HRQL measures in chronic pain trials

Gerold Stucki MD, MS, Alarcos Cieza, PhD

Gerold Stucki, MD, MS

Professor and Chair, Department of Physical Medicine and Rehabilitation

Director, WHO ICF Research Branch

University of Munich

81377 Munich, Germany

Phone: 0049 89 7095 4050

Fax: 0049 89 7095 8836

E-Mail: gerold.stucki@phys.med.uni-muenchen.de

Alarcos Cieza, PhD

Group leader, ICF Core Set development

WHO ICF Research Branch

Department of Physical Medicine and Rehabilitation

University of Munich

81377 Munich, Germany

Phone: 0049 89 7095 4050

Fax: 0049 89 7095 8836

E-Mail: alarcos.cieza@phys.med.uni-muenchen.de

Introduction

HRQL was selected to be one out of six core outcome dimensions to be considered in clinical trials on pain management at the first IMMPACT meeting. However, no recommendations regarding specific measures were made. It was decided that the process towards recommendations should be based on a comprehensive review.

HRQL measures are increasingly used to describe and evaluate functioning and health in clinical trials (Liang et al.,1990, Bombardier et al., 1995). The main purpose of HRQL measures in clinical trials is to describe the burden of disease of the population studied. Because HRQL measures focus on activities and participation which may be considered the components of functional health most relevant to patients and society and because activities and participation are relevant to all conditions they allow to compare effects across conditions, populations and

interventions. Generic health profiles including the SF-36 (Ware et al., 1992), the NHP (Hunt et al., 1985), or the WHODAS (Ustun et al., submitted) are best suited for this purpose. HRQL indices or utility measures such as the SF-6D and the EQ-5D used to measure the societal burden and the economic benefit of interventions across conditions, populations and settings.

HRQL measures may also provide an overview of the expected and unexpected changes in functioning and health associated with an intervention. While condition or symptom specific measures in chronic pain trials typically cover pain and aspects of physical function, HRQL measures cover symptoms and aspects of functioning and health not strongly associated with a condition, e.g. sleep functions and vitality. HRQL measures may also detect side-effects or complications of treatment which are not covered by symptom and condition specific measures. HRQL measures may also be more suited to discriminate patients with comorbidities. E.g. in a study with osteoarthritis the SF-36 was better at distinguishing those reporting a non-musculoskeletal comorbidity from those who did not than the condition specific instrument WOMAC (Brazier et al, 1999).

It is a matter of debate whether HRQL measures should also be useful and used as primary study endpoints. If this would be expected from a HRQL measure used in chronic pain trials, an excellent responsiveness for pain and physical function would be required. However, a number of studies that have compared the responsiveness of HRQL measures among themselves and with condition specific measures have generally found the latter to be more sensitive. It could be argued that for a HRQL measure to be useful for chronic pain trials, the comprehensive coverage of functioning and health is more important than its sensitivity to change for pain and physical function.

These considerations are important when deciding on the relative importance of the criteria for selecting specific HRQL measures to be recommended in chronic pain trials. The first concern when selecting and recommending a HRQL measure is truth, ore does the measure cover what it is supposed to measure (Boers et al., 1998, Liang et al., 1981). There are many ways how to establish truth. However, the first and most important concern is face and content validity: does the measure cover the spectrum of health problems which clinicians, clinician scientists, patients and regulatory agencies expect to be covered?

A problem when comparing the content of different HRQL measures is the varying use of concepts, scales and items. With the newly available International Classification of Functioning, Disability and Health or ICF (WHO 2001) it is now better possible to compare the representation of items and scales of the measures (Stucki et al. 2002, Cieza et al. 2002). The ICF was

approved in May 2001 by the world health assembly and is the successor of the ICIDH. Using established linkage rules (Cieza et al., 2002) items from specific instruments can be linked to the best corresponding ICF categories and the representation of the ICF components body functions and structures, activities and participation and contextual factors can be examined.

Based on the ICF it is also possible to define ICF Core Sets or short lists of categories to be measured in clinical practice or clinical studies. In a cooperation project between the WHO-ICF Research Branch at Munich and the CAS (Classification, Assessment and Surveys) team at WHO such ICF Core Sets for a number of chronic conditions including musculoskeletal conditions (osteoarthritis, rheumatoid arthritis, osteoporosis, low back pain), chronic pain, stroke, depression, obesity, diabetes, chronic obstructive lung disease, coronary artery disease and breast cancer are currently being developed (Stucki et al., 2002). The development process involves delphi procedures with experts, systematic reviews examining the use of outcome categories in clinical trials, primary data collection in patients with chronic pain using the ICF checklist (a 12-page short version of the full classification) and finally a consensus process with international experts. The pilot ICF Core Sets developed in this process will be tested for their content validity in different populations and regions, settings and interventions. It is important to note that the ICF and ICF Core Sets are not measures but do only provide a description of functioning and health. Based on the ICF Core Sets e.g. the ICF Core Set for chronic pain it is possible to examine whether and to what extent a HRQL or a condition specific measure covers the prototypical spectrum of problems encountered in patients with chronic pain.

The next consideration when selecting and recommending a HRQL measure is practicability (Boers et al., 1998). Since HRQL are regularly used together with condition and symptom specific measures and possibly utility measures, the respondent burden must be minimal. In practice and research applications, the time needed to complete a questionnaire should not exceed 10 to 15 minutes to ensure compliance (Stoll et al., in press). In general, self-administered questionnaires are more practical than instruments requiring a trained interviewer. However, in multicultural populations, in elderly patients or where literacy levels are variable, a standardized interview might be preferable.

The final consideration is cross-sectional and longitudinal discrimination or sensitivity to change. When comparing the discriminative ability of HRQL measures head-to-head comparisons ideally in a number of populations, settings and interventions including improvement and worsening are most informative

The objective of this review is provide information and arguments for the selection of HRQL measures to be used in clinical trials involving patients with chronic pain. The specific

aims are 1) to identify the HRQL measures that have been used in chronic pain trials, 2) to summarize the general characteristics of the identified measures and the use of the measure in chronic pain trials, 3) to examine the content validity of the instruments based on the common framework ICF and newly established linkage rules, 4) to examine the practicability of the HRQL instruments 5) to compare the discriminative ability of the instruments in head-to-head comparisons and 6) to make suggestions regarding currently possible recommendations, further research and an alternative approach based on the ICF to define what should be measured in each clinical trial.

Identification of HRQL measures used in chronic pain trials

We performed a non-systematic search using mesh and non mesh terms referring to: 1) pain conditions including chronic pain, fibromyalgia, low back pain, migraine, cancer pain; 2) terms used for health status measures including HRQL, 3) specific names of HRQL measures and 4) head-to-head comparisons. HRQL measures were also identified from systematic reviews performed in the context of the ICF Core Set project (Stucki G et al. 2002). Selected studies are presented after the description of each identified HRQL measure. Table 1 provides a summary of the described generic health-status measures. With the exception of the EQ-5D and the WHOQOL-BREF all instruments have been developed in anglo-saxon countries. All measures have been validated cross-culturally in a number of languages and the number of adaptations is constantly increasing (Anderson et al. 1996). Therefore, no detailed account regarding the currently available cross-cultural adaptations is provided in this report.

Description of the HRQL measures and use in pain trials

Spitzer's Quality of Life Index (QL-I) was developed for use by physicians in patients with cancer and chronically-ill patients. It is widely used by cancer specialists interested in the health status and quality of life of their patients before and after therapy. QL-I is one of the earliest scales to be designated a "quality-of-life" measurement (Spitzer et al., 1981). The QL-I consists of five items with three options for replies. The item responses are scored 0, 1 or 2, giving an overall score of 0 to 10. The scale can be summed into a single score, or each item can be calculated separately. Each item represents a different domain of life functioning. The respondents choose the item in terms of the statement's applicability to themselves: activity, daily living, health, support, and outlook. The scale also comprises a visual-analogue rating scale to rate quality of life.

The scoring of the QL-I is simple. The emphasis of the QL-I is on practicality; it is brief and easy to administer in a clinical setting. It has proved acceptable to clinicians and has been widely cited in the literature. The development of the QL-I was carefully undertaken, involving extensive consultation with patients and clinicians. The approach has served as a model for subsequent investigators (Wood-Dauphinee et al., 1991).

The instrument was valid for assessing HRQL in a study with cancer patients suffering from pain, in which two different care settings were compared (Ventafridda et al., 1989).

The **Sickness Impact Profile** (**SIP**) (Bergner et al., 1976) is a widely-used, general health-status instrument containing 136 items that have true or false answers. Scores use predetermined weights based on rater panel estimates of relative severity of dysfunction. The categories of ambulation, body care (including continence), and mobility are combined in a physical dimension, and four categories (emotional behavior, social interaction, alertness behavior, and communication) in a psychosocial dimension. The remaining categories are work, sleep and rest, eating, home management, recreation, and pastimes. Although the SIP can be used successfully in general practice setting, it is a long instrument that takes about 30 minutes to complete.

The SIP has been validated for use in chronic back pain (Deyo et al., 1983, Kopec et al. 1995). In a study of 107 patients with chronic low back pain the sensitivity of the SIP was considered sufficient to detect pre/posttreatment changes (Follick et al., 1985).

The **Quality of Well-Being Index** (**QWB**) and an earlier version, the Index of Well-Being, (Kaplan et al., 1976) is a health index that summarizes a person's current symptoms and disability in a single number that represents a judgment of the social undesirability of the problem. The QWB assess mobility, physical activity, and social activity, but not pain. A trained interviewer asks what the patient did because of illness during the last six days. The respondent is placed into one level for each scale, giving 45 possible combinations plus death, making 46 function levels. Scoring for particular functions is based on preference weights derived from the normal population. The QWB has the advantage of being a ratio scale with a true zero point which makes it possible to calculate quality-adjusted life-years in cost-benefit studies. The interview takes about 20 minutes.

Up to quite recently one of the major concerns about the QWB was that it must be administered by a trained interviewer. A self-administered form, which is known as the Quality of Well-Being Self-administered (QWB-SA) version 1.04 has therefore been developed. The QWB-SA has 58 symptoms/problems and 71 questions overall. The primary difference between both versions is the increased number of mental health items in the symptom/problem subscale of the QWB-SA. The QWB-SA can be completed in about 10 minutes. Initial studies have demonstrated the good psychometric properties of the QWB-SA (Kaplan et al., 1997). The QWB-SA can also be used to calculate quality-adjusted life-years (Sieber et al., 2000).

The QWB was valid, but not as sensitive to change as other measures in a clinical trial in patients with rheumatoid arthritis (Bombardier et al., 1986) and has been validated for patients with fibromyalgia (Kaplan et al., 2000).

The interviewer-administered QWB has been compared with the self-administered form (QWB-SA) for patients with migraine. Both versions appear to have sensitivity to migraine severity, and the ability to quantify an effect in multiple quality-of-life domains (Sieber et al., 2000).

The **WHOQOL**-BREF quality-of-life assessment was developed by the WHOQOL Group and represents an abbreviated version of the WHOQOL-100, which was developed in collaboration with 15 international field centers simultaneously in an attempt to develop a quality-of-life assessment that would be applicable cross-culturally (The WHOQOL Group, 1994). The WHOQOL-BREF contains a total of 26 questions. It includes one item for general quality of life, one item for health-related quality of life, and 24 items belonging to four domains (physical capacity, psychological, social relationships, and environment (The WHOQOL Group, 2001). In constructing the WHOQOL-BREF, data were taken from twenty field stations located in 18 countries. When comparing the psychometric characteristics of both instruments, the WHOQOL-BREF and the WHOQOL-100, the WHOQOL-BREF has proven to be a useful alternative to the WHOQOL-100 in evaluating improvement in quality of life following major therapeutic interventions for physical, psychological, and environmental domains. However, social aspects of quality of life are better addressed by the WHOQOL-100 (Carroll et al., 2000).

The instrument was valid and sensitive to change in a clinical trial with patients with low back pain (Muller et al, 2001).

The **WHO Disability Assessment Schedule (WHODAS II)** is a new 36-item instrument developed to assess activity limitations and participation restrictions in six domains (understanding and communicating, getting around, self care, getting along with people, life activities, and participation in society). The WHODAS II is different from the other measures of

health status in that it is based on an international classification system, i.e. The International Classification of Functioning Disability and Health or ICF (WHO, 2001). The WHODAS II is applicable across cultures; and treats all disorders at parity when determining the level of functioning (Ustun et al. submitted). In extensive testing, the WHODAS II has been shown to have a test-retest reliability ranging from 0.91-0.95 (within class correlations) across samples from the general population, from different geographical regions and with different physical disorders (e.g. drug, alcohol, and mental disorders) (Epping-Jordan et al., submitted).

At this point, the WHODAS II has not been used in patients with chronic pain. However, the WHODAS II has been used in patients with ankylosing spondylitis (AS) in which pain is a leading symptom and strongly associated with the burden of disease as measured with the WHODAS II (Tubergen et al., 2003). The WHODAS II was found to be a useful instrument for measuring disability in AS. Scores on the WHODAS II were significantly correlated with all AS-oriented questionnaires on disease activity, physical functioning, and quality of life. These findings were similar to the external criterion, the SF-36, thereby further validating the WHODASII.

The **Nottingham Health Profile (NHP)** (McDowell et al., 1978) has two parts. Part I contains 38 items grouped into six sections: physical abilities, pain, sleep, social isolation, emotional reactions, and energy level. Part II provides a brief indicator of handicap and contains seven items that record the effect of health problems on occupation, jobs around the house, personal relationships, social life, sex life, hobbies, and holidays. Part II is optional, as some items, like work and sex life, may not be applicable. Yes/no responses are used throughout. The NHP is self administered and takes approximately 10-15 minutes to complete (Hunt et al., 1985).

Since the NHP does not include items indicating positive well-being, community surveys typically find that around two thirds of the population record no problems on the NHP. Hunt and McKenna subsequently suggested that the NHP should not be used as a survey instrument (Hunt et al., 1992). Thus, the NHP is designed to represent severe problems, giving individuals with minor difficulties little room to improve. However, the NHP is one of the more frequently used measures, especially in Europe and continues to be used and tested and is considered to be of considerable usefulness as a clinical instrument. It has been applied to a variety of people in medical and non-medical settings. Its strengths include simplicity and broad coverage (Hunt et al., 1985).

The NHP together with the SF-36, it is one of the most popular existing health-status measures. A study aimed to evaluate and compare the psychometric properties of the SF-36

and the NHP in a group of patients with chronic peripheral neuropathic pain (PNP) shows that the SF-36 performed better on psychometric testing than did the NHP. However, it was also shown that the NHP contains dimensions such as sleep and more pain items which might be of particular importance in the PNP population (Meyer-Rosberg et al., 2001).

Similarly, in a head-to head comparison of the SF-36, the NHP and the EQ-5D in a survey of patients with migraine, lower though generally acceptable reliability estimates were found for the NHP scales as compared to the SF-36 (Essink-Bot et al.,1997). It was suggested that the differences in scale reliabilities noted between the SF-36 and the NHP may be due, in part, to the type of data generated by the two instruments (ie, the SF-36 yields polytomous data, the NHP dichotomous data). It was also argued that the NHP sacrifices some internal consistency for the sake of simplicity (Essink-Bot et al., 1997).

In a study with 3294 pain patients from 13 pain facilities in Germany both the SF-36 and the NPH discriminated well between headache and back pain patients and between several pain grades Gerbershagen (Gerbershagen et al., 2002).

The **Medical Outcome Study Short Form 36 (SF-36)** derives from a larger battery of questions administered in the Medical Outcomes Study (Ware et al.,1992). The SF-36 includes eight multi-item scales containing 2-10 items each and a single item to assess health transition. The scales cover the dimensions of physical health, mental health, social functioning, role functioning, general health, pain, and vitality. The SF-36 is the most widely-used general health-status instrument and is suitable for subjects aged 14 years and older and takes approximately ten minutes to complete.

Studies show excellent psychometric properties. As mentioned under the NHP, the reliability seems to be better than with the NHP (Essink-Bot et al., 1997, Meyer-Rosberg et al., 2001). The SF-36 has been used widely in studies with patients with chronic pain e.g. in musculoskeletal conditions. Some of them are described in the paragraph on head-to-head comparisons.

Two summary scales can also be obtained – the Physical Component Summary Score (**PCS**) and the Mental Component Summary Score (**MCS**) (Ware et al., 1992). Further evaluation of these summary scales provided the foundation for the construction of an instrument that is much shorter than the SF-36 (Ware et al., 1995). This new short form, the SF-12, uses 12 items from the SF-36 and demonstrates satisfactory reproducibility of the PCS and MCS with correlation coefficients of 0.93 to 0.97 on cross validation with the whole instrument. The new short form is likely to perform well enough for monitoring general

populations. It does not, however, allow scoring of individual SF-36 subscales, such as bodily pain or social functioning.

A preference based measure of health known as the **SF-6D** can be derived from the SF-36 for use in economic evaluation (Brazier et al., 2002). The SF-6D has six dimensions. Each dimension has between two and six ranked statements or levels, such as "You have no pain" through "You have pain that interferes with your normal work (both outside the home and housework) extremely". An SF-6D health state is defined by selecting one statement from each dimension, starting with physical functioning and ending with vitality. A total of 18,000 health states can be defined in this way.

Hollingworth et al. (2002) compared in a study with low back pain patients the practicality and construct validity of the SF-6D and five further preference scores derived from the SF-36 with directly elicited time trade off (TTO) and visual analogue scale (VAS) scores. All SF-36 derived preferences values including the one derived with the algorithm used for the SF-6D exhibited good practicality and construct validity in patients with LBP. However, all of them showed a minimum threshold implying a potential floor effect for severely ill patients. Directly elicited TTO values have less power to distinguish among patients with differing severity of LBP (Hollingworth et al., 2002).

The European Quality of Life instrument (**EQ-5D**) is a brief, self-administered, two-page questionnaire. The first page contains five items describing health status across five dimensions (mobility, self-care, usual activity, pain/ distress, and depression/anxiety) (the EQ-5D). The second page has a visual analogue rating scale on which the respondent marks an assessment of his/her overall health (The Euroqol Group, 1990). Each dimension is divided into three levels which, when taken together, define a total of 243 (3^5) unique health states. The responses to the five items in the EQ-5D can be scored using a utility-weighted algorithm (Williams, 1995), which has been recommended for use in economic evaluation. The EQ-5D, therefore, provides two single-index measures of health, the rating scale, and the EQ-5D index, ranging from 0-100 (Brook, 1997).

The EQ-5D has been shown to be sensitive in patients with LBP. The scales which discriminated best between patients who improved and those who deteriorated at 3 months were the EQ-5D and the SF-36 bodily pain and emotional role subscales (Suarez-Almazor et al., 2000).

Content comparison of the HRQL measures

The ICF as the common framework

The International Classification of Functioning, Disability and Health (ICF) (WHO, 2001) is a multipurpose classification belonging to the WHO family of international classifications. The ICF was designed to record and organize a wide range of information about health and health-related states in a standardized, common language, thereby facilitating communication about health and health care in various disciplines and scientific fields worldwide. It makes the comparison of data across countries, health-care disciplines, services, and time possible.

The ICF has two parts, each containing two separate components. Part 1 covers FUNCTIONING and DISABILITY and includes the components: Body Functions (b) and Structure (s) and Activities and Participation (d). Part 2 covers CONTEXTUAL FACTORS and includes the components: Environmental Factors (e) and Personal Factors.

In the ICF classification, the letters b, s, d, and e, which refer to the components of the classification, are followed by a numeric code starting with the chapter number (one digit), followed by the second level (two digits) and the third and fourth levels (one digit each). The component letter with the suffix of two, three, or four digits corresponds to the code of the so-called categories. Categories are the units of the ICF classification. Within each chapter, there are individual two-, three- or four-level categories. An example selected from the component Body Functions is presented in the following:

b2	Sensory functions and pain	(first level)
b280	Sensation of pain	(second level)
b2801	Pain in body part	(third level)
b28013	Pain in back	(fourth level).

Within each component, the categories are arranged in a stem/branch/leaf scheme. Consequently, a lower-level category shares the attributes of the higher-level categories to which it belongs, i.e., the use of a lower-level (more detailed level) category automatically implies that the higher-level category is applicable, but not the other way around.

At the end of each embedded set of third- or fourth-level categories and at the end of each chapter, there are "other specified" categories (uniquely identified by the final code 8). These categories allow the coding of aspects that are not included in any other specific categories. For example, at the end of the set of the fourth-level categories embedded in the third-level category b2801 Pain in body part, one finds the category b28018 Pain in body part, other specified.

Linkage of items to the ICF

Linking rules have been developed to link health-status measures to the ICF in a specific and precise manner (Cieza et al., 2002). On the basis of these linking rules, each item of an outcome measure should be linked to the most precise ICF category.

If one item encompasses different constructs, the information in each construct should be linked. For example, the item "I have some problems washing or dressing myself" of the EQ-5D the concepts "washing" and "dressing" have been linked to the ICF.

– The response options of an item are linked if they refer to additional constructs.

– If the information provided by the item is not sufficient for making a decision about the most appropriate ICF category, then this item should be linked "nd" (not definable).

– The abbreviation nd-gh (not definable-general health) is used for items concerning health in general, and the abbreviation nd-qol (not definable-quality of life) is used for items concerning the quality of life of patients in general.

– If an item is not contained in the ICF classification, then this item is assigned "nc" (not covered by the ICF).

Consensus between health professionals was used to decide which ICF category should be linked to each item/concept of the questionnaires. To resolve disagreements between the two health professionals concerning the selected categories, a third person trained in the linking rules was consulted. In a discussion led by the third person, the two health professionals that linked the item stated their pros and cons for the linking of the concept under consideration to a specific ICF category. Based on these statements, the third person made an informed decision.

Results of the linkage

Table 2 shows the difficulty of comparing the scales covered in the HRQL instruments. Table 3 shows an attempt to better compare the scales by using the structure with chapters and categories of the ICF. Table 4, 5, and 6 show a comparison of the items in the instruments using the ICF categories as a reference and ordered by component, i.e., body functions, activities and participation, and environmental factors. For practical reasons, only the instruments with less than 40 items are shown in tables 4 to 6. Thus, the SIP and the QWB are not presented.

As can be seen in Table 3, the ICF categories were generally linked to just one item/content from each health-status instrument. However, some instruments were linked to the same category more than once. This could indicate that the ICF did not differentiate in greater detail, and, therefore, several items or contents of items from a specific health-status instrument

had to be linked to the same ICF category. For example, in the SF-36, the ICF category b152 Emotional functions was chosen to link many different feelings: "feeling depressed or anxious", "emotional problems", "very nervous", "I felt so down in the dumps nothing could cheer me up", "I felt calm and peaceful", "I felt downhearted and blue", and "Have you been a happy person?". If there were different categories for different feelings, the named items would have been to different categories.

Representation of body functions:

As shown in Table 3 energy level (b1300) which may include positive and negative aspects, such as fatigue or energy, is represented in the QL-I, the WHOQOL-BREF, the NHP and the SF-36/SF-6D, but not in the WHODAS or the EQ-5D. Energy level seems relevant both from a generic or HRQL perspective and from the perspective of patients suffering from chronic pain.

The same holds true for sleep functions, which was only covered by the WHOQOL-BREF and the NHP and which was again selected into the ICF Core Set for chronic pain.

Emotional functions are relevant to patients with chronic pain and are covered by all examined measures. They are covered into more detail by the NHP and the SF-36/SF-6D than, for example by the WHODAS II.

Pain is covered more often and into more detail by the NHP than by the SF-36/SF-6D and in the EQ-5D. It is not addressed by the WHODAS II or in the QL-Index.

All these aspects of body function have been included in the ICF Core Set for chronic pain.

Representation of activities and participation:

Carrying out daily routine, again relevant to patients with chronic pain, is not contained in the QL-I, the NHP, or in the EQ-5D while it is covered by the others. Mobility is typically considered a relevant aspect of physical function in patients with chronic pain but is scarcely represented with the exception of the NHP and the SF-36. The EQ-5D only covers "walking" and the WHODAS only covers "walking long distances".

With respect to self-care, all instruments, with the exception of the WHOQOL-BREF and the EQ-5D, cover dressing. The QL-I and the WHODAS more broadly cover other aspects of self care.

At least one of the categories included in the chapter domestic life (d6) is covered in all instruments with the exception of the WHOQOL-BREF. However, the WHODAS, the NHP, and

the SF-36 cover this area more detailed than the other instruments. Housework, again highly relevant to patients with chronic pain is contained in the WHODAS, the NHP, and the SF-36.

With the exception of the SF-36/SF-6D, all instruments contain to some degree interpersonal interactions and relationships. The QL-I, WHOQOL-BREF, WHODAS, and NHP cover this domain very broadly. The corresponding items/concepts of the three instruments are linked at the chapter level to the ICF. While the WHODAS emphasizes relationships with friends and intimate and sexual relationships the EQ-5D focuses on family relationships.

It is important to emphasize that the social functioning scale of the SF-36 covers social activities with family, friends, neighbors, or groups, which, based on the ICF framework, are linked to community life, recreation, and leisure. However, the SF-36 does not cover interpersonal interaction and relationships at all. For patients with chronic pain, the issue of relationships most certainly is of importance. Accordingly, the ICF Core Set for chronic pain addresses family relationships (b760) and intimate relationships (b770).

All instruments cover aspects of work, but they are covered in more detail by the QL-I. With the exception of the QL-I and the SF-6D recreation and leisure (d920) is addressed in all instruments. Within the chapter community, social and civic life (d9) the SF-36 additionally covers the category socializing (d9250) and the NHP the category hobbies (d9204).

All these aspects of activities and participation have been included in the ICF Core Set for chronic pain.

The WHODAS covers aspects that may not be relevant to patients with chronic pain including understanding and communication e.g. "starting a conversation" and "sustaining a conversation".

Representation of environmental factors:

Environmental factors are covered very heterogeneously. Only the WHOQOL-BREF and the NHP covered the ICF category drugs (e1101), which represents an ICF category contained in the ICF Core Set for chronic pain.

Based on the ICF Core Set for chronic pain the support and relationship of the family and health professionals, the attitudes of immediate family members as well as social security services, systems and policies can be considered important aspects in patients with chronic pain. Accordingly, they have been selected into the ICF Core Set for chronic pain. However, these aspects are currently not covered by any of the instruments (with the exception of the QL-I that contains one item referring to the support and relationships of immediate family).

Representation of "physical function":

When considering physical functioning from a broad perspective, its key elements are: energy and drive functions, sleep functions, emotional functions, mobility, self-care, and housework. HRQL measures which comprehensively cover the experience of patients with chronic pain should arguably address these aspects. However, the QL-I and the SF-36 do not cover sleep functions. The WHOQOL-BREF covers neither self-care nor housework and covers mobility only very scarce. The WHODAS II and the EQ-5D do not cover energy level and sleep functions. The only instrument covering all these elements of physical function, as well as pain, is the NHP.

Feasibility

The major concern regarding practicability is the length of the measures. The time to complete for the SIP is considerably longer than for the other measures. Since the shorter measures including the SF-36, the NHP and the WHODAS appear to retain the psychometric properties of the longer instruments they are clearly to be preferred. All three instruments are self-administered but can also been used using an interview format. For the WHODAS there is also a proxy version which is advantageous e.g. in patients with stroke but which offers no advantage in patients with chronic pain. The preference measure SF-6D takes as long as the SF-36 since it is derived from the SF-36 using an algorithm. The EQ-5D takes even less time to complete then the SF-36. Therefore, both preference measures have a high feasibility with respect to time to administer.

In the interesting study by Essink-Bot et al (Essink-Bot et al., 1997) in patients with migraine and comparing the NHP, the SF-36 and EQ-5D, the feasibility of the measures (ie, the ease with which they can be completed by respondents) was examined indirectly by calculating rates of missing values. Importantly, the length of an instrument did not appear to have any direct bearing on the frequency of missing responses. For example, the highest rate of missing values was observed for the EQ-5D, one of the shortest of the instruments investigated. The NHP had the lowest missing value rate, lower than the proportions of missing data reported for the UK version of the instrument, This may reflect the simple, dichotomous response choices used consistently throughout the NHP, as well as the low demands placed on the respondents' reading skills through the use of short, uncomplicated sentences. The missing value rates for the SF-36 observed in the study were comparable to those reported for the UK and the US versions of the instrument (McHorney et al., 1994, Brazier et al., 1992)

In a study including the SF-36 and the EQ-5D in patients with OA of the knee, the high response and completion rate achieved, and the absence of any adverse comments from respondents, suggested that these instruments may be acceptable to this elderly population. (Brazier et al., 1999).

Discriminative ability in head-to-head comparisons

All measures described in the paragraph on the measures and summarized in table 1 are "psychometrically sound" and "well validated". According to published evidence and descriptions in leading books (Spilker, 1996) and reviews (Anderson et al. 1996) they have been shown to be valid, reliable and sensitive to change in a number of populations, conditions, settings and interventions. However, when comparing measures head-to-head comparison are needed. Unfortunately, in a non-systematic search we found only a limited number of head-to-head comparisons of generic measures in general and only few in patients with chronic pain or conditions with pain as a major complaint.

In the head-to-head comparison of the SF-36 with the NHP and the EQ-5D the SF-36 was best able to discriminate between groups formed on the basis of migraine status and work disability days (Essink-Bot ML 1997). These results were found to be in agreement with earlier reports of the ability of the SF-36 to discriminate between patients with minor versus major medical conditions and between patients with physical health versus psychiatric conditions (McHorney CA et al1993). It was rightly cautioned, that the discriminative ability based on cross-sectional analyses does not necessarily imply that an instrument will also be responsive to changes in health status over time.

In a study of patients with back pain (Suarez-Almazor ME 2000) including the SF-36 and the EQ5D, most SF-36 subscales appeared to have a floor effect for those patients who had deteriorated, with smaller effect sizes or change in the opposite direction. Despite being a much simpler measure, the VAS of the EQ-5D appeared to perform better than most SF-36 subscales in discriminating among those who improved and those who became worse. The EQ-VAS, however, was not very stable for patients reporting no change.

In a head to head comparison of the SF-36, the NHP and the SIP in workers with musculoskeletal conditions, the SF-36 was most responsive to the clinical change experienced by the subjects investigated, whether estimated according to the change in health status, or the sampling strategy (Beaton et al., 1997). The effect sizes were comparable to those reported by Katz in 1992 (Katz JN et al. 1992). However, like in the study by Katz (Katz 1992) the sample

size was not adequate to find a statistically significant difference between the questionnaires (Beaton DE et al. 1997).

In a study of patients with OA undergoing either clinical rheumatology care or knee replacement (Brazier JE 1999) the EQ-5D was able to discriminate on the basis of severity for patients with OA of the knee attending a rheumatology clinic, and was comparable in terms of responsiveness to the best-performing dimensions of the SF-36 in the knee replacement group. However, the EQ-5D was noticeably less responsive to change in the rheumatology clinic group than many of the dimensions of the SF-36. This was attributed to the fact that it is based on a more crude description of status in any given dimension, which makes it efficient for large changes, but less so for the more subtle and diverse changes experienced by the rheumatology clinic group. The EQ-5D rating scale was found to be even simpler, but its performance was inconsistent. It proved unable to distinguish between severity groups, and remarkably unresponsive to the changes following total knee replacement.

Discussion

General remarks and limitations

There are many considerations when deciding on specific HRQL measures to be used in chronic pain trails. Some of them have been addressed in this review. It is important to recognize that the suggestions made in this report are based on the limited evidence from relatively few studies in a limited number of subsets of patients and conditions undergoing a limited number of interventions. Also, our suggestions are based on relatively few true head-to-head comparisons which are most informative regarding the cross-sectional and longitudinal discrimination or sensitivity.

Recommendations of HRQL measures

The SF-36 and the NHP are certainly HRQL measures that can be recommended. Both instruments reasonably satisfy the discussed criteria including truth with a focus on content validity, practicability and cross-sectional and longitudinal discrimination.

While the SF-36 is somewhat more reliable and probably more sensitive to change, the NHP has the least missing values. Both measures comprehensively cover aspects of physical and emotional function. Pain is covered more broadly in the NHP and sleep functions are covered only by the NHP but not the SF-36. Since pain per definition is covered by specific measures there is no need for a HRQL measure used on chronic pain trials to cover pain more broadly. Also, the level of differentiation of the NHP with regard to sleep covering onset of sleep,

maintenance of sleep, and quality of sleep does not seem necessary for a HRQL measure. However, sleep may not be included in condition specific measures selected for chronic pain trials and sleep may therefore not be covered when using the SF-36. An advantage of the NHP may be its coverage of interpersonal interactions and relationships and some important environmental factors including drugs, support and relationships. Because of their relative brevity both measures can be administered in combination with condition and symptom specific measures and utility measures such as the EQ-5D.

While the WHODAS may cover many important aspects, excluding body functions, which could be measured with condition-specific measures, it may have the disadvantage in that it covers aspects like problem solving and communication that may be not be relevant to patients with chronic pain. If an instrument includes aspects that are not relevant to a patient population, respondents may refuse to complete the questionnaire, leading to missing values. Also, the WHODASII covers some aspects of physical function only scarcely.

The SIP and the QWB hardly fulfill the criterium of feasibility to be used in all clinical trials on pain management. However, the SIP may be useful in patients with a number of comorbid conditions, in hospitalized patients or patients in nursing homes. In these settings the SIP may be preferable to the SF-36 which may have a floor effect in these settings.

The QL-I is of questionable usefulness in patients with chronic pain mainly because of its questions using many examples summarizing a broader concept. This is does not seem ideal for a health status profile.

Despite its brevity, the EQ-5D may not be recommended to be used as the only HRQL measure in clinical trials. It does not give the broad picture available from a profile measure such as the SF-36 and therefore lacks content validity. It is also not as responsive as the other measures particularly for smaller changes. Its feasibility with respect to missing values has also been questioned.

Both the EQ-5D or the SF-6D derived from the SF-36 are options for economic evaluations. In the case of the EQ-5D in combination with a profile measure and in the case of the SF-6D obviously derived from the SF-36. However, at this point there is only limited evidence regarding the usefulness of the EQ-5D and no evidence regarding the performance of the SF-6D in pain patients. One of the advantages of the SF-6D over the EQ-5D could be a result of the much larger size of its descriptive system and, hence, a possibly greater degree of sensitivity. Nevertheless, this must be weighed against the inconsistencies between the coefficients at the upper levels of some SF-6D dimension. The sensitivity of the SF-6D needs to

be compared to other preference-based measures before drawing any conclusion on this point (Brazier J et al., 2002).

ICF Core Sets of functioning, disability and health

Health status measures and new versions of existing measures seem constantly to emerge and are constantly evolving with respect to their contents and psychometric properties. Therefore, any recommendation regarding a specific instrument is likely to be soon outdated. To avoid this problem at least to some extent, it would be preferable to first define **what should be measured** and only then to recommend **how to measure** or which measure to use.

This approach has been successfully taken in rheumatology by the OMERACT group (Boers et al. 1998). If enough care is being taken to define what should be measured, it should form the basis for a solid and stable recommendation being adhered to for many years. Based on the solid set of recommended categories to be measured, the constantly evolving best measurement options can be tested against this set and may then be recommended.

When defining a Core Set of categories of functioning and health for chronic pain trials, a common framework is needed and available with the new ICF. Based on the experience from the ICF Core Set project the prototypical spectrum of problems encountered in patients with chronic diseases is covered in the ICF. Also, the linkage of the items of measures to the ICF is feasible. Therefore it would be possible to define an ICF based Core Set which can be used to test the content validity of conditon specific and HRQL measures. More practically the pilot ICF Core Sets for chronic pain developed by the ICF research branch at Munich and the CAS (Classification, Assessment and Surveys) team at WHO could be refined in a delphi and or consensus process involving the IMMPACT participants. An important additional source of information is the representation of the currently recommended condition and pain specific pain measures and the recommended HRQL measures. As we have shown in this review, based on the ICF Core Sets e.g. the ICF Core Set for chronic pain it is possible to examine whether and to what extent HRQL measures or condition specific measures cover the prototypical spectrum of problems encountered in patients with chronic pain.

It would also be possible to operationalize the ICF Core Sets to be used as common and comparable short standard measure in clinical trials together with more refined and specific measures best suited to the study question, population and intervention. Because comparisons can be made based on such a short set of items all currently and upcoming suitable measures could be recommended. Such an approach has been suggested by Deyo and colleagues who came up with a 6 item standardized core set of questions and questionnaires to be used in all

studies for back pain (Deyo RA et al. 1998). They suggested the use of this set together with an expanded, more precise battery of measures.

Future research to close the gap

It would be worthwhile to also examine the overlap of the HRQL measures with the recommended condition and symptom specific measures. Such a comparison is possible based on the demonstrated approach with linkage of the items to the ICF. Most currently used condition specific measures for musculoskeletal conditions and chronic pain have already been linked and manuscripts about this linkage have been submitted. Ideally, there would be only a minimal overlap of condition and symptom specific measures to minimize the respondent burden.

If a recommendation of „what to measure" as the the basis for the recommendation „how to measure" is considered, it is suggested to review, test and possibly modify the developed pilot version of the ICF Core Set for chronic pain. For example, it may be reasonable to reduce the number of recommended categories to a true „minimum" ICF Core Set. Based on such a „minimum" ICF Core Set the presented results regarding the content validity can be verified.

If the operationalization of the ICF Core Set categories to be used independently is considered an option, such an operationalized ICF Core Set should be tested alongside the other measures in a head-to-head comparison.

Finally, a head-to-head comparison of the performance of the SF-36, the NHP, the WHODASII and the EQ-5D in head-to-head comparisons in a number of populations (including the elderly), pain conditions, settings and interventions would be most informative. The comparisons should include an analysis of the cross-sectional and longitudinal discrimination or sensitivity and the feasibility including the examination of missing values and acceptance with patients (e.g. for the WHODAS scales understanding and communication). The sensitivity to change should be studied for both improvement and worsening since studies in chronic pain patients may evaluate the effectiveness to prevent pain exacerbations. The sensitivity to change should be studied not only for large but also for smaller effect sizes. The comparison should be used to define the smallest detectable and minimal clinically important differences relevant to sample size calculations.

References

1. Anderson RT, Aaronson NK, Bullinger M, McBee WL. A Review of the Progress Towards Developing Health-Related Quality-of-Life Instruments for International Clinical Studies and Outcomes Research. Pharmacoeconomics 1996; 10(4): 336-55

2. Beaton DE, Hogg-Johnson S, Bombardier C. Evaluating Changes in Health Status: Reliability and Responsiveness of Five Generic Health Status Measures in Workers with Musculoskeletal Disorders. J Clin Epidemiol 1997; 50(1): 79-93

3. Bergner M, Babbitt RA, Pollard WE, Martin DP, Gilson BS. The Sickness Impact Profile: validation of a health status measure. Medical Care 1976; 14: 57–67

4. Boers M, Brooks P, Strand C V ,Tugwell P. The OMERACT filter for outcome measures in rheumatology (editorial). Journal of Rheumatology 1998, 25:198-199

5. Bombardier C, Melfi CA, Paul J, Green R, Hawker G, Wright J ei al. Comparison of a generic and a disease-specific measure of pain and physical function after knee replacement surgery. Med Care 1995; 33 (suppl.): AS131—44

6. Bombardier C, Ware J, Russell IJ et al. Auranofin therapy and quality of life in patients with rheumatoid arthritis: results of a multicenter trial. American Journal of Medicine 1986; 81: 565–78.

7. Brazier J.E, Harper R, Munro J, Walters S.J., Snaith M.L. School for Health and Related Research and Institute for Bone and Joint Medicine, Medical School, University of Sheffield, Sheffield, UK. Generic and condition-specific outcome measures for people with osteoarthritis of the knee. Rheumatology 1999; 38:870-7

8. Brazier JE, Harper R, Jones NMB, et al. Validating the SF-36 health survey questionnaire: New outcome measure for primary care. BMJ 1992; 305:160-4

9. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ 2002; 21:271-92

10. Brook R with the EuroQol group. EuroQol: the current state of play. Health Policy 1996; 37:53-72; Dolan P. Modeling valuations for EuroQol health states. Med Care 1997; 35:1095-108

11. Carroll RE, Schmith K, Couston M, Cossar JA, Hayes PC. A comparison of the WHOQOL-100 and the WHOQOL-BREF in detecting change in quality of life following liver transplantation. Quality of Life Research 2000; 9:121-124

12. Cieza A, Brockow T, Ewert T, Amman E, Kollerits B, Chatterji S, Bedirhan Üstün B, Stucki G. Linking Health-Status measurements to the International Classification of Functioning, Disability and Health. J Rehabil Med 2002; 34: 1-6

13. Deyo, RA and Diehl AK. Measuring physical and psychosocial function in patients with low back pain. Spine 1983; 8: 635-42

14. Deyo RA, Battie M, Beurskens AJ, Bombardier C, Croft P, Koes B, Malmivaara A, Roland M, Von Korff M, Waddell G. Outcome Measures for Low Back Pain Research. A Proposal for Standardized Use. Spine 1998; 23:2003-13

15. Epping-Jordan JE, Prieto L, Doyle P, Chatterji S, Ustun B. Reliability of the World Health Organization disability assessment schedule 11 (WHODAS II): subgroup analysis. Submitted

16. Essink-Bot ML, Krabbe PF, Bonsel GJ, Aaronson NK. An empirical comparison of four generic health status measures. The Nottingham Health Profile, the Medical Outcomes Study 36-item Short-Form Health Survey, the COOP/WONCA charts, and the EuroQol instrument. Med Care 1997; 35(5): 522-37

17. Fahrenberg J, Myrtek M, Wilk D, Kreutel K. Multidimensionale Erfassung der Lebenszufriedenheit: Eine Untersuchung an Herz-Kreislauf-Patienten. Psychotherapie, Psychosomatik, Medizinische Psychologie 1986: 347-54

18. Fairbank JC, Couper J, Davies JB, O'Brien JP. The Oswestry low back pain disability questionnaire. Physiotherapy 1980; 66(8): 271-73

19. Feeny D, Furlong W, Boyle M, Torrance GW. Multi-attribute health status classification systems. Health Utilities Index. Pharmacoeconomics 1995; 7(6):490-502

20. Follick, M.J., Smith, T.W. and Ahern, D.K., 1985. , The Sickness Impact Profile: A global measure of disability in chronic low back pain. Pain 21: 67-76

21. Gerbershagen HU, Lindena G, Korb J, Kramer S. Health-related quality of life in patients with chronic pain. Schmerz 2002; 16(4): 271-84

22. Hollingworth W, Deyo RA, Sullivan SD, Emerson SS, Gray DT, Jarvik JG. The practicality and validity of directly elicited and SF-36 derived health state preferences in patients with low back pain. Health Econ 2002; 11(1): 71-85

23. Hunt SM, McEwen J, McKenna SP. Measuring health status: a new tool for clinicians and epidemiologists. J R Coll Gen Pract 1985; 35(273): 185-8

24. Hunt SM, McKenna SP. Validating the SF-36. BMJ 1992, 12:305:645

25. Kaplan RM, Bush JW, Berry CC. Health status: types of validity for an index of well-being. Health Services Research 1976; 11: 478–507

26. Kaplan RM, Schmidt SM, Cronan TA. Quality of well being in patients with fibromyalgia. J Rheumatol 2000; 27:785-9

27. Kaplan RM, Sieber WJ, Ganiats TG. The Quality of Well-being Scale: comparison of the interviewer-administered version with a self-administered questionnaire. Psychol Health 1997; 12: 783-91

28. Katz JN, Larson MG, Phillips CB, Fossel AH, Liang MH. Comparative measurement sensitivity of short and longer health status instruments. Med Care 1992; 30(10): 917-25.

29. Kopec JA, Esdaile JM. Functional disability scales for back pain. Spine 1995 1; 20(17):1943-49

30. Liang MH and Jette AM. Measuring functional ability in chronic arthritis: a critical review. Arthritis and Rheumatism 1981; 24: 80–6

31. Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. Med Care 1990; 28:632-42

32. McDowell IM, Martini CJM, Waugh W. A method for self-assessment of disability before and after hip replacement operations. British Medical Journal 1978; 2: 857–9

33. McHorney CA, Ware JE, Lu JFR, 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(1): 40-66

34. McHorney CA, Ware JE, 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(3):247-63

35. Osterhaus JT, Townsend RJ, Gandek B, Ware JE. Measuring the functional status and well-being of patients with migraine. Headache 1994; 34(6):337-43

36. Meyer-Rosberg K, Burckhardt CS, Huizar K, Kvarnstrom A, Nordfors LO, Kristofferson A. A comparison of the SF-36 and Nottingham Health Profile in patients with chronic neuropathic pain. Eur J Pain 2001; 5(4): 391-403

37. Muller K, Schwesig R, Leuchte S, Riede D. Coordinative treatment and quality of life - a randomised trial of nurses with back pain. Gesundheitswesen 2001; 63(10): 609-18

38. Saxena S, Carlson D. Billington R, Orley J. The WHO quality of life assessment instrument (WHOQOL-Bref): The importance of its items for cross-cultural research. Qualty of Life Research 10 2001, 10: 711-21

39. Sieber WJ, David KM, Adams JE, Kaplan RM, Ganiats TG. Assessing the impact of migraine on health-related quality of life: An additional use of the quality of well-being scale-self-administered. Headache 2000; 40(8): 662-71

40. Spilker B, Quality of Life and Pharmacoeconomics in Clinical trials, 2nd edition, Lippincott-Raven Publishers, 1996

41. Spitzer WO, Dobson AJ, Hall J, Chesterman E, Levi J, Shepherd R, Battista RN, Catchlove BR. Measuring the quality of life of cancer patients: a concise QL-index for use by physicians. J Chronic Dis 1981; 34(12):585-97

42. Stoll T, G. Stucki. Outcome Assessment in Rheumatology. Chapter 2.1. In: Oxford Textbook of Rheumatology. 3rd edn. Oxford University Press, in press

43. Streiner DL, Norman GR. Health measurement scales: A practical guide to their development and use. Oxford: Oxford Medical Publications, 1989

44. Stucki G, Ewert T, Cieza A. Value and application of the ICF in rehabilitation medicine. Disability and Rehabilitation 2002; 24(17): 932-8

45. Suarez-Almazor ME, Kendall C, Johnson JA, Skeith K, Vincent D. Use of health status measures in patients with low back pain in clinical settings. Comparison of specific, generic and preference-based instruments. Rheumatology 2000; 39(7): 783-90

46. The Euroqol Group. Euroqol-a facility for the measurement of health-related quality of life. Health Policy 1990; 16:199-208

47. The WHOQOL Group, 1994. The development of the World Health Organization quality of life assessment instrument (the WHOQOL). In J. Orley and W: Kuyken (Eds.) Quality of Life Assessment: International Perspectives. Heidelber: Springer Verlag

48. The WHOQOL Group. Development of the World Health Organization WHOQOL-BREF quality of life assessment. Psychol Med 1998; 28(3): 551-58

49. Tubergen A, R Landewe, L Heuft-Dorenbosch, A Spoorenberg, D van der Heiide, H van der Tempel, S van der Linden. Assessment of disability with the World Health Organisation Disability Assessment Schedule II in patients with ankylosing spondylitis. Ann Rheum Dis 2003; 62:140-45

50. Ustun TB, Chatterji S, Rehm J, Kennedy C, Prieto L, Epping-Jordan JE, et al. World Health Organization Disability Assessment Schedule 11 (WHODAS II): development, psychometric testing, and applications. Submitted

51. Ventafridda V, De Conno F, Vigano A, Ripamonti C, Gallucci M, Gamba A. Comparison of home and hospital care of advanced cancer patients. Tumori 1989 31; 75(6):619-25

52. Ware JE and Sherbourne CD. The MOS 36-item short-form health survey (SF-36). A. Conceptual framework and item selection. Medical Care 1992; 30: 473–83

53. Ware JE, Kosinski M, Keller SD. SF-12. How to score the SF-12 physical and mental health summary scales. The Health Institute, New England Medical Center, Boston, 1995

54. Williams A. The measurement and valuation of health: a chronicle. Discussion Paper 136. Centre for Health Economics, York Health Economics Consortium, NHS Centre for Reviews and Dissemination. York: University of York, 1995

55. Wood-Dauphinee S, William Ji. The Spitzer Quality of Life Index_ its performance as a measure. In: Osoba D (Ed.) Effect of cancer on quality of life. Boca Raton, Flirida: CRC Press, 1991: 169-84

56. World Health Organization. International Classification of Functioning, Disability and Health: ICF. Geneva: WHO, 2001