

# Chapter 17: Patient-reported outcomes

---

Authors: Donald L Patrick, Gordon H Guyatt and Catherine Acquadro on behalf of the Cochrane Patient Reported Outcomes Methods Group.

Copyright © 2008 The Cochrane Collaboration. Published by John Wiley & Sons, Ltd under “The Cochrane Book Series” Imprint.

This extract is made available solely for use in the authoring, editing or refereeing of Cochrane reviews, or for training in these processes by representatives of formal entities of The Cochrane Collaboration. Other than for the purposes just stated, no part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except under the terms of the Copyright, Designs and Patents Act 1988 or under the terms of a licence issued by the Copyright Licensing Agency Ltd, 90 Tottenham Court Road, London W1T 4LP, UK, without the permission in writing of the copyright holders.

Permission to translate part or all of this document must be obtained from the publishers.

This extract is from *Handbook* version 5.0.1. For guidance on how to cite it, see Section 17.9. The material is also published in Higgins JPT, Green S (editors), *Cochrane Handbook for Systematic Reviews of Interventions* (ISBN 978-0470057964) by John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, Telephone (+44) 1243 779777; Email (for orders and customer service enquiries): cs-books@wiley.co.uk. Visit their Home Page on [www.wiley.com](http://www.wiley.com).

## Key points

- Patient-reported outcomes (PROs) are reports coming directly from patients about how they feel or function in relation to a health condition and its therapy without interpretation by healthcare professionals or anyone else.
- PROs can relate to symptoms, signs, functional status, perceptions, or other aspects such as convenience and tolerability.
- Items reflecting the concepts included in a PRO questionnaire are elicited from the target population; patient involvement in questionnaire generation is essential for content validity.
- A glossary is provided on the PRO Methods Group web site ([www.cochrane-pro-mg.org](http://www.cochrane-pro-mg.org)) for finding definitions of terms unfamiliar to authors.
- PROs are not only important when more objective measures of disease outcome are not available but also to represent what is most important to patients about a condition and its treatment.
- PROs can be continuous or categorical. Techniques are available to pool both kinds of measures.
- Review authors may need to do background reading about PROs to ensure they understand those chosen for inclusion into trials, in particular their validity and ability to detect change.
- A checklist is provided in this chapter on issues relating to PROs that authors should consider before incorporating PROs into their reviews and ‘Summary of findings’ tables.
- If completed reviews fail to record PROs when they were chosen as important outcomes in the review protocol, then they should be highlighted in the review as a deficiency in the current research on efficacy of treatment.

## 17.1 What are patient-reported outcomes?

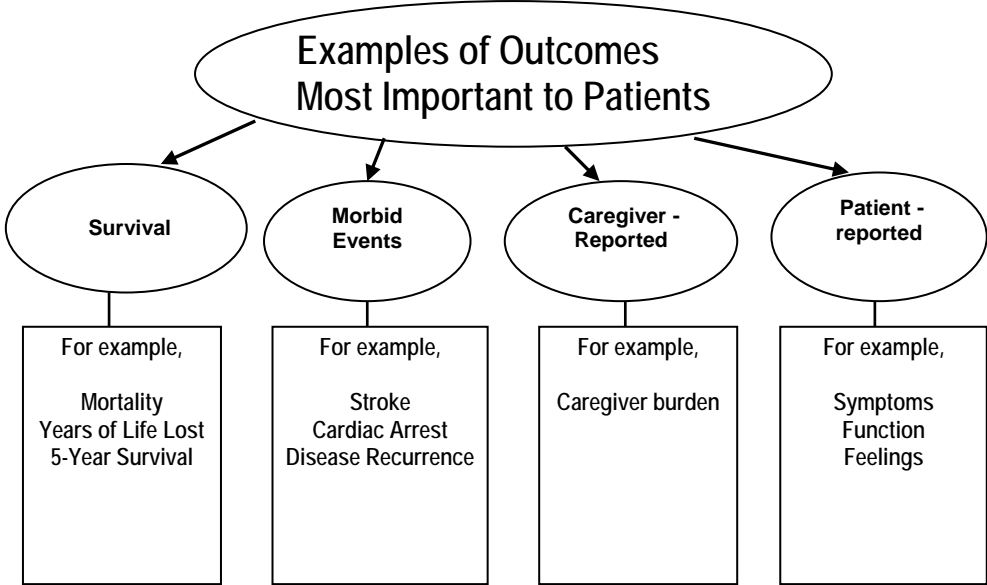
Patient-reported outcomes (PROs) are any reports coming directly from patients about how they function or feel in relation to a health condition and its therapy, without interpretation of the patient’s

responses by a clinician, or anyone else. PROs include any treatment or outcome evaluation obtained directly from patients through interviews, self-completed questionnaires, diaries or other data collection tools such as hand-held devices and web-based forms (US Food and Drug Administration 2006). Proxy reports from caregivers, health professionals, or parents and guardians (necessary in some conditions such as advanced cancer and cognitive impairment) cannot be considered PROs and should be considered as a separate category of outcomes.

PROs provide patients’ perspective on treatment benefit; directly measure treatment benefit beyond survival, disease, and physiologic markers; and are often the outcomes of greatest importance to patients. Reports from patients may include the signs and symptoms reported in diaries, the evaluation of sensations (most commonly classified as symptoms), reports of behaviours and abilities (most commonly classified as functional status), general perceptions or feelings of well-being, and other reports including satisfaction with treatment, general or health-related quality of life, and adherence to treatments. Reports may also include adverse or side effects (see Chapter 14).

PROs are sometimes used as primary outcomes in clinical trials, particularly when no surrogate measure of direct benefit is available to capture the patient’s well-being. More often, PROs complement primary outcomes such as survival, disease indicators, clinician ratings and physiologic or laboratory-based measures. Figure 17.1.a shows those outcomes that are considered most often as important to patients within a classification of all outcomes.

**Figure 17.1.a: Classification of clinical trial outcomes with illustration of those most important to patients**



PROs may be collected using a measure (or instrument) that is disease-specific, condition-specific or generic. Disease-specific measures describe severity, symptoms, or functional limitations specific to a particular disease state, condition or diagnostic grouping (e.g. arthritis or diabetes). Condition-specific measures describe patient symptoms or experiences related to a specific condition or problem (e.g. low-back pain) or related to particular interventions or treatments (e.g. knee-replacement or coronary artery bypass graft surgery). Generic measures are designed for use with any illness group or population sample.

A glossary on PROs is available from the Cochrane Patient Reported Outcomes Methods Group web site (see [Box 17.9.a](#)).

## 17.2 Patient-reported outcomes and Cochrane reviews

Systematic review authors will select PROs for inclusion depending on the scope and aims of their review. PROs are most important when externally observable patient-important outcomes are unavailable, or rare. For many conditions, including pain, functional disorders, sexual dysfunction and insomnia, no satisfactory biological measures are available. Conditions in which outcomes are known only to the patients themselves, such as pain intensity and emotions, demand PROs as primary outcomes. PROs are also important when observable outcomes are available, because they reflect directly what is important to patients.

An important early part of the systematic review process is to define and list all patient-important outcomes that are relevant to their question (Guyatt 2004) (see Chapter 5, Section 5.4.1). This step is highly germane to the measurement of PROs. Many primary studies fail to measure aspects of perceived health and quality of life that are very important to patients. When this is the case, evidence regarding impact of interventions on PROs may be much weaker than evidence regarding impact on disease indicators such as morbidity or mortality. In the extreme, there may be a line in a ‘Summary of findings’ table that is blank, that is, for instance, a line specifying health-related quality of life (HRQL) that is blank because no study addressed this issue directly. The careful prior consideration of all patient-important outcomes and inclusion as a blank row in a ‘Summary of findings’ table will highlight what is missing in outcome measurement in the eligible randomized trials and other studies.

It is important that review authors understand the nature of the PROs used in the studies included in their review, and communicate this information to the reader. In clinical trials, investigators use many instruments to capture PROs, and methods for developing, validating, and analysing PRO data are diverse.

## 17.3 Health status and quality of life as PRO outcomes

Health status and quality of life outcomes are an important category of PROs. Published papers often use the terms ‘quality of life’ (QOL), ‘health status’, ‘functional status’, ‘health-related quality of life’ (HRQOL) and ‘well-being’ loosely and interchangeably, despite clear definitions of terms (see [Table 17.3.a](#)).

Different types of instruments are available for measuring health status and quality of life (see [Table 17.3.b](#)). These may yield an overall score or **indicator number** (representing impact of the intervention on physical or emotional function, for instance), an **index number** (again an overall score, but weighted in terms of anchors of death and full health), a **profile** (individual scores of dimensions or domains), or a **battery** of tests (multiple outcome assessing different concepts): see [Table 17.3.b](#).

HRQOL can be measured using generic or specific instruments, or a combination of both. If investigators were interested in going beyond the specific illness and possibly making comparisons between the impact of treatments on HRQOL across diseases or conditions, they may have chosen generic HRQOL measures that cover all relevant areas of HRQOL (including, for example, mobility, self-care, and physical, emotional, and social function), and are designed for administration to people with any kind of underlying health problems (or no problem at all). These instruments are sometimes called health profiles; the most commonly used health profiles are short forms of the instruments used in the Medical Outcomes Study (Tarlov 1989, Ware 1995). Alternatively (or in addition) randomized

trials and other studies may have relied on instruments that are specific to function (e.g. sleep or sexual function), a problem (e.g. pain), or a disease (e.g. heart failure, asthma, or irritable bowel syndrome).

Elicitation of concepts and items for a PRO questionnaire should come from qualitative research with patients, family members, clinical experts, and the literature. For a guide to using qualitative methods, see Chapter 20. Involvement of patients in PRO questionnaire development is essential to ensure content validity. The concepts that are included and measured in an included study can only be determined by examining the actual content of items or questions included in an instrument claiming to measure quality of life or health-related quality of life. The *concept* is the ‘thing’ being measured. Concepts may relate to an individual item or to a subset of items that refer to the same concept, often referred to as domains. For example, an item measuring pain, a sensation known only to the patient, would be a symptom and the symptom concept that is being measured can be labelled as pain. An item assessing difficulty walking up stairs would be a concept related to physical functioning and might be labelled walking up stairs or as part of physical function. The labelling of concepts varies widely among researchers and there is no agreed-upon classification of concepts. Nonetheless, each item, subdomain, domain, or overall score addresses one or more concepts, which authors can identify from the content, e.g. language, used in the label for an item, domain, or overall score.

Review authors may gain considerable insight from what the authors of the original PRO development studies write about the nature or sources of items chosen for inclusion in a specific instrument. Unfortunately review authors will often find themselves reading between the lines of published clinical trial results to try and get a precise notion of the concepts or constructs under consideration. They may, to gain a full understanding, have to make at least a brief foray into the articles that describe the development and prior use of the PRO instruments included in the primary studies.

For example, authors of a Cochrane review of cognitive behavioural therapy (CBT) for tinnitus included quality of life as an outcome (Martinez-Devesa 2007). Quality of life was assessed in four trials using the Tinnitus Handicap Questionnaire, in one trial the Tinnitus Questionnaire, and in one trial the Tinnitus Reaction Questionnaire. The original sources are cited in the review. Citations to articles on the psychometric properties are also available in MEDLINE for all three instruments and could easily be identified with a search using the Google search engine. Information on the items and the concepts measured are contained in these articles, and review authors were able to compare the content of the instruments.

Another issue to consider in understanding what is being measured is how the PRO instruments are weighted. Many specific instruments weight items equally when producing an overall score. Utility instruments designed primarily for economic analysis put great stress on item weighting, attempting to present HRQOL as a continuum anchored between death and full health. Readers interested of the issues we have laid out in the previous paragraph can look to an old, but still useful summary (Guyatt 1993).

**Table 17.3.a: Definitions of selected terms related to quality of life**

<b>Term</b>	<b>Definition</b>
Functional status	An individual's effective performance of or ability to perform those roles, tasks, or activities that are valued (e.g. going to work, playing sports, or maintaining the house).
Health-related quality of life (HRQOL)	Personal health status. HRQOL usually refers to aspects of our lives that are dominated or significantly influenced by

	our mental or physical well-being.
Quality of life (QOL)	An evaluation of all aspects of our lives, including, for example, where we live, how we live, and how we play. It encompasses such life factors as family circumstances, finances, housing and job satisfaction. (See also health-related quality of life).
Well-being	Subjective bodily and emotional states; how an individual feels; a state of mind distinct from functioning that pertains to behaviours and activities.

**Table 17.3.b: A taxonomy of health status and quality-of-life measures adapted from Patrick and Erickson (Patrick 1993).**

Measure	Strengths	Weaknesses
<b>Types of Scores Produced</b>		
Single indicator number.	Global evaluation; Useful for population.	May be difficult to interpret.
Single index number.	Represents net impact; Useful for cost-effectiveness.	Sometimes not possible to disaggregate contribution of domains to the overall score.
Profile of interrelated scores.	Single instrument; Contribution of domains to overall score possible.	Length may be a problem; May not have overall score.
Battery of independent scores.	Wide range of relevant outcomes possible.	Cannot relate different outcomes to common measurement scale; May need to adjust for multiple comparisons; May need to identify major outcome.
<b>Range of Populations and Concepts</b>		
Generic: applied across diseases, conditions, populations, and concepts.	Broadly applicable; Summarizes range of concepts; Detection of unanticipated effects possible.	May not be responsive to change; May not have focus of patient interest; Length may be a problem; Effects may be difficult to interpret.
Specific: applied to individuals, diseases, conditions, populations, or concepts/domains.	More acceptable to respondents; May be more responsive to change.	Cannot compare across conditions or populations; Cannot detect unanticipated effects.
<b>Weighting System</b>		

Utility: preference weights from patients, providers, or community.	Interval scale; Patient or consumer view incorporated.	May have difficulty obtaining weights; May not differ from equal weighting, which is easier to obtain.
Equal weighting: items weighted equally or from frequency or responses.	More familiar techniques; Appears easier to use.	May be influenced by prevalence; Cannot incorporate tradeoffs.

## 17.4 Issues in the measurement of patient-reported outcomes

### 17.4.1 Validity of instruments

Validity has to do with whether the instrument is measuring what it is intended to measure. Empirical evidence that PROs measure the domains of interest allows strong inferences regarding validity. To provide such evidence, investigators have borrowed validation strategies from psychologists who for many years have struggled with determining whether questionnaires assessing intelligence and attitudes really measure what is intended.

Validation strategies include:

- content-related: evidence that the items and domains of an instrument are appropriate and comprehensive relative to its intended measurement concept(s), population and use;
- construct-related: evidence that relationships among items, domains, and concepts conform to *a priori* hypotheses concerning logical relationships that should exist with other measures or characteristics of patients and patient groups; and
- criterion-related (for a PRO instrument used as diagnostic tool): the extent to which the scores of a PRO instrument are related to a criterion measure.

Establishing validity involves examining the logical relationships that should exist between assessment measures. For example, we would expect that patients with lower treadmill exercise capacity generally will have more shortness of breath in daily life than those with higher exercise capacity, and we would expect to see substantial correlations between a new measure of emotional function and existing emotional function questionnaires.

When we are interested in evaluating change over time, we examine correlations of change scores. For example, patients who deteriorate in their treadmill exercise capacity should, in general, show increases in dyspnoea, whereas those whose exercise capacity improves should experience less dyspnoea. Similarly, a new emotional function measure should show improvement in patients who improve on existing measures of emotional function. The technical term for this process is testing an instrument's construct validity.

Review authors should look for, and evaluate the evidence of, the validity of PROs used in their included studies. Unfortunately, reports of randomized trials and other studies using PROs seldom review evidence of the validity of the instruments they use, but review authors can gain some reassurance from statements (backed by citations) that the questionnaires have been validated previously.

A final concern about validity arises if the measurement instrument is used with a different population, or in a culturally and linguistically different environment, than the one in which it was developed (typically, use of a non-English version of an English-language questionnaire). Ideally, one would have evidence of validity in the population enrolled in the randomized trial. Ideally PRO measures should be re-validated in each study using whatever data are available for the validation, for instance, other endpoints measured. Authors should note, in evaluating evidence of validity, when the population assessed in the trial is different from that used in validation studies.

### **17.4.2 Ability of an instrument to measure change**

When we use instruments to evaluate treatment effects, they must be able to measure differences between groups, if differences do in fact exist. Randomization should ensure that participants in experimental and control intervention groups begin studies with the same status on whatever concept or construct the PRO is designed to measure. PROs must be able to detect what is important to patients and distinguish among participants who remain the same, improve, or deteriorate over the course of the trial. This is sometimes referred to as responsiveness, or sensitivity to change.

An instrument with a poor ability to measure change can result in false-negative results in which the experimental intervention improves how patients feel, yet the instrument fails to detect the improvement. This problem may be particularly salient for generic questionnaires that have the advantage of covering all relevant areas of HRQOL, but the disadvantage of covering each area superficially. In studies that show no difference in PROs between experimental and control intervention, lack of instrument responsiveness is one possible reason.

## **17.5 Locating and selecting studies with patient-reported outcomes**

Searching methods for PROs are the same as for other outcomes (see Chapter 6). Usually all reports retrieved by the review's search strategy will be examined to identify those that include the PROs of interest. Sometimes a separate, additional, PRO search might be used to supplement the standard strategy. For example, if a review of randomized trials and other studies in the area of asthma did not yield studies using PROs, a separate search could be performed to include search terms specific to PROs used in asthma, such as 'asthma-specific quality of life'. However, this relies on there being mention of the PROs in the electronic record within the databases searched.

Index terms for PROs differ between the major bibliographic databases. Review authors cannot rely on a single index or subheading search term to identify studies addressing PROs. Multiple search terms are usually necessary. For example, Maciejewski et al. used the following MEDLINE index terms in their systematic review to estimate the effect of weight-loss interventions on health-related quality of life in randomized trials (Maciejewski 2005): 'Contingent valuation'; 'Health status'; 'Health-related Quality of Life'; 'Psychological aspects'; 'Psychosocial'; 'Quality of life'; 'Self-efficacy'; 'SF-36'; 'Utility'; 'Well-being'; 'Willingness to pay'. Free-text searches should also include as many relevant synonyms as possible. The search needs to combine index terms and free text terms and is likely to take several iterations.

Review authors may find it useful to design and use a separate section of the data collection form used in the systematic review to include review of PRO methods and results. An example of such a form can be found on our web site: [www.cochrane-pro-mg.org/documents.html](http://www.cochrane-pro-mg.org/documents.html). Review authors should attend to alternative ways of collecting data from instruments: in particular, whether they can collect

data in forms that facilitate analysis of data both in the form of continuous variables and dichotomous outcomes.

## 17.6 Assessing and describing patient-reported outcomes

Table 17.6.a presents selected issues specific to PROs that review authors should consider in incorporating PROs into their reviews. Authors may want to consider describing PROs in detail, according to this checklist, in the ‘Characteristics of included studies’ table or as an Additional table.

**Table 17.6.a: A checklist for describing and assessing PROs in clinical trials**

Based on Chapter 7 of Patrick and Erickson, a Users’ Guide to the Medical Literature, CDC guidance for evaluation of community preventive services, and criteria used by the Medical Outcomes Trust (Patrick 1993, Guyatt 1997, Zaza 2000, Lohr 2002).

<p>1. What were PROs measuring?</p> <ul style="list-style-type: none"> <li>a. What concepts were the PROs used in the study measuring?</li> <li>b. What rationale (if any) for selection of concepts or constructs did the authors provide?</li> <li>c. Were patients involved in the selection of outcomes measured by the PROs?</li> </ul>
<p>2. Omissions</p> <ul style="list-style-type: none"> <li>a. Were there any important aspects of health (e.g., symptoms, function, perceptions) or quality of life (e.g. overall evaluation, satisfaction with life) that were omitted in this study from the perspectives of the patient, clinician, significant others, payers, or other administrators and decision makers?</li> </ul>
<p>3. If randomized trials and other studies measured PROs, what were the instruments’ measurement strategies?</p> <ul style="list-style-type: none"> <li>a. Did investigators use instruments that yield a single indicator or index number, a profile, or a battery of instruments?</li> <li>b. If investigators measure PROs, did they use specific or generic measures, or both?</li> <li>c. Who exactly completed the instruments?</li> </ul>
<p>4. Did the instruments work in the way they were supposed to work – validity?</p> <ul style="list-style-type: none"> <li>a. Had the instruments used been validated previously (provide reference)? Was evidence of prior validation for use in this population presented?</li> <li>b. Were the instruments re-validated in this study?</li> </ul>
<p>5. Did the instruments work in the way they were supposed to work – ability to measure change?</p> <ul style="list-style-type: none"> <li>a. Are the PROs able to detect change in patient status, even if those changes are small?</li> </ul>
<p>6. Can you make the magnitude of effect (if any) understandable to readers? (You must!)</p> <ul style="list-style-type: none"> <li>a. Can you provide an estimate of the difference in patients achieving a threshold of function or improvement, and the associated number needed to treat (NNT).</li> </ul>

## 17.7 Comparability of different patient-reported outcome measures

Investigators may choose different instruments to measure PROs, either because they use different definitions of a particular PRO or because they choose different instruments to measure the same PRO. For example, an investigator may choose to use a generic instrument to measure functional status or a different disease-specific instrument to measure functional status. The definition of the outcome may or may not differ. Review authors must decide how to categorize PROs across studies, and when to pool results. These decisions will be based in the characteristics of the PRO, which will need to be extracted and reported in the review.

On many occasions, studies using PROs will make baseline and follow-up measurements and the outcome of interest will thus be the difference in change from baseline to follow-up between intervention and control groups. Ideally then, to pool data across two PROs that are conceptually related, one will have evidence of strong longitudinal correlations of change in the two measures in individual patient data, and evidence of similar responsiveness of the instruments. Further supportive evidence could come from correlations of differences between treatment and control, or difference between before and after measurements, across studies. If one cannot find any of these data, one could fall back on cross-sectional correlations in individual patients at a point in time.

For example, the two major instruments used to measure health-related quality of life in patients with chronic obstructive disease are the Chronic Respiratory Questionnaire (CRQ) and the St. George's Respiratory Questionnaire (SGRQ). Correlations between the two questionnaires in individual studies have varied from 0.3 to 0.6 in both cross-sectional (correlations at a point in time) and longitudinal (correlations of change) comparisons (Rutten-van Mólken 1999, Singh 2001, Schünemann 2003, Schünemann 2005).

In a subsequent investigation, investigators examined the correlations between mean changes in the CRQ and SGRQ in 15 studies including 23 patient groups and found a correlation of 0.88 (Puhan 2006). Despite this extremely strong correlation, the CRQ proved more responsive than the SGRQ: standardized response means of the CRQ (median of the standardized response means 0.51, IQR 0.19 to 0.98) were significantly higher ( $P < 0.001$ ) than those associated with the SGRQ (median of the standardized response means 0.26, IQR -0.03 to 0.40). That is, in situations when both instruments were used together in the same study, the CRQ yielded systematically larger treatment effects. As a result, pooling results from trials using these two instruments could lead to underestimates of treatment effect in studies using the SGRQ.

Most of the time, unfortunately, detailed data such as those described in the previous paragraph will be unavailable. Investigators must then fall back on intuitive decisions about the extent to which different instruments are measuring the same underlying construct. For example, the authors of a meta-analysis of psychosocial interventions in the treatment of pre-menstrual syndrome faced a profusion of outcome measures, with 25 PROs reported in their nine eligible studies. They dealt with this problem by having two investigators independently examine each instrument – including all domains – and group them into six discrete conceptual categories; discrepancies were resolved by discussion to achieve consensus. The pooled analysis of each category included between two and six studies.

Meta-analyses of studies using different measurement scales will usually be undertaken using standardized mean differences (SMDs; see Chapter 9, Section 9.2.3). However, SMDs are highly problematic when the focus is on comparing change from baseline in intervention and control groups, because standard deviations of change do not measure between-patient variation (they depend also on the correlation between baseline and final measurements; see Chapter 9, Section 9.4.5.2).

Similar principles apply to studies in which review authors choose to focus on available data that are presented in dichotomous fashion, or from which review authors can extract dichotomous outcome data with relative ease. For example, investigators studying the impact of flavanoids on symptoms of haemorrhoids found that eligible randomized trials did not consistently use similar symptom measures; all but one of 14 trials, however, recorded the proportion of patients either free of symptoms, with symptom improvement, still symptomatic, or worse (Alonso-Coello 2006). In the primary analysis investigators considered outcomes of patients free of symptoms and patients with symptomatic/some improvement as equivalent, and pooled each outcome of interest based on the *a priori* expectation of a similar magnitude and direction of treatment effect.

This left a question of how to deal with studies that reported that patients experienced ‘some improvement’. The investigators undertook analyses comparing the approach of dichotomizing including ‘some improvement’ as a positive outcome and as a negative outcome (similar to no improvement). Dichotomizing outcomes is often very useful, particularly for making results easily interpretable for clinicians and patients. Imaginative and yet rigorous ways of dichotomizing will result in summary statistics that provide useful guides to clinical practice.

The use of multiple instruments for measuring a particular PRO, and experimentation with multiple methods for analysis, can lead to selective reporting of the most interesting findings and introduce serious bias into a systematic review. Review authors focusing on PROs should be alert to this problem. When only a small number of eligible studies have reported a particular outcome, particularly if it is a salient outcome that one would expect conscientious investigators to measure, authors should note the possibility of reporting bias (see Chapter 10).

## 17.8 Interpreting Results

### 17.8.1 Study summaries focusing on a single patient-reported outcome

When a meta-analysis includes studies reporting only a single PRO, presented as a continuous variable, a pooled result will generate a mean difference. The problem with this mean difference is that clinicians may have difficulty with its interpretation. For example, if told that the mean difference between rehabilitation and standard care in a series of randomized trials using the Chronic Respiratory Questionnaire was 1.0 (95% CI 0.6 – 1.5), many readers would have no idea if this represents a trivial, small but important, moderate, or large effect.

The systematic review author can aid interpretation by reporting the range of possible results and the range of mean results in treatment and control groups in the studies. Most useful, however – if it is available – is an estimate of the smallest difference that patients are likely to consider important (the minimally important difference or MID). There are a variety of methods for generating estimates of the MID, including use of global ratings of change (Guyatt 2002). Ideally, review authors will present estimates of the MID in the abstract. For example, investigators examining the impact of respiratory rehabilitation in patients with chronic lung disease on health-related quality of life reported, in their abstract, that “for two important features of HRQL, dyspnea and mastery, the overall effect was larger than the MCID: 1.0 (95% CI 0.6-1.5) and 0.8 (0.5-1.2), respectively, compared with an MCID of 0.5.” (Lacasse 1996).

While this is very helpful, it potentially tempts clinicians to make inappropriate inferences. If the MID is 0.5 and the mean difference between treatments is 0.4, clinicians may infer that nobody benefits from the intervention. If the mean difference is 0.6, they may conclude that everyone benefits. Both inferences may be misguided. First, they ignore the uncertainty (confidence intervals) around the point

estimate. More importantly, they ignore the variation (standard deviation) in responses across individuals.

It is also possible for investigators to provide a ‘responder’ definition to help interpret outcomes (see Chapter 12, Section 12.6.1). It is useful to know the definition that characterizes an individual patient as a responder to treatment. Such a responder definition is based upon pre-specified criteria backed by empirically derived evidence supporting the responder definition as a measure of benefit. Methods for defining a responder include: (1) a pre-specified change from baseline on one or more scales; (2) a change in score of a certain size or greater (e.g. a 2-point change on an 8-point scale); and (3) a percentage change from baseline.

## 17.8.2 Study summaries using more than one patient-reported outcome

As the discussion in Section 17.8.1 pointed out, when pooling across PROs the mean difference is no longer a possible measure of effect and we therefore replace it with the standardized mean difference (SMD) (see Chapter 9, Section 9.2.3). Unfortunately, there are no fully satisfactory ways of providing a sense of the magnitude of effect in a PRO when one has had to resort to SMD to generate a summary. One can offer readers standard rules of thumb in interpretation of effect sizes (for instance 0.2 represents a small effect, 0.5 a moderate effect, and 0.8 a large effect (Cohen 1988) or some variation (<0.41 = small, 0.40 to 0.70 = moderate, >0.70 = large). Another, perhaps even less satisfactory, approach suggests that a standardized mean difference of 0.5 approximates, in many cases, a minimal important difference (Norman 2003).

General methods of reporting and interpreting PROs, and other clinical outcomes, with respect to drawing inferences and conclusions are discussed in Chapter 12 (Section 12.6).

## 17.8.3 When studies do not address patient-reported outcomes

Many primary studies fail to measure aspects of perceived health and quality of life that are very important to patients. When this is the case, evidence regarding interventions’ impact on PROs may be much weaker than evidence regarding impact on disease indicators morbidity or mortality. In the extreme, no study may address PROs directly. The careful prior consideration of all patient-important outcomes will highlight what is missing in outcome measurement in the eligible randomized trials and other studies. This omission should be highlighted in the reviews authors’ conclusions as an implication for future research.

## 17.9 Chapter information

**Authors:** Donald L Patrick, Gordon H Guyatt and Catherine Acquadro on behalf of the Cochrane Patient Reported Outcomes Methods Group.

**This chapter should be cited as:** Patrick D, Guyatt GH, Acquadro C. Chapter 17: Patient-reported outcomes. In: Higgins JPT, Green S (editors). *Cochrane Handbook for Systematic Reviews of Interventions*. Version 5.0.1 [updated September 2008]. The Cochrane Collaboration, 2008. Available from [www.cochrane-handbook.org](http://www.cochrane-handbook.org).

**Acknowledgements:** Jason Busse, Peter Fayers, Toshi Furukawa, Madeleine King and Milo Puhan provided comments on drafts.

### Box 17.9.a: The Cochrane Patient Reported Outcomes Methods Group

The main objective of the Patient Reported Outcomes Methods Group (PRO MG) is to advise Cochrane authors about when and how to incorporate health status and quality-of-life data

into systematic reviews. Some Cochrane Review Groups have encountered difficulties when incorporating PRO data in reviews. Examples of such difficulties include pooling and interpreting data and evaluating the validity of PRO scales.

The PRO MG aims to:

- refine methods of literature search on PRO studies;
- develop methods for systematically reviewing HRQL studies;
- refine methods for meta-analysis of PRO studies (in collaboration with the Statistical Methods Group);
- refine methods for use of PRO measures in economic evaluations in collaboration with the Campbell-Cochrane Economics Methods Group; and
- advise on software development.

The group gives advice to the Cochrane Collaboration Steering Group upon request, convenes workshops on health and patient-reported outcomes issues and methods, in response to the needs of the Collaboration, and prepares recommendations for this *Handbook*. Members of the group will take part in the preparation of Cochrane reviews and will give advice to authors through written material and training workshops. Members of the group will help review authors to develop protocols and reviews where it has been decided to include PRO outcomes.

*Web site:* [www.cochrane-pro-mg.org/](http://www.cochrane-pro-mg.org/)

## 17.10 References

### **Alonso-Coello 2006**

Alonso-Coello P, Zhou Q, Martinez-Zapata MJ, Mills E, Heels-Ansdell D, Johanson JF, Guyatt G. Meta-analysis of flavonoids for the treatment of haemorrhoids. *British Journal of Surgery* 2006; 93: 909-920.

### **Cohen 1988**

Cohen J. *Statistical Power Analysis in the Behavioral Sciences* (2nd edition). Hillsdale (NJ): Lawrence Erlbaum Associates, Inc., 1988.

### **Guyatt 1993**

Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. *Annals of Internal Medicine* 1993; 118: 622-629.

### **Guyatt 1997**

Guyatt GH, Naylor CD, Juniper E, Heyland DK, Jaeschke R, Cook DJ. Users' guides to the medical literature. XII. How to use articles about health-related quality of life. Evidence-Based Medicine Working Group. *JAMA* 1997; 277: 1232-1237.

### **Guyatt 2002**

Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR, Clinical Significance Consensus Meeting Group. Methods to explain the clinical significance of health status measures. *Mayo Clinic Proceedings* 2002; 77: 371-383.

### **Guyatt 2004**

Guyatt G, Montori V, Devereaux PJ, Schünemann H, Bhandari M. Patients at the center: in our practice, and in our use of language. *ACP Journal Club* 2004; 140: A11-A12.

**Lacasse 1996**

Lacasse Y, Wong E, Guyatt GH, King D, Cook DJ, Goldstein RS. Meta-analysis of respiratory rehabilitation in chronic obstructive pulmonary disease. *The Lancet* 1996; 348: 1115-1119.

**Lohr 2002**

Lohr K. Assessing health status and quality-of-life instruments: attributes and review criteria. *Quality of Life Research* 2002; 11: 193-205.

**Maciejewski 2005**

Maciejewski ML, Patrick DL, Williamson DF. A structured review of randomized controlled trials of weight loss showed little improvement in health-related quality of life. *Journal of Clinical Epidemiology* 2005; 58: 568-578.

**Martinez-Devesa 2007**

Martinez-Devesa P, Waddell A, Perera R, Theodoulou M. Cognitive behavioural therapy for tinnitus. *Cochrane Database of Systematic Reviews* 2007, Issue 1. Art No: CD005233.

**Norman 2003**

Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Medical Care* 2003; 41: 582-592.

**Patrick 1993**

Patrick DL, Erickson P. *Health Status and Health Policy: Quality of Life in Health Care Evaluation and Resource Allocation*. New York (NY): Oxford University Press, 1993.

**Puhan 2006**

Puhan M, Soesilo I, Guyatt GH, Schünemann HJ. Combining scores from different patient reported outcome measures in meta-analyses: when is it justified? *Health and Quality of Life Outcomes* 2006; 4: 94.

**Rutten-van Mólken 1999**

Rutten-van Mólken M, Roos B, Van Noord JA. An empirical comparison of the St George's Respiratory Questionnaire (SGRQ) and the Chronic Respiratory Disease Questionnaire (CRQ) in a clinical trial setting. *Thorax* 1999; 54: 995-1003.

**Schünemann 2003**

Schünemann HJ, Griffith L, Jaeschke R, Goldstein R, Stubbing D, Guyatt GH. Evaluation of the minimal important difference for the feeling thermometer and the St. George's Respiratory Questionnaire in patients with chronic airflow obstruction. *Journal of Clinical Epidemiology* 2003; 56: 1170-1176.

**Schünemann 2005**

Schünemann HJ, Goldstein R, Mador MJ, McKim D, Stahl E, Puhan MA, Griffith LE, Grant B, Austin P, Collins R, Guyatt GH. A randomised trial to evaluate the self-administered standardised chronic respiratory questionnaire. *European Respiratory Journal* 2005; 25: 31-40.

**Singh 2001**

Singh SJ, Sodergren SC, Hyland ME, Williams J, Morgan MD. A comparison of three disease-specific and two generic health-status measures to evaluate the outcome of pulmonary rehabilitation in COPD. *Respiratory Medicine* 2001; 95: 71-77.

**Tarlov 1989**

Tarlov AR, Ware JE, Jr., Greenfield S, Nelson EC, Perrin E, Zubkoff M. The Medical Outcomes Study. An application of methods for monitoring the results of medical care. *JAMA* 1989; 262: 925-930.

**US Food and Drug Administration 2006**

US Food and Drug Administration. Guidance for Industry. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims [February 2006]. Available from: <http://www.fda.gov/cber/gdlns/probl.htm> (accessed 1 January 2008).

**Ware 1995**

Ware JE, Kosinski M, Bayliss MS, McHorney CA, Rogers WH, Raczek A. Comparison of methods for the scoring and statistical analysis of SF-36 health profile and summary measures: summary of results from the Medical Outcomes Study. *Medical Care* 1995; 33: AS264-AS279.

**Zaza 2000**

Zaza S, Wright-De Agüero LK, Briss PA, Truman BI, Hopkins DP, Hennessy MH, Sosin DM, Anderson L, Carande-Kulis VG, Teutsch SM, Pappaioanou M, Task Force on Community Preventive Services. Data collection instrument and procedure for systematic reviews in the Guide to Community Preventive Services. *American Journal of Preventive Medicine* 2000; 18 (Suppl 1): 44-74.