# Patient reported outcomes: general principles of development and interpretability

**Dianne Bryant[1], Holger Schünemann[2], Jan Brożek[3], Roman Jaeschke[4], Gordon Guyatt[5]**

[1] University of Western Ontario, London, Ontario, Canada

[2] Department of Epidemiology, Istituto Nazionale Tumori Regina Elena, Roma, Italy

[3] Department of Medicine, Jagiellonian University School of Medicine, Kraków, Poland

[4] Department of Medicine, McMaster University, Hamilton, Ontario, Canada

[5] Department of Clinical Epidemiology & Biostatistics, McMaster University, Hamilton, Ontario, Canada

**Abstract:** Direct measurement of how people are feeling and the extent to which they are functioning in daily activities (generally as patient reported outcomes) is critical to judging the benefit of health interventions in chronic conditions. Selection of an appropriate instrument will reflect a comprehensive understanding of the condition of interest and a thorough knowledge of the expected benefits and harms of the proposed intervention. We provide a brief discussion about different ways that health and health measurement have been defined, including the International Classification of Function, Disability and Health (ICF), Health Related Quality of Life (HRQOL) and cost-to-benefit analyses. We outline important properties (reliability, validity, and responsiveness) that a measurement instrument must demonstrate depending on the purpose of measurement, and provide insight as to how to interpret the results of studies that report patient reported outcomes.

**Key words:** health, outcomes, quality of life

## INTRODUCTION

Motivation for including patient reported health outcomes as the primary measure of the effectiveness of an intervention has increased over the last two decades. Indeed, entire programs of research are dedicated to the science of developing, refining, and testing instruments to measure patient important health-related outcomes. Direct measurement of how people feel and the extent to which they can function in their daily activities is replacing physiologic or laboratory tests as primary outcomes of interest in clinical studies for chronic disease populations. This shift is motivated by the realization that changes in physiologic endpoints often bear a limited relation with changes in patient reported health status making them inappropriate surrogates for patient-important endpoints [1-5].

Measurement of health from the patient's perspective is important when the goal of treatment is to improve how the patient is feeling, rather than to prolong life. However, even when the primary objective is to prolong life or to reduce the incidence of seemingly straightforward outcomes such as stroke or myocardial infarction, measurement from the patient's perspective may be important to capture the variability in patient's function and feelings: e.g. a mild versus severe stroke, large versus small infarct or painful versus painless death.

Consider the GOAL study, in which patients with asthma were treated either with inhaled corticosteroids (ICS) or ICS and long acting beta-agonists (LABA) where the goal of treatment was to achieve total (essentially no symptoms) or well control (minor and easily controlled symptoms) asthma. One of the instruments used to measure the severity of asthma and the effect of treatment was the Asthma Quality of Life Questionnaire (AQLQ) [6,7]. The AQLQ consists of 32 questions in four domains: activity limitation, symptoms, emotional function and environmental stimuli. Responses in each domain and an overall score are graded on a 7-point scale, where 1 represents "total impairment" and 7 represents "no impairment". The AQLQ was administered in this study before treatment was administered (baseline) and after the commencement of treatment at weeks 4, 12, 24, 36, 48 and 52. AQLQ scores are presented as the mean of each domain, as well as an overall score. How should clinicians interpret these results so that they can be used to guide practice?

This paper will explain terms relevant to the understanding of health and health measurement; provide an overview of the key measurement properties, a brief overview of some of the more common methods used to determine when important change has occurred and how to interpret the results of studies that report patient reported outcomes.

**Correspondence to:**
Dianne Bryant, MSc, PhD, Assistant Professor Faculty of Health Sciences, Elborn College, Room 1438, University of Western Ontario, London, ON, Canada, N6G 1H1, phone: 519-661-2111, fax: 519-661-3866, e-mail: dianne.bryant@uwo.ca
**Received:** April 18, 2007. **Accepted in final form:** July 3, 2007.
**Conflict of interest: none declared.**
Pol Arch Med Wewn. 2007; 117 (4): 125-131

## How is health described and measured?

### The World Health Organization

The World Health Organization (WHO) defines health as "a state of complete physical, mental, and social well-being" [8]. The WHO's International Classification of the Consequences of Disease, Impairment, Disability and Handicap (ICIDH) [9], more recently titled, International Classification of Functioning, Disability and Health (ICF) [10], was developed to provide a standard language and framework to describe and measure health and health-related states.

Within the ICF system, health outcomes are classified according to the effect upon body function, body structure, limitations in activities, and limitations in participation. Health outcomes that measure body function include measures of physiological functions of body systems (e.g. ejection fraction, glucose level, depression, pain, etc), whereas outcomes that measure body structures include measures of anatomical parts and their components (e.g. x-ray to measure fracture healing, computed tomography to measure tumor size, etc). Activity is defined as the performance of a task or action. Participation is the involvement of an individual in meaningful, fulfilling and satisfying activities that are socially or culturally expected of that person. Impairments can be thought of as problems with body functions or structures. Having an impairment of a body structure (e.g. disc hernia) or function (e.g. range of motion) may contribute to limitations in activities, including activities of daily living, walking, or driving a car, that might also contribute to restrictions in participation. Comprehensive assessment of patient health will include measures of body systems and function, as well as limitations in activities and participation.

### Health-Related Quality of Life

Health Related Quality of Life (HRQOL) instruments measure the broad concept of health (physical, mental, and social well-being) by inquiring into the extent of difficulty with activities of daily living, (including work, recreation, and household management), and ensuing difficulties in relationships with family, friends, and social groups; capturing not only the ability to function within these roles, but also the degree of satisfaction derived from doing them.

Within the construct of HRQOL, it is common to come across the terms *generic* and *disease-specific*. A generic instrument measures general health status including physical symptoms, function, and emotional dimensions of health relevant to all health states, including healthy individuals[11]. Generic HRQOL instruments are useful when measuring the impact of a specific illness or injury across different diseases, severities, and interventions [11]. The disadvantage of these types of measures is that because of their broad scope they are often not sensitive enough to detect small, but important, changes in health within the specific population under study. A number of previously widely used health profiles such as the Sickness Impact Profile (SIP) [12,17], and the Nottingham Health Profile (NHP) [18,23] are now of largely historical interest; health profiles developed from the Medical Outcomes Study, including the 36-Item Short-Form Health Survey (SF-36) [24-24] and 12-Item Short-Form Health Survey (SF-12) [27] have come to dominate the field of generic health status measurement.

In contrast, compared to generic health instruments, disease-specific measures are tailored to inquire about specific aspects of health that are affected by the disease of interest (for example, specific to asthma). Disease-specific instruments are usually more responsive to small but important changes in health than the generic instruments (Wiebe S, Guyatt G, Weaver B, Matijevic S, Sidwell C. Comparative responsiveness of generic and specific quality-of-life instruments. J Clin Epidemiol. 2003; 56: 52-60) Consequently, because they are so focused, disease-specific HRQOL instruments cannot be used to compare the impact of one illness to another, and in some cases disease-specific measures are so specific that comparisons between different populations within the same disease may not be possible (e.g. instruments tailored specifically for pediatric versus adult populations).

## Cost-to-benefit analyses

When making decisions on behalf of patient groups, decision-makers weigh the benefits and risks of treatment, but must also consider whether the benefits are substantial enough to warrant the health care resources expended to provide them. Such considerations may also play a role in individual patient decision-making. An economic analysis can inform these decisions. The main distinction between economic analyses and other studies is the explicit measurement and valuation of both resource consumption and patient-important benefit and harm. This paradigm of health measurement is rooted in decision-theory [28-30] and includes a quantitative technique of specifying the alternatives or choices that are available to the patient, information or knowledge of relevant events and their probabilities, and preferences or utilities (measures of the desirability of various outcomes) to the patient.

To compare the costs and benefits of different treatments for different diseases (required if, for instance, one is making a decision about allocating scarce resources to a new drug for cancer or a treatment for children with autism) necessitates being able to measure benefits and harms of alternative interventions using the same units. One way of creating the same units is through the concept of utility – the value people place on health benefits, and avoiding poor health outcomes. Therefore, similar to generic instruments, utility outcomes can be used to measure the impact of different interventions across different diseases.

There are a few common approaches to measuring preferences or utility. One method evaluates the preferences of the individual patient by asking them to make a decision un-

der uncertainty (measure of utility); for examples include the Standard Gamble [31]. During administration of the Standard Gamble, the patient imagines that there is an intervention that will result in a return to perfect health but that there is a risk of death with the intervention. The patient is asked to specify the largest probability of death they would be willing to accept before declining the intervention and choosing to remain in their current (sub-optimal) health state. The larger the probability of death that the patient is willing to accept, the lower value the respondent places or their current health state. The value or utility of the present health state – as in all utility measures – is placed on a continuum between death (typically give a value of 0) and full health (typically given a value of 1.0). In this setting one year of life with the utility of 0.5 is worth half a year adjusted for its quality – hence concept of QALY (quality adjusted live year).

The Time Trade-Off [32], asks patients to imagine living their life in their current health state and to contrast this with the alternative of perfect health in exchange for a shorter lifespan (preference-based measured). The administrator provides alternatives of years of life in the present health state versus years of life in perfect health. The more years a patient is willing to sacrifice in exchange for a return to perfect health, the worse the patient perceives their current health state. Utility is calculated by subtracting the number of years sacrificed from the number of years of life remaining divided by the number of years remaining. The number of years remaining is estimated using actuarial tables. Another common preference-based measure is the Feeling Thermometer (FT), where patients rate their health status using a visual analogue scale presented in the form of a thermometer from 0 (worst) to 100 (best) [33-35].

The final approach we will describe focuses on the preferences of the general population, using rating scales such as the Quality of Well-Being Scale [36], Health Utilities Index (HUI) [37-41], European Quality of Life Scale (EuroQoL-5D) [42,43], where patients are asked to rate their ability to function in physical, emotional and social aspects of life. Here, patients report on their health state rather than the value they put on their health state. The patient's utility is assigned on the basis of a mathematical model using preference ratings of health states that have been derived from a random sample of the general population. Some have suggested that generic health scores from health profiles such as the SF-12 can be mapped onto utility scales, though the validity of these methods remains controversial [44-50].

## What are the properties of a good measurement instrument?

The choice of outcome measure should align itself with a study's objectives. The intent may be to discriminate between patients with different disease severity at a point in time (e.g. whose asthma is impairing function to a greater degree and who to a lesser degree), to predict patient outcome (e.g. functional status may predict mortality in heart failure patients) or to evaluate change following an intervention (e.g. which asthma patients have improved and which have not). To be useful for application in a research and clinical setting for the first two purposes (discrimination according to severity and predicting outcome), instruments must be valid (measure what they are supposed to measure – discriminative validity) and reliable (provide consistent ratings between repeated measures in a stable population). If the intention is to evaluate change after treatment, the instrument must be valid (longitudinal validity) and responsive (able to detect important change, even if the magnitude of change is small).

### Validity

An assessment of the validity of a new instrument is an evaluation of the extent to which the instrument measures what it was intended to measure. With respect to patients with asthma, you need an instrument capable of discriminating between patients with asthma who have varying degrees of control and functional disability. An invalid tool might appear to be measuring functional ability, but if poorly constructed, may in fact be measuring satisfaction with medical care, or patients' emotions about their current situation.

With respect to investigations into the effectiveness of new interventions that report quality of life – how do clinicians know whether the instrument that investigators have selected measures aspects of life that patient's value? There are several ways that investigators might go about demonstrating this. They may include a description of how an instrument was developed. Instruments with the greatest potential for validity will have consulted with patients (and perhaps clinician experts or patients' family members) who have experience with the disease and how the disease affects their lives from a physical, mental and social standpoint (this is the approach used in the development of the Asthma Quality of Life Questionnaire (AQLQ) used in GOAL study).

Alternatively, investigators may elect to cite articles that describe the development and testing phases in detail – though ideally, investigators will include a description of the instrument that includes sufficient detail to obviate the need to review the citation itself. In some cases, the investigators will describe the content of the questionnaire or include the instrument in an appendix (more common in online versions of the article than in hard copy) so that clinicians can use their own experience to decide whether what is being measured is important to patients (*face validity*).

There are several strategies that the developers of a new instrument may use to provide empirical evidence of the validity of the outcome measure. For example, the authors may describe an investigation into the *criterion validity* of the instrument to assess whether the instrument behaves the way it should when compared to a gold standard measurement. Since there is no gold standard reference for quality of life, this will be unusual. The only circumstance when criterion validity is relevant is when investigators try to develop a shorter measure

of an existing instrument, in which case the longer, already existing measure serves as a gold standard.

*Construct validity* assesses the extent to which the instrument relates to other measurements in the way that it should. Types of construct validity include, convergent and discriminant validity. *Convergent validity* examines the degree to which interpretations of scores on the instrument being tested are similar to the interpretation of scores on other instruments that theoretically measure similar constructs. For example, a new emotional function measure should correlate highly to an existing measure of emotional function. *Discriminant validity* examines the degree to which the construct (e.g. health related quality of life) does not correlate to a dissimilar construct. For example, if the theory was that quality of life is not related to intelligence then there should not be a strong correlation between the two measures.

The appropriate way to measure validity for discriminative instruments is by looking at the correlations between measures at a single point in time (do asthma patients with better control and higher functional status do better on a respiratory testing, and do those patients with poorer functional capacity do less well). Such correlations reflect an instrument's *cross-sectional construct validity*.

The appropriate way to measure validity for evaluative instruments is by looking at the correlations in change over time between measures (do asthma patients with improved functional or emotional status also show improvement on a spirometry, and do those with deterioration in functional capacity demonstrate decrements in respiratory function tests). Such correlations reflect an instrument's *longitudinal construct validity*.

## Reliability

Reliability refers to the extent to which an instrument discriminates between individuals in a population in a consistent manner when respondents are in stable health. Reliability is relevant for discriminative and predictive instruments. The mathematical relationship that defines reliability can be explained by the ratio of the variability in scores between patients to the total variability (i.e. between and within patient variability). Scores obtained on a reliable instrument will demonstrate relatively small differences in scores upon repeated administrations in patients who are stable in their condition (i.e. small within person variability). Reliability will always appear to be greater when measured in a heterogenous population with greater variability in scores between patients (e.g. includes patients with no limitations to those with severe limitations) than in a homogeneous population.

An instrument free of random error will have a reliability of 1.0 as long as there is some between-patient variability. As the amount of random error increases in relation to the between-patient variability, the measure of reliability will approach 0. Common expressions of the magnitude of reliability are *Kappa*, when the scale is categorical and *intraclass correlation coefficient* (ICC) when the scale is continuous. There are several potential influences that may affect the reliability of an instrument including learning effects, regression to the mean, alterations in mood, circumstance and conditions of administration, and the length of time between assessments. It is also possible that real changes have occurred between consecutive assessments. The most important frequently neglected determinant of reliability is the variability in patient's status on the underlying attribute.

Different techniques to measure the reliability of an instrument include test-retest, inter-rater and internal consistency reliability. *Test-retest reliability* is a measure of the magnitude of the agreement between ratings in repeated administrations of the instrument in a population with a stable health condition. *Inter-rater* reliability is a measure of the magnitude of the agreement between ratings given by different raters administering the same instrument in a population with a stable health condition. *Internal consistency reliability* assesses the homogeneity of the items that make up the instrument. The internal consistency reliability coefficient (R) is used to calculate the standard error of measurement (SEM), which provides an easily defined estimate of the reproducibility of individual measurements ($SEM = \sigma (1\text{-}R)^{1/2}$) and can be used to determine whether true change has occurred within an individual ($\sqrt{2}$ x SEM) [51]. Internal consistency is very limited as a measure of reliability because it relates only to the correlation between items on a single administration, and makes no attempt to assess the degree of variability on repeated administration of a measure.

## Responsiveness

Responsiveness, sometimes called sensitivity to change, has been defined in the past as the ability of an instrument to measure true change in the state being measured regardless of whether it is relevant or meaningful to the patient or clinician [52]. More recently responsiveness has been defined as the ability of the instrument to detect change in the state being measured that is important to the patient) even if that difference is small [52,53].

The smallest difference in score in the outcome of interest that informed patients or informed proxies perceive as important, either beneficial or harmful, and that would lead the patient or clinician to consider a change in the management is defined as the minimally important difference (MID) [54,55]. The magnitude of change that constitutes an MID for most patient reported outcome measures is not self-evident, creating difficulties with interpreting the results of studies that report changes in patient reported outcomes. In studies that show no difference in HRQL when patients receive a treatment versus a control intervention, clinicians should look for evidence that the instrument has been shown to be responsive to small or moderate-sized effects in a similar population in previous investigations. In the absence of this evidence it is unknown whether the intervention was ineffective or whether

the instrument was not responsive. For example, in the GOAL study, the authors provide a description of the MID, with appropriate citations to articles that report the measurement properties of the questionnaire, as a within-subject change of 0.5 points on either the overall AQLQ score or any of the individual domains is considered the minimum change to be considered clinically meaningful [6,7,56].

## How do I interpret the results of a study that reports patient reported outcomes?

Physicians often have limited familiarity with methods of measuring how patients feel or their ability to do the things the need or want to do. At the same time, published articles recommend administering or withholding treatment on the basis of its impact on patients' well-being. Thus, if a measure is to be clinically useful, its scores must be interpretable. Interpretability is greatly enhanced if we know the magnitude of the change in score that is important – the MID.

Strategies to define important change have included distribution-based approaches and anchor-based approaches. In general, distribution-bases approaches relate the magnitude of the effect to some measure of variability. For example in a simple before/after comparison one could calculate the difference between scores before and after treatment divided by the standard deviation of scores at baseline; we call the resultant statistic an "effect size". In a parallel groups design one calculates the difference in scores between the treatment and control group divided by the standard deviation of the change that patients experienced during the study to generate the effect size.

A rough rule of thumb for interpreting effects sizes is that changes of a magnitude of 0.2 represent small changes, 0.5 moderate changes and 0.8 large changes [57]. Interpretation using effect sizes remains problematic because it is sensitive to the homogeneity of the distribution of the sample of patients who participated in the study (i.e. estimates of variability will vary from study to study). In other words, the same difference between treatment and control will appear as a large effect size if the sample is homogenous (patients are similar and thus there is a small between-patient standard deviation) and as a small effect size if the sample is heterogeneous (patients are dissimilar and thus there is a large between-patient standard deviation).

On the other hand, anchor-based approaches involve comparing the magnitude of the change observed on a patient reported outcome to an anchor or independent standard that is itself interpretable. The anchor may be defined by achieving change on some external criteria; for example changing category increasing on a well-known classification system for disease or functional severity (e.g. moving from New York Heart Association Functional Classification III to II) or moving in or out of a diagnostic category (e.g. from depressed to non-depressed, or the reverse).

Another common anchor-based approach follows patients longitudinally and asks patients to report whether they got better, stayed the same or got worse. If better or worse, patients rate the degree of change – for example, they may rate the degree of change from 1 (minimal change) to 7 (a very large change), where 1 to 3 indicates a small but important change. In the most common way of using this approach, the investigators estimate the MID as the average of the change scores on the patient reported outcome that corresponds to a small but important change (that is, the average change in patients who have rated themselves as 1 to 3 on the degree of change rating).

One application of the MID is to help compute the proportion of patients benefiting from an intervention by at least the minimally patient-important amount. In this application, investigators compare the proportion of patients benefiting in the treatment group to the proportion of patients benefiting in the control group. The difference in proportions can be converted to a Number Needed to Treat (NNT) – an expression of the number of patients that need to be treated to achieve an important benefit in a single patient – found by calculating the reciprocal of the difference in the proportion of patients in each group who experience meaningful change.

For instance, a randomized trial of respiratory rehabilitation in patients with chronic lung disease measured health-related quality of life using the Chronic Respiratory Questionnaire (CRQ) [58]. The MID for measures of both dyspnea and fatigue on the CRQ is 0.5 on a scale of 1 to 7 [59]. In this trial, the difference between treatment and control in dyspnea was 0.6 and the difference in fatigue was 0.45 (both in favor of the intervention). Given the MID of 0.5 one might be tempted to conclude that the intervention had an important effect on dyspnea but not on fatigue. The additional proportion with an improvement or deterioration in dyspnea greater than 0.5 was, however, 0.24 (yielding a number needed to treat of 4.1 for one additional patient to have an important benefit in day-to-day dyspnea). The corresponding numbers for fatigue were 0.23 and 4.4.

Knowing whether the results of a study that report patient reported outcomes are relevant to your clinical practice means understanding the patient's experience of the disease. Even the most common symptoms of a chronic disease do not affect all of those afflicted and different patients will cope with symptoms in various ways. Ideally, one would measure the effect of treatment on the individual patient. When a clinician is using the results of a clinical trial that reports patient reported outcomes it is likely to be more informative to relay the results of disease-specific outcomes than more generic instruments. For instance, in the example above one might tell patients contemplating putting the time and energy (and possibly the money) involved in a respiratory rehabilitation program that their chances of achieving an important improvement are approximately 25%.

Returning to the example of the GOAL study, it is evident from the results that the majority of patients with asthma who participated in this study experienced major improvement in their quality of life compared to their baseline scores; well over

80% of all patients improved by more than 0.5 points per question. Achieving total control was associated with an almost 2 point per question improvement – demonstrating major clinical importance. In addition, changes in the questionnaire were able to differentiate among different levels of control – total, well, and not-well controlled. This study, which concentrated on patient-specific outcomes (rather than physiological only), showed also, that optimizing asthma treatment may markedly improve patients quality of life, reaching essentially normal levels in more than half of the patients.

## SUMMARY

Consideration of the impact of treatment on patients' quality of life is important to clinicians when making informed decisions about treatment options. For the purpose of evaluating the merits of studies reporting patient reported outcomes, it is important to understand the general principles of validity, reliability and responsiveness as well as how to interpret the results so that they can be applied to a clinical setting.

## REFERENCES

1. Jaeschke R, Guyatt GH, Willan A, et al. Effect of increasing doses of beta agonists on spirometric parameters, exercise capacity, and quality of life in patients with chronic airflow limitation. Thorax. 1994; 49: 479-484.
2. Juniper EF, Svensson K, O'Byrne PM, et al. Asthma quality of life during 1 year of treatment with budesonide with or without formoterol. European Respiratory Journal. 1999; 14: 1038-1043.
3. Stratford PW, Kennedy D, Pagura SM, Gollish JD. The relationship between self-report and performance-related measures: questioning the content validity of timed tests. Arthritis Rheum. 2003; 49: 535-540.
4. Galatz LM, Ball CM, Teefey SA, et al. The outcome and repair integrity of completely arthroscopically repaired large and massive rotator cuff tears. J Bone Joint Surg Am. 2004; 86: 219-224.
5. Roddey TS, Cook KF, O'Malley KJ, Gartsman GM. The relationship among strength and mobility measures and self-report outcome scores in persons after rotator cuff repair surgery: impairment measures are not enough. J Shoulder Elbow Surg. 2005; 14 (1 Suppl S): S95-S98.
6. Juniper EF, Guyatt GH, Epstein RS, et al. Evaluation of impairment of health related quality of life in asthma: development of a questionnaire for use in clinical trials. Thorax. 1992; 47: 76-83.
7. Juniper EF, Guyatt GH, Willan A, Griffith LE. Determining a minimal important change in a disease-specific Quality of Life Questionnaire. J Clin Epidemiol. 1994; 47: 81-87.
8. Preamble to the Constitution of the World Health Organization as adopted by the International Health Conference. Official Records of the World Health Organization, no. 2, p. 100 and entered into force on 7 April 1948 19-22 June, 1946; signed on 22 July 1946 by the representatives of 61 States; New York.
9. World Health Organization. International Classification of Impairments, Disabilities, and Handicaps: A manual of classification relating to the consequences of disease. Geneva, World Health Organization, 1980.
10. World Health Organization. Towards a common language for functioning, disability, and health: ICF The International Classification of Impairment, Disability and Health. Geneva, World Healtlh Organization, 2002. Report No: WHO/EIP/GPE/CAS/01.3.
11. Jackowski D, Guyatt G. A guide to health measurement. Clinical Orthopaedics & Related Research. 2003; 413: 80-89.
12. Bergner M, Bobbitt RA, Carter WB, Gilson BS. The Sickness Impact Profile: development and final revision of a health status measure. Med Care. 1981; 19: 787-805.
13. Bergner M, Bobbitt RA, Kressel S, et al. The sickness impact profile: conceptual formulation and methodology for the development of a health status measure. Int J Health Serv. 1976; 6: 393-415.
14. Bergner M, Bobbitt RA, Pollard WE, et al. The sickness impact profile: validation of a health status measure. Med Care. 1976 ; 14: 57-67.
15. de Bruin AF, Buys M, de Witte LP, Diederiks JP. The sickness impact profile: SIP68, a short generic version. First evaluation of the reliability and reproducibility. J Clin Epidemiol. 1994; 47: 863-871.
16. de Bruin AF, Diederiks JP, de Witte LP, et al. The development of a short generic version of the Sickness Impact Profile. J Clin Epidemiol. 1994; 47: 407-418.
17. de Bruin AF, Diederiks JP, de Witte LP, et al. Assessing the responsiveness of a functional status measure: the Sickness Impact Profile versus the SIP68. [Review] [38 refs]. J Clin Epidemiol. 1997; 50: 529-540.
18. Hunt SM, McEwen J. The development of a subjective health indicator. [Review] [62 refs]. Sociol Health Illn. 1980; 2: 231-246.
19. Hunt SM, McKenna SP, McEwen J, et al. A quantitative approach to perceived health status: a validation study. J Epidemiol Community Health. 1980; 34: 281-286.
20. Hunt SM, McKenna SP, McEwen J, et al. The Nottingham Health Profile: subjective health status and medical consultations. Soc Sci Med. 1981;15: 221-229.
21. Hunt SM, McKenna SP, Williams J. Reliability of a population survey tool for measuring perceived health problems: a study of patients with osteoarthrosis. J Epidemiol Community Health. 1981; 35: 297-300.
22. Hunt SM, McEwen J, McKenna SP. Measuring health status: a new tool for clinicians and epidemiologists. J R Coll Gen Pract. 1985; 35: 185-188.
23. McKenna SP, McEwen J, Hunt SM, Papp E. Changes in the perceived health of patients recovering from fractures. Public Health. 1984; 98: 97-102.
24. McHorney CA, Ware JE Jr, Raczek AE. The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. Med Care. 1993; 31: 247-63.
25. McHorney CA, Ware JE Jr, Lu JF, Sherbourne CD. The MOS 36-item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. Med Care. 1994; 32: 40-66.
26. Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care. 1992; 30: 473-483.
27. Ware JE Jr, Kosinski M, Keller SD. A 12-Item Short-Form Health Survey: Construction of Scales and Preliminary Tests of Reliability and Validity. Medical Care. 1996; 34: 220-233.
28. Wiebe S, Guyatt G, Weaver B, et. al. Comparative responsiveness of generic and specific quality-of-life instruments. J Clin Epidemiol. 2003; 56: 52-60.
29. Influence Diagrams. In: Howard RA, Matheson JE, eds. The Principles and Applications of Decision-Analysis Menol Park: Strategic Decisions Group, 1984.
30. Ledley RS, Lusted LB. Reasoning foundations of medical diagnosis. Science. 1959; 130: 9-21.
31. Sox HC. Decision analysis: a basic clinical skill? N Engl J Med. 1987; 316: 271-272.
32. Von Neumann J, Morgenstern O. Theory of Games and Economic Behavior. 3rd ed. New York, Wiley, 1953.
33. Torrance GW, Thomas WH, Sackett DL. A utility maximization model for evaluation of health care programs. Health Serv Res. 1972; 7: 118-133.
34. Puhan MA, Guyatt GH, Montori VM, et al. The standard gamble demonstrated lower reliability than the feeling thermometer. J Clin Epidemiol. 2005; 58: 458-465.
35. Schünemann HJ, Griffith L, Jaeschke R, et al. Evaluation of the minimal important difference for the feeling thermometer and the St. George's Respiratory Questionnaire in patients with chronic airflow obstruction. J Clin Epidemiol. 2003; 56: 1170-1176.
36. Schünemann HJ, Griffith L, Stubbing D, et al. A clinical trial to evaluate the measurement properties of 2 direct preference instruments administered with and without hypothetical marker states. Medical Decision Making. 2003; 23 : 140-149.
37. Kaplan RM, Anderson JP. The Quality of Well-Being Scale! Rationale for a Single Quality of Life Index. In: Walkee SR, Rosser R, eds. Quality of Life. Assessment and Application. London, MTP Press, 1988: 51-77.
38. Torrance GW, Furlong W, Feeny D, Boyle M. Multi-attribute preference functions. Health Utilities Index. [Review] [83 refs]. Pharmacoeconomics. 1995; 7: 503-520.
39. Feeny D, Furlong W, Boyle M, Torrance GW. Multi-attribute health status classification systems. Health Utilities Index. [Review] [58 refs]. Pharmacoeconomics. 1995; 7: 490-502.
40. Boyle MH, Furlong W, Feeny D, et al. Reliability of the Health Utilities Index – Mark III used in the 1991 cycle 6 Canadian General Social Survey Health Questionnaire. Qual Life Res. 1995; 4: 249-257.
41. Boyle MH, Furlong W, Feeny D, et al. Reliability of the Health Utilities Index – Mark III used in the 1991 cycle 6 Canadian General Social Survey Health Questionnaire. Qual Life Res.1995; 4: 249-257.
42. Torrance GW, Feeny DH, Furlong WJ, Multiattribute utility function for a comprehensive health status classification system. Health Utilities Index Mark 2. Med Care. 1996; 34: 702-722.
43. EuroQol – a new facility for the measurement of health-related quality of life. The EuroQol Group. [see comment]. Health Policy 1990; 16: 199-208.
44. Brooks R. EuroQol: the current state of play. [Review] [32 refs]. Health Policy. 1996; 37: 53-72.
45. Bosch JL, Halpern EF, Gazelle GS. Comparison of preference-based utilities of the Short-Form 36 Health Survey and Health Utilities Index before and after treatment of patients with intermittent claudication. Med Decis Making. 2002; 22: 403-409.
46. Brazier JE, Roberts J. The estimation of a preference-based measure of health from the SF-12. Med Care. 2004; 42: 851-859.
47. Sherbourne DC, Unutzer J, Schoenbaum M, Duan N, Lenert LA, Sturm R, et al. Can utility-weighted health-related quality-of-life estimates capture health effects of quality improvement for depression? Med Care. 2001; 39: 1246-1259.

48. Gabriel SE, Kneeland TS, Melton LJ 3rd, et. al. Health-related quality of life in economic evaluations for osteoporosis: whose values should we use? Med Decis Making. 1999; 19: 141-148.
49. O'Shea K, Bale E, Murray P. Cost analysis of primary total hip replacement. Ir Med J. 2002; 95: 177-180.
50. Pickard AS, Wang Z, Walton SM, Lee TA. Are decisions using cost-utility analyses robust to choice of SF-36/SF-12 preference-based algorithm? Health Qual Life Outcomes. 2005; 3: 11.
51. Shelbourne KD, Rask BP. The Sequelae of Salvaged Nondegenerative Peripheral Vertical Medial Meniscus Tears with ANterior Cruciate Ligament Reconstruction. Arthroscopy. 2001; 17: 270-274.
52. Stratford PW, Goldsmith CH. Use of the standard error as a reliability index of interest: an applied example using elbow flexor strength data. Physical Therapy. 1997; 77: 745-750.
53. Liang MH. Longitudinal construct validity: establishment of clinical meaning in patient evaluative instruments.[see comment]. [Review] [37 refs]. Med Care. 2000; 38 (9 Suppl): 84-90.
54. Kirshner B, Guyatt G. A methodological framework for assessing health indices. J Chronic Dis. 1985; 38: 27-36.
55. Schünemann HJ, Puhan M, Goldstein R, et al Measurement properties and interpretability of the Chronic respiratory disease questionnaire (CRQ). COPD. 2005; 2: 81-89.
56. Schünemann HJ, Guyatt GH. Commentary – goodbye M(C)ID! Hello MID, where do you come from? Health Serv Res. 2005; 40: 593-597.
57. Jones PW. Interpreting thresholds for a clinically significant change in health status in asthma and COPD [see comment]. Eur Respr J. 2002; 19: 398-404.
58. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.
59. Guyatt GH, Berman LB, Townsend M, et al A measure of quality of life for clinical trials in chronic lung disease. Thorax. 1987; 42: 773-778.
60. Schünemann HJ, Puhan M, Goldstein R, et al. Measurement properties and interpretability of the Chronic respiratory disease questionnaire (CRQ). COPD. 2005; 2: 81-89.