

Proxy and Caregiver Measures

The standards presented in this report focus on patient self-reported outcome measures. However, some patients are unable to self-report (e.g. cognitively impaired patients or infants), and as such some consideration of proxy measures is required. When developing and selecting proxy measures consideration needs to be given to the difficulties for a proxy reporter to know how the patient is feeling and to avoid having their own opinions influence their reporting of the patient's health. Therefore, proxy measures should be limited to observable events or behaviors (FDA, 2009; ISOQOL, 2009). The development and selection of proxy measures for PCOR may require some unique minimum standards, and this will need to be addressed in PCORI's future work.

For patients who receive informal care, the impact of their disease and / or treatment on their carer may have an impact on their decision making. The impact of a patient's disease or treatment on a caregiver's health-related quality of life is recognized by institutions such as the National Institute for Health and Clinical Excellence (NICE, 2008) in their health technology appraisals. The caregiver voice could be considered in PCOR through the use of caregiver outcome measures. If these measures are used in PCOR, the same qualitative and quantitative standards detailed in this report should be applied to their development. However, further research will be required to establish how the caregiver voice should be incorporated in PCOR as there are a number of unresolved issues surrounding the use of such data (e.g. cultural differences, how do you define a caregiver?, how many caregiver perspectives should be considered?).

Patient Involvement Beyond the Development of a PROM

In light of the shift towards patient-centered outcomes, it is crucial that the evidence base surrounding outcomes research adequately engages and reflects the views of patients. This can be achieved in both retrospective and prospective research:

- There may be occasions where the study design does not employ the use of PRO instruments (e.g. retrospective review of medical charts, medical databases, and claims data in comparative effectiveness research (CER)). In such occasions there are a number of ways which investigators can ensure the patient voice is not lost. Investigators may consider an exploratory stage to review a sample of original CER data for reference to patients' subjective experience. For example, investigators may consider reviewing patient notes compiled by nursing/medical staff to extrapolate any incidence where salient patient reported items have been recorded. However investigators need to first consider the appropriate level of ethical review before reviewing any material beyond the original purpose for which it was collected. Alternatively, patients could be involved in the decision making process around which clinical outcomes or available clinical data should be included in CER, based on their experience of a disease or treatment.
- When PROM data are available in retrospective analysis patients could equally be involved in selecting the outcomes most relevant to their experience of disease or treatment for inclusion in CER.
- Similarly in prospective studies, investigators should consider including lay representatives (patient or care-giver) in study steering groups or consult patient associations for each stage of this type of research: from guiding research questions to reviewing final reports. Inclusion of patients and/or carers would ensure patient perspective is identified and addressed. Furthermore, adopting this approach of patient or care-giver inclusion will enhance validity and relevance of research. It is

helpful if the role of any lay representative is clearly defined in an agreed document from the outset.

Communication of PRO Research to Patients

In the past, healthcare decision making relied mostly on the physician's clinical experience and data from medical tests. More recently, healthcare has been moving toward a "patient-centered model" that allows for the active participation of patients in their own care with guidance from healthcare professionals. This shift toward patient centeredness means a wider range of outcomes from the patient's perspective should be measured in order to better understand the benefits and risks of healthcare interventions (Stanton, 2002). Involving patients in the development of PROMs and incorporating concepts that are meaningful to them based on their specific experiences with their disease/treatment is very important and can help ensure better quality data. However, communication between patients and clinicians based on this data, and patients' ability to interpret the data resulting from PROMs should also be considered.

PROMs can help facilitate clinician-patient interactions and relationships because they can shed light on the needs and concerns of patients (Lohr and Zebrack, 2009). However, researchers must ensure that the PROM results are meaningful to clinicians and patients in order to contribute to their communication and decision making processes. Depending on the clinicians own comfort level with PROMs, it could be challenging for them to determine what their patients are able to understand and interpret. Incorporating the use of PROMs into clinical training could have a positive impact on clinicians understanding and attitudes towards the use of PROMs in research and decision making. However, in the meantime it is essential that research is conducted to ensure PROM data is meaningfully interpreted and clearly presented to maximize its use in clinical decision making and ensure the patients' voice is not lost outside of the research setting.

Some research into the presentation of PROMs results has been conducted (e.g. McNair et al., 2010); however, while this is a vital part of communication the meaningful interpretation of PROM data should be required before presentation is considered. As discussed in the interpretation of meaningful change standard, further research is required to involve patients as active participants of interpretation of PROMs for use in PCOR rather than being passive participants.

Ideally, easily interpretable and clearly presented PROM data relevant to diseases/treatments would be readily available to patients through internet searches or other easily accessible means in order to facilitate patients' decision making. This could lead to patients' improved self-efficacy and greater satisfaction with care.

Feasibility of Use in Low Literacy and/or Non-English Speaking Patients

In addition to focusing on concepts most meaningful to patients, in order to be patient centered, PROMs must maintain this meaning by being comprehensible across the patient population in which they are to be used (Basch et al, 2011). In real-world and non-experimental settings, more than many other research contexts, this is likely to include patients with low literacy levels due to lower levels of educational attainment or poor English language fluency where English is not the primary language spoken.

Low literacy is an inability to read or write well enough to perform necessary tasks in society or at work. The U.S. Department of Education's National Adult Literacy Survey (National Center for Education Statistics, 2002) categorized the US population into 5 levels of literacy with the lowest (level 1) broadly defined as less than fifth-grade reading and comprehension skills. Low literacy is an important factor in use of PROMs in real-world or non-experimental settings as completion of PROMs is traditionally reliant on patient comprehension of written words and approximately a fifth of American adults read and comprehend below a fifth-grade level, with the figure rising to approximately half the US population if level 2 literacy (reading and comprehending at 5^{th}-7^{th} grade levels) is included in the definition of low literacy (National Center for Education Statistics, 2002). Moreover, low literacy is associated with poor health status (Weiss et al, 1994) and therefore is an issue of even greater concern when conducting research in patient populations. Many PROMS are not suitable for use in populations that include patients with low literacy. This raises concerns over the validity of PROM data from patients with low levels of literacy, that the experiences of these patients are not adequately captured, and the potential for missing data due to patients with low literacy being unable to complete PROMs, with the risk that the patient voice may be lost for patients with low literacy.

Feasibility of use in non-English speaking patients is a related issue. According to the 2010 American Community Survey 4.6% of all US households have no one aged 14 and over who speaks English only or "very well", with the majority of these speaking Spanish or Asian and Pacific Island languages. Literacy in English is likely to be low in a significant proportion of individuals in the US for whom English is not their first language or the main language spoken in their homes. To some extent this can be addressed through use of appropriate alternative language versions. Many existing PROMs have alternative language versions available and there is available guidance for translation of PROMs (Wild et al, 2005). However, translation does not address the possibility of low levels of literacy in other languages.

There is no easy way to address the issue of low literacy in PROM selection, development or use. Patients with lower levels of educational attainment and of various racial/cultural/linguistic backgrounds should be included in cognitive debriefing samples to explore the ease with which these patients complete the PROM and identify specific areas of misunderstanding or difficulty. Indeed, involvement of patients with lower levels of educational attainment was required for Patient-Reported Outcomes Measurement Information System (PROMIS) in qualitative item development work (DeWalt et al, 2007) and was also a requirement in development of the PRO version of the Common Terminology Criteria for Adverse Events (PRO CTCAE, Hay et al, 2010). However, this raises its own challenges as these patients may not be among those most likely to volunteer to participate in cognitive debriefing interviews, particularly if recruitment strategies and materials rely on the written word. In addition, areas of misunderstanding and difficulties identified by cognitive debriefing with low literacy patients may not easily be addressed through traditional instrument development options of rewording instructions, items and response options.

Some modes of PROM administration offer the potential to be less reliant on patient comprehension of written words but these are not without limitations. For example, using a telephone based interactive voice response (IVR) platform removes the requirement to read items and response options; however the types of questions and response options that are suitable for IVR data capture are limited. Study protocols could also be adapted to allow PROMs to be read to patients by site staff to enable those this low literacy to respond verbally but this may suffer from similar limitations and has resource implications. ePRO and web-platforms offer the potential for more dynamic presentation of items and response options including interactive or animated images and audio-tracks that might replace or accompany text to aid comprehension in low literacy patients. However, the extent to which these adaptations succeed in achieving higher levels of comprehension or PROM completion rates among patients with low literacy is unclear. Additionally, adaptations like these would also constitute a significant level of modification of most existing PROMs requiring revalidation of measurement properties or evaluation of equivalence (Coons et al, 2009).

Oxford Outcomes

An ICON plc Company