Author's response to reviews

Title: Measuring patient reported outcomes: Moving beyond misplaced common sense to hard science

Authors:

Stephen P McKenna (smckenna@galen-research.com)
Lynda C Doward (ldoward@galen-research.com)

Version: 2 Date: 22 November 2010

Author's response to reviews: see over
Measuring patient reported outcomes: Moving beyond misplaced common sense to hard science

Stephen P McKenna and Lynda C Doward

Abstract

Interest in the patient’s views of their illness and treatment has increased dramatically. However, our ability to measure such issues appropriately lags far behind the level of interest and need. Too often such measurement is considered to be a simple and trivial activity that just requires the application of common sense. However, good quality measurement of patient reported outcomes is a complex activity requiring considerable expertise and experience. This review considers the most important issues related to such measurement and details how instruments should be developed, validated and adapted for use in additional languages.

Commentary

Questionnaires are ubiquitous throughout life these days. Medicine is no different with the patient rightly seen as a client whose views are crucial to gaining a clear understanding of anything from the quality of service provision to treatment effectiveness. Patients are increasingly regarded as one of the key stakeholder groups in medicine that, alongside regulators, payers and clinicians, can influence access to, and reimbursement for, pharmaceutical products. Much of the information on patient views is collected via questionnaires. Many, if not most, of these are hastily prepared by clinical or other professionals wishing to answer specific questions that they consider to be important. Unfortunately, the development and application of such questionnaires is often regarded as a matter of ‘common sense’ requiring little scientific consideration. However, in this area of research, common sense is commonly non-sense! As a result many of the questionnaires patients are asked to complete in clinical practice and trials are of poor quality and collect information that is of scant relevance to the patient and ultimately of little value.

Questionnaires used to elicit information from patients in the medical arena are referred to as patient reported outcome (PRO) measures. PROs are far more than a mechanism for gathering opinion. They are designed to measure a specific concept (construct) in a standardised way. Thus, they provide a means of quantifying qualitative information. In reality there is a great deal of science involved in producing good quality PRO measures. Indeed, the PRO development process requires careful consideration of several key issues as set out in Figure 1.

When selecting a PRO it is important that evidence is available to show that each of these key issues has been considered and addressed during instrument development and testing.

Where measures are required for use in different languages or cultures there are added considerations;
The purpose of this article is to highlight current standards in the development and testing of PRO measures in order to guide practitioners and researchers in their selection and use.

**What do PROs measure?**

PRO is an umbrella term that covers a range of different types of outcome (see Table 1). PRO scales should not be confused with clinical-rating scales, where a clinician completes a form to rate disease severity or treatment effects. The common link between PROs is that they collect information directly from the patient without interpretation by clinicians or others [1-3]. However, this does not imply that all PROs measure issues that are of concern or importance to the patient.

Measures of symptoms, activity limitations, health status, health-related quality of life (HRQL) and quality of life (QoL) completed by patients are all examples of PROs [1,4]. More recently, PROs have also been used in clinical trials to address issues of patient satisfaction, compliance and treatment preferences. These outcomes each represent distinct measurement constructs that should not be confused. Indeed, the term PRO was coined in 2000 specifically to avoid misuse of and confusion surrounding the term ‘quality of life’. It had been (and occasionally still is) common practice for instrument developers to refer to any scale as a measure of QoL even where it was clearly designed to address a different construct [5].

To summarise, PROs that assess symptoms (impairment) or activity limitations (disability or functional limitations) address issues of primary interest to the clinician. HRQL measures are made up of scales that assess symptoms and activity limitations. In contrast, QoL scales determine outcomes that are of primary concern to the patient.

The most widely applied QoL model is concerned with the extent to which disease and its treatment prevents an individual from meeting his or her needs. Individuals are driven or motivated by their needs and the fulfilment of these provides satisfaction and a good QoL [6]. Consequently, QoL is good when most needs are fulfilled and poor when few needs are satisfied. The QoL issue is whether the individual is able to fulfil his or her needs. Functioning is only important insofar as it permits need fulfilment. For example, employment has the objective of earning a salary but it also leads to the fulfilment of a number of basic human needs (see Figure 2). Satisfaction of these needs leads to a good QoL.

Measures of satisfaction differ from HRQL and QoL as they are concerned with the process of treatment rather than its outcome. The assessment of utility involves valuing preferences for specific outcomes. Most utility assessment employs generic measures such as the EQ-5D [7] or Health Utility Index [8] which are in fact measures of HRQL. Recently, utility valuations have been derived from responses to disease-specific QoL instruments, providing more accurate measurement of this outcome [9-11].
Generic versus disease-specific PROs

Irrespective of the construct assessed a PRO scale may be generic or disease-specific. As its name implies, a generic instrument is intended to be used in any disease population. Some of the more widely known HRQL instruments are generic. Examples include the Sickness Impact Profile (SIP; [12]), the Nottingham Health Profile (NHP; [13]), the Short Form 36 (SF-36; [14]) and the EuroQoL (EQ-5D; [7]). Such instruments usually assess several domains and provide a profile of scores.

Traditionally, generic instruments were used to provide comparisons between diseases or to compare data to population normative values. However, it is now generally accepted that such comparisons are scientifically flawed and invalid, as questionnaire items work in different ways with different patient groups [15]. This means that as generic measures cannot allow comparisons to be made between the impacts of different diseases or between healthy and diseased populations they no longer have a clear role in measuring health outcomes.

A second major problem with generic instruments is that they are not designed to capture areas of concern to specific patient populations. This raises two issues. First, they are likely to include items that are irrelevant for certain patient groups. For example, questions that address physical functioning or bodily pain will only be relevant if they are a feature of the disease under study. Asking patients to answer questions that are irrelevant is likely to alienate respondents and increase the potential for missing or inaccurate responses. Secondly, they are likely to miss issues that are a specific feature of the disease under study. As a result generic scales lack the responsiveness needed to measure change associated with effective treatment.

As a result of the acknowledgement that generic measures are of little value in the measurement of health, generic instruments are no longer developed. The most widely used generic measures (such as the SIP, SF-36 and NHP) were developed in the early 1970s. The five items included in the EQ-5D were taken from existing generic measures and so are of the same vintage. The way in which patients conceptualise their problems and the language with which they express themselves can change in a generation. Moreover, certain issues may become less important with time. For example, lack of mobility may be compensated for by advances in technology. Furthermore, the generic health status instruments have not benefited from improvements in test construction methodology and scaling techniques. Consequently, the reliability and responsiveness of the generic measures fall far short of that required for instruments included in clinical trials.

Disease-specific questionnaires are developed to address those aspects of outcome that are important for a particular patient population. Thus, patients are only asked questions that are relevant, meaningful and acceptable to them. Addressing the relevant areas of concern for the group under study maximizes respondent acceptability and minimizes missing data. Consequently, disease-specific instruments possess greater potential for showing differences between competing therapies. A criticism that is often levied at the use of disease-specific scales is the lack of comparability across diseases. This is a particular issue for reimbursement authorities who have a requirement to assess the comparative benefits of treatment reimbursement across disease areas. However, as has been noted above the use of generic scales does not provide a valid basis for comparison across diseases. Recent advances in scaling theory are being applied to
address this issue. It is now feasible to use disease specific measures to allow across-disease comparisons, providing the instruments are based on the same theoretical model of the construct measured.

Use of PROs within the medical sector

Figure 3 shows the different types of PRO outcomes used in medicine. The diagram reflects the fact that most PROs currently used assess HRQL. The assessment of QoL is relatively rare, despite the term being widely used in research reports and publications.

The widespread use of HRQL measures gives some cause for concern. First, the term itself is misleading and unhelpful insofar as it implies that QoL is being measured. Bradley [16] argues that ‘clinicians may be misled into thinking that findings based on a [HRQL] instrument indicate that treatments do not damage QoL when all the data reveal is that treatments do not damage perceived health’. Indeed the focus on HRQL provides a framework for assessing interventions predominantly from a clinical rather than patient perspective. Secondly, HRQL scales do not necessarily address issues of primary concern to the patient. The focus of HRQL on the patient’s ability to fulfil roles deemed ‘normal’ takes no account of the fact that patients with chronic disease adapt to their condition; often replacing activities that they can no longer perform with others that are equally satisfying. Patients may give up functions that become problematic and take up other leisure activities in order to maintain their QoL. For example, while muscular degenerative disease patients may experience ambulatory problems, they can still remain independent and thus maintain a reasonable level of QoL through the use of a walking frame or wheelchair. HRQL measures are unable to cope with such adaptation making it difficult for severely ill or disabled patients to show improvement, even following effective interventions.

QoL is the primary outcome of relevance and importance to patients. This is particularly true for patients with chronic disease where therapies cannot promise a cure or an extension to life. QoL is not intended to be an aid to diagnosis or a guide to the most appropriate intervention for a specific patient. However, its careful assessment should be able to determine which of alternative interventions is superior for patients as a group within the context of a clinical trial.

The needs-based approach to QoL assessment has a number of advantages for measurement of the impact of disease and its treatment. Rather than asking directly about a function it is possible to enquire about the needs that could be satisfied by that function. For example, questions about sexual performance are frequently left unanswered in questionnaires due to irrelevance or unacceptability. The needs approach allows questions to be asked about needs related to sexual functioning that can also be satisfied in other ways such as love, intimacy, touch and sharing with another person. The needs-based also copes well with patient adaptation. A chronically ill person can maintain a reasonable level of QoL by remaining independent through the use of aids and / or assistance. Patients who have activity limitations can still be shown to have a good QoL, as the concern here is the degree to which they can meet their needs, irrespective of how this is achieved.
Measures developed using the needs approach are disease-specific (or could be more appropriately described as disease-relevant). This allows them to focus on the specific needs interfered with by the disease and hence makes them highly relevant and acceptable to the patient. As specific needs may be affected by different illnesses it is possible to develop valid methods of making comparisons between the impacts of different diseases through the application of Rasch analysis (see below).

A further advantage of the needs-based measures is that they assess the single construct of need satisfaction allowing the construction of unidimensional scales or indices of QoL. A major problem of HRQL measures is that they collect information on a range of different types of outcome. Consequently they provide a profile of scores – see for example the Nottingham Health Profile (NHP; [13]), Sickness Impact Profile (SIP; [12]) and SF-36 [14]. It is not possible to compare scores on the different sections of the profile and it is certainly unacceptable simply to add together responses to the different sections to give a single score – although this is common practice in outcome measurement.

Selecting and using PRO measures for clinical trials and studies

The inclusion of poorly designed or inadequately targeted instruments in a clinical trial or study can have serious consequences. Furthermore, ethical questions are raised by asking patients to complete measures that are incapable of demonstrating treatment effects. It is strongly recommended that expert help is sought in selecting PROs. Too often the choice is based on issues that are helpful rather than of scientific importance. PROs may be selected because they are commonly used, used by a competitor or they are available in a wide range of languages. While such factors can be helpful they are minor compared with what the questionnaire actually measures and how well it does this.

When selecting a PRO measure it is first necessary to determine the construct(s) that have to be assessed in order to meet the objectives of the study. Having done this the next stage is to find PROs that measure these constructs well. It is not advisable to rely on, or to be limited to, the PROs listed in databases such as OLGA [17] or PROQOLID [18]. Such sources of information are often selective and / or omit important measures. Furthermore, they rely on test authors to provide information on the quality of the measures listed without providing any commentary on the acceptability of testing methods used or the appropriateness of any conclusions drawn. A thorough search of the medical literature should be made in order to find available measures and evaluate their suitability for use in the trial. This will often generate a host of potential PROs that will vary considerably in terms of the care with which they were developed and their psychometric quality.

Selecting the most appropriate questionnaire requires consideration of several key quality standards. These cover the development processes, instrument scaling, psychometric properties and cultural translation and adaptation processes (Figure 4; [4,19]). These standards are described in detail in the Appendix.
It is increasingly common for trials to include PROs. In almost all cases these are secondary endpoints although the European Medicines Agency (EMA) [3] and Food and Drink Administration (FDA) [20] state that such scales could be primary endpoints. Both organisations emphasise that the outcome measures selected for a study should be well targeted to the specific patient population - which fundamentally rules out the use of generic PROs. It is noticeable that both bodies now consider the most widely used PRO, the SF-36, to be unsuitable for making claims about the value of treatments. Indeed, this measure has such poor psychometric properties that it has never proved to be a valuable instrument for showing differences between treatments (see for example, [21,22]).

Where the instrument is used as a clinical endpoint in a trial and / or is intended to be used to support a product label claim or to provide information for inclusion in the Summary of Product Characteristics it is necessary to agree in advance with the appropriate authorities that data collected with the measure will be acceptable to them. This generally involves providing a detailed briefing book. The briefing book must include information on how each item was generated and the reasons for rejecting items. Evidence is required for the whole testing procedure and the development and validation of all language versions of the measure to be used in the trial. Problems will occur with older measures where such information is unlikely to be available and / or the development methodology was inadequate. Where a new measure is being developed for a specific trial it is prudent to keep the authority informed at each stage of instrument development. The EMA now has a Biomarker qualification system that allows PRO instruments to be evaluated. Once qualification has been achieved the EMA will accept all data collected in a trial that uses the measure.

Sufficient time should be allowed to ensure that the required language versions of a measure have been developed and validated (see below). Very often poor quality translations are produced relying on simple forward-backward translation techniques rather than an approach that involves relevant patients. Adapting measures appropriately is a time consuming procedure that needs to be built into trial planning.

Once an instrument has been selected it is crucial that its value and reasons for its use are clear to everyone involved in the trial. If this is not done data collection with the measure will be of debatable value. Staff involved in the trial at each centre will require training on the application of the measure and how to deal with problems that might arise.

Expert help is also required in order to analyse PRO data collected in a trial and to interpret findings. Too often articles reporting the outcome of PRO assessment in clinical trials are of a poor quality with the authors lacking the background necessary to report the findings in an appropriate way. A common error is to compare scores obtained in the trial with those for healthy populations or other disease groups. The application of Rasch analysis clearly shows that such comparisons are generally meaningless and must be avoided.
Development and validation of PRO measures

Where a search fails to identify a high quality PRO scale it requires a new PRO questionnaire to be developed. There are four key stages in instrument development:

1. Identification of measurement model
2. Generation of questionnaire content
3. Content refinement and item reduction
4. Scaling and psychometric evaluation

1. Identification of measurement model

All PRO instruments should be based on a stated model or theory of the construct being measured. Measures of symptoms or functioning may well be based on the WHO ICIDH classification for impairments and activity limitation respectively. A measure of QoL is likely to be based on the needs-based model of QoL. The model employed should be reported in the instrument development publication.

2. Generation of questionnaire content

The development of PRO instruments is a highly skilled activity best undertaken by specialists in measurement and psychometrics. It is particularly important that their content is generated by researchers experienced in qualitative interviewing techniques. Content for all PRO scales should be derived from interviews with relevant experts or patients. Thus, if a measure of QoL specific to endometriosis is required the content will be derived from qualitative interviews conducted with women experiencing the problem. Such interviews are not intended to explore issues identified from the literature or clinical experts. Both the relevant concerns and the wording used in the items must be generated during these interviews. This is the most crucial and skilled stage of instrument development. If good quality items are not identified the resulting instrument will be poor.

The interviews, which may last several hours, should be audio-recorded and transcriptions produced. It is generally found that 30 to 35 interviews are sufficient in order to generate items for a scale. Additional interviews tend not to identify new issues of importance. Interviewees will generally raise specific functions that are problematic for them. The skill of the interviewer is to probe such responses carefully to understand how the patient's life is impaired by such restricted functioning. The needs-based model of QoL grew out of such probing in the development of the Quality of Life in Depression Scale (QLDS; [6]). Depressed patients who were unable to be employed reported problems with the structuring their days, with identity and status and reduced social interaction (see Figure 2).

Qualitative analysis of the transcripts allows the construction of a PRO outcome model for the disease. This will identify the issues / needs that are relevant for assessment in the disease studied. The analysis will also identify potential items for inclusion in the measure. Where possible it is preferable to keep the wording used by interviewees for the items - although minor changes may be necessary. Items are then best expressed as statements made by patients, such as; "I've lost interest in food" or "I feel dependent on other people." Having the items in this form leads to a response format of 'yes' / 'no' or
‘true’ / not true. This is a natural way of responding to items that should enquire into issues that are clear cut. The application of modern psychometric models (such as Rasch analysis) indicates that increasing the number of possible responses for an item does not increase the sensitivity of the scale. Instead the final set of items should represent different amounts of the construct measured in the same way as the marks on a ruler denote different lengths.

3. Content refinement and item reduction

Patient interviews will identify a large set of potential items. Content validity is assessed by comparing the issues covered by the items to the outcome model and other sources of information about the impact of the disease. The first stage of item reduction involves ensuring that items are clearly expressed. For example, they should only address one issue, avoid duplication, be potentially capable of change with effective treatment (for example, avoiding statements such as I worry that my illness will become worse) and that they will apply to all respondents. Items that are not relevant are poor as they lead to ambiguous responses.

The next stage is to test the draft questionnaire with a new set of relevant patients by means of cognitive debriefing interviews. These will explore interviewee’s ability to understand and complete the measure and ensure that items are considered relevant. In this way the face validity of the measure will be established. Changes to wording can still be made at this stage and items can be removed or added as a result of the interviews.

4. Scaling and psychometric evaluation

Formal testing of the questionnaire for dimensionality, reproducibility and construct validity is undertaken by means of a test-retest survey. In most European countries the survey can be conducted by post. While test-retest reliability (reproducibility) can be assessed with a sample of around 50, the need to determine the dimensionality of the scale means that a sample of 100 or more is preferable.

Traditionally classical measurement theory has been used to evaluate the quality of measures. Emphasis was placed on internal consistency, factor analysis, correlational analyses and reliability. In recent years quality instrument development has been based on Item Response Theory and, specifically, Rasch analysis. In order to achieve fundamental measurement certain properties are required. These are that;

- the numerical properties of order (one mark on the scale represents more or less of the construct than another),
- addition (points on scales may be added together), and
- specific objectivity (the calibration of the scale is independent of the persons used to calibrate it and vice versa) are met.

Where data fit the Rasch model these properties are confirmed and fundamental measurement follows. Rasch is an unidimensional model that has two main assertions; that an individual’s score on a scale will only be dependent on (a) the severity of the item and (b) the level of impairment (in the construct measured) of the respondent. For example, in the context of QoL assessment, individuals with poor QoL will be more likely to affirm items representing severe impairment of QoL [23-25].
The ability of a scale to provide fundamental measurement should be established prior to the more commonly reported psychometric attributes. Rasch analysis has particular value in guiding item reduction. Traditional methods of item reduction that rely on item-total correlations and/or indices of internal consistency can have unfortunate effects on the sensitivity of measures and their ability to provide valid scores for individuals who are mildly or severely impaired in terms of the construct assessed.

Items that do not fit the Rasch model assess a construct other than that being measured and should be removed from the scale. Others will be discarded where they show DIF - indicating that responses are influenced by some extraneous factor (such as gender or age). DIF indicates that the item has bias – for example, showing that men are more likely to affirm the item irrespective of the level of their impairment or QoL.

As the Rasch analysis provides a value for each item in the scale it identifies items that are redundant (have the same value) and, consequently, are also candidates for removal. Where two items have the same value only one contributes to the total score on the measure.

While fit to the Rasch model is evidence of the construct validity of the scale it is still important to show that the scale is sound in terms of the classical properties of reproducibility and construct validity.

**Adapting PRO measures**

If the required language versions are known from the outset instrument development should be conducted in parallel in these countries [26,27]. However, this information is rarely available and it is more common for subsequent adaptations to be required. Again it is necessary to allow sufficient time for such adaptations to be produced as the process is time consuming.

Translation procedures

Translating PROs is a complex task that cannot be undertaken lightly without the risk of producing poorly-performing questionnaires. It is commonly stated that forward-backward translation is the gold standard translation methodology [28]. However, there is no evidence to support this view – it is merely a statement of belief. When such translation work was first handed to translators the need was felt by the test developers to assess the quality of the new version by some sort of ‘scientific’ method. This led to the introduction of back-translation. However, such a methodology raises the hackles of translators and not only because it casts unwelcome doubts on their abilities. If the translation is good, then the back-translation may well look nothing like the source questionnaire. Consequently, little information of any value is obtained by conducting the backward translation, while a lot of misleading impressions can result. The answer must be to produce quality throughout the translation procedure, rather than checking it a posteriori.

Rather than relying on forward-backward translation a dual panel methodology has been developed and is now commonly employed [see text box]. A recent study has shown that the ‘dual panel’ methodology produces translations that are more acceptable to patients in the new country than the use of forward-backward translation [29].
It is important to remember that this is only the start of the adaptation process. The new translation should then be tested by means of face-to-face interviews with several relevant patients to ensure that the adapted version has face and content validity. Finally, the psychometric properties of the adapted questionnaire must be established with new patient samples. This requires a test-retest survey to be conducted for each new language version produced. Such retesting is rarely undertaken but is necessary in order to show that the new language version works in the same way as the original - evidence required by the FDA [20].

Conclusions

The development, administration, analysis and adaptation of PRO measures are highly skilled, specialist areas. Too often non-specialists are given the tasks of determining which outcomes should be included in clinical studies and trials and how these should be measured. Unfortunately this largely explains why few such studies provide useful data. Such a situation represents a waste of resources and the opportunity to show the benefits of expensive new treatments. A more professional approach to assessing patient reported outcomes is needed. Of particular concern is the paucity of QoL studies undertaken - given that high quality measures are available for specific to several diseases [see for example; 30].

PRO measures needed for use in clinical trials should be given the same consideration as clinical outcome measures. Furthermore, sufficient time and resources need to be devoted to ensuring that they are available in the languages required for the study. Where a suitable measure is not available even more time is required to develop the appropriate PRO(s). Too often very expensive clinical trials waste the opportunity to assess QoL or other PRO outcomes appropriately due to lack of planning or an unwillingness to pay for the development work. In reality the cost of such work is minimal in comparison to the overall cost of the trial.

The development of outcome measures is far from a commonsense procedure and is dependent on expertise and experience. The [text box] lists some of the issues covered in this review. Many if not most of the points listed are counter to the commonsense view on outcome measurement and instrument development.

Given the expressed desire of organizations such as the FDA and EMA to be made aware of the benefits of treatment from the consumer’s perspective and the need to convince payers of the added benefit of new treatments, it is to be hoped that more attention will be paid in future to the assessment of the effects of new interventions from the patients’ perspective.

The science of patient-reported outcome measurement is developing quickly. For too long out-dated generic HRQL measures such as the SF-36, NHP and EQ-5D have been relied on in clinical studies. It is now well understood that such measures are inadequate for showing change over time or the different impacts of alternative interventions. Greater emphasis is now placed on measurement models, disease-specific measurement and the application of Item Response Theory rather than Classical Test Theory. Well developed measures are now generally of a better quality and are more sensitive than clinical outcome measures, even those considered to be objective.
Where it is considered to be important to assess the views of patients it is essential that great care is given to selecting and applying the best measures available. Too often measures are selected because they are commonly used, they were used by competitors in previous trials or because the physician is familiar with the measure and believes that he/she can interpret scores. None of these are valid reasons for selecting outcome measures. It is to be hoped that greater attention will be paid to the quality of available measures, what they actually measure and to enlisting the help of experts when selecting and using measures.

Abbreviations

Differential item functioning: DIF; European Medicines Agency: EMA; Food and Drink Administration: FDA; Health-related quality of life: HRQL; Nottingham Health Profile: NHP; Patient reported outcome measures: PRO; Quality of life: QoL; Quality of Life in Depression Scale: QLDS; Sickness Impact Profile: SIP.

Authors' contributions

SPM drafted the document. LCD helped to draft the manuscript. Both authors read and approved the final manuscript.

Authors' information

SPM is Director of Research and LCD is Principal Researcher at Galen Research. Since 1988 both authors have worked in the field of instrument development and adaptation. During that time they have created over 30 disease-specific patient-reported outcome measures, many of which are considered to be the instrument of choice for clinical trials and for monitoring patients in clinical practice.
References


**Recommendations for the production of high quality adaptations**

The two panel method is recommended for producing translations. The following recommendations are made in order produce high quality translations:

- Recruit ‘translators’ who currently live in the target country and who have good English.
  - Hold meeting in the country for which the measure is required.
  - Five to seven people enable fruitful discussion.
  - Preferable to exclude professional translators.

- An instrument developer should attend this meeting to explain intent of the items and their specific meanings in the context of the questionnaire.

- Inform group of model underlying the questionnaire, how it was developed, its design and content and target audience.

- Inform them of the translation requirements (in particular accessibility and acceptability of wording).

- Group should work as team with a co-ordinator whose task is to check that none of the parameters are neglected (in particular, structural and metric aspects that could be overlooked).

- Allow adequate time for the meeting to explore all issues fully.

Once the translated version of the instrument is agreed have it assessed by a lay panel again working as a group:

- The co-ordinator involved in the first panel should work with this panel also to ensure that the original meaning of items and questionnaire structure is maintained.

- Results of this meeting should be used to make final decisions about the wording of the questionnaire.

The whole procedure should be reported in detail, in particular explaining translation choices and changes made following lay panel testing. This not only informs on the process undertaken but also constitutes a thorough final review.
### A new commonsense for patient-reported outcome assessment

Do not rely on instrument databases for PRO identification and selection

HRQL consists of symptoms and functions

HRQL is not the same as quality of life (QoL)

The needs-based model of QoL is the most widely employed in medical research

True QoL has rarely been measured in clinical studies and trials

The content of QoL measures must be derived from relevant patients

PRO scales must be simple to administer, complete and score

Simple 2-point response formats are preferable to multiple response formats

All PROs used in clinical trials should be disease-specific

Generic PROs do not allow the impact of different diseases on patients to be compared

Population norms for PROs are invalid

Think twice before selecting generic utility measures such as the EQ-5D that have limited psychometric quality

QoL is an unidimensional construct

Data collected with PRO scales must fit the Rasch model

High reliability (reproducibility) is crucial to the accuracy of a PRO scale

Forward-backward translation is a flawed methodology – creating unnecessary work

Think carefully before using PRO scales developed in the Western world in Asia and Africa

Evidence is required of the scalability, reproducibility and construct validity of all language versions of a PRO measure used in a clinical trial
Figure 1: Key considerations for PRO questionnaire development

Figure 2: Employment related needs

Figure 3: Current usage of PRO outcomes in medical research

Figure 4: Brief checklist for assessing the quality of PRO instruments
Table 1. Types of patient reported outcome measures

<table>
<thead>
<tr>
<th>Type of PRO</th>
<th>Constructs assessed</th>
<th>Examples of coverage / domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symptoms</td>
<td>Impairment</td>
<td>• Pain</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Fatigue</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Anxiety</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Depression</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Incontinence</td>
</tr>
<tr>
<td>Functioning</td>
<td>Disability</td>
<td>• Bathing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Dressing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Walking</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Ability to work</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Activities of daily living (e.g. personal care)</td>
</tr>
<tr>
<td>Health Status</td>
<td>Combination of impairment, disability and occasionally</td>
<td>• Symptoms and functions as above</td>
</tr>
<tr>
<td>(HRQL)</td>
<td>some QoL</td>
<td></td>
</tr>
<tr>
<td>Quality of life</td>
<td>QoL</td>
<td>• Needs-based QoL</td>
</tr>
<tr>
<td>Utility</td>
<td>Combination of impairment, disability or QoL</td>
<td>• Symptoms &amp; functions as above</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Activities of daily living (e.g. personal care)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Needs-based QoL</td>
</tr>
</tbody>
</table>