Health and Quality of Life Outcomes

Commentary

A brief guide to the selection of quality of life instrument

Michael E Hyland*

Address: Department of Psychology, University of Plymouth, Plymouth PL4 8AA

Email: Michael E Hyland* - M.Hyland@plymouth.ac.uk

* Corresponding author

Published: 03 July 2003

Health and Quality of Life Outcomes 2003, **1**:24

Received: 19 June 2003
Accepted: 03 July 2003

This article is available from: http://www.hqlo.com/content/1/1/24

Abstract

There are numerous quality of life (QOL) scales. Because QOL experts are often partial to their own scales, researchers need to be able to select scales for themselves. Scales best suited for longitudinal purposes (clinical trials and audit) have different properties to those suited for cross-sectional studies (population and correlational studies and clinical use). The reason and logic of these differences is explained. For longitudinal use, researchers need to consider the relationship between item set, population and treatment; scales can be short, floor and ceiling effects must be avoided, and there should be extended response options. For cross-sectional use scales should have a wide range of items, should be longer, and there are no adverse floor and ceiling effects, and response options can be simpler to allow a larger set of items.

Introduction

There are numerous quality of life (QOL) instruments available to researchers, but little guidance for selection between them [1]. This choice is made more difficult by the fact that experts are frequently partial to their own scales [2]. Although researchers may feel daunted by the need to choose for themselves, this task is surprisingly straightforward once the rules underlying QOL scale performance are understood. The purpose of this paper is to explain those rules.

Purpose of scale

The optimum properties of a QOL scale are determined by the purpose for which it is put, in the same way that the selection of a surgical instrument is determined by its use. There is no such thing as a 'best scale' in an absolute sense, only scales best suited to a particular purpose. Several years ago, Guyatt, Kirshner and Jaeschke [3] suggested that QOL scales can be validated in terms of two purposes: longitudinal comparison and cross-sectional comparison. Within each of these two types of use, it is possible to make a further division based on whether the scale is to be used for research purpose (i.e., infrequently and for a specific research project) or whether the scale is to be used in clinical practice (i.e., is used frequently and without the benefits of research funding).

This paper uses these two classifications (longitudinal versus cross-sectional and research versus clinical) to examine the properties of scales which are most suited for the following purposes

1. Longitudinal comparison in randomised clinical trials (RCTs).

2. Longitudinal comparison where the quality of provision of treatment is being audited by health managers

3. Cross-sectional comparison for statistical purposes.

4. Cross-sectional comparison for clinical purposes.

QOL scales can be used for other purposes, for example for resource allocation between different diseases, but this purpose is not covered here.

Longitudinal comparison in RCTs

The purpose of a QOL scale in a RCT is to be able to detect important changes in the patient's QOL. A good QOL scale for use in RCTs is therefore one that is good at detecting change. A QOL scale is nothing more than a set of items. Those items can be likened to a shopping basket of experiences selected from the supermarket of possible life experiences. A good longitudinal QOL scale for RCT use is one containing items measuring all important aspects of QOL for the population under study and most of these items are sensitive to change that would be expected from the treatments studied. Sensitivity will be determined by three factors: the items themselves, the treatment and the population.

If the items of a QOL scale are analysed individually in a clinical trial, it invariably happens that items vary in the extent to which they demonstrate improvement, with some items actually showing a small deterioration. By adopting a cut off point, the item set can be divided for convenience into two groups – the 'shifting items' which demonstrate improvement beyond a criterion and the 'non-shifting items' that show no change or a deterioration in QOL. Sensitivity to change is function of the proportion of shifting and non-shifting items. Thus, a good longitudinal scale is a scale that has the many shifting items.

Whether an item is shifting or not depends on several factors. The most obvious is purely statistical. If a patient does not report a problem with a particular item, then that patient can not improve on that item. This fact is particularly important if the majority of patients have mild QOL impairment. Items where QOL deficits are reported only by patients with severe morbidity are unlikely to shift in a population with mild morbidity – there are many cases where failure to achieve QOL improvement is because a criterion of sufficient QOL deficit at baseline has not been employed. Patients with few recorded problems seldom provide evidence of improvement. On the other hand, if an item is experienced as a problem by all patients because it is a characteristic of the disease, then a treatment that cannot achieve a cure is unlikely to remove that problem. Items exhibiting floor or ceiling effects are poor shifters – good shifting items tend to be midrange in terms of frequency of the reported problem where at least half of the patients note impairment in their baseline response to the item. Note that floor and ceiling effects are population dependent. An item may exhibit a floor effect with a mild sample of patients, because few patients report the problem, but is a useful item in a severe sample of patients.

Floor and ceilings can be inferred in part from the distribution of responses to an item within a specified population, but content also makes a difference. If an item is irrelevant to members of a population, then there is little chance it will show improvement in a longitudinal study. For example, research in asthma [4] suggests that items relating to sport shift more in younger populations, whereas those relating to mobility problems shift more in older populations. This is because older people are more likely to find sports items irrelevant whereas younger asthmatics seldom have mobility problems. The relevance of an item can be highly population specific. If, for instance, a patient never does gardening because he lives in a high rise tower block, then an item on whether his disease adversely affects gardening is unlikely to shift after any treatment. Similarly items like shovelling snow in the backyard are not going to shift in populations living in temperate climates.

One way of improving the relevance of items to a population is to individualise either items or the whole scale to the individual. For example, patients can be asked to nominate 5 activities affected by their disease and then use these individualised items for purposes of rating [5]. Individual quality of life scales often have good longitudinal properties, though individualisation can create problems if when the scale is used for crossectional purposes.

Item relevance becomes particularly important when comparing disease specific with generic scales. Suppose a generic scale containing items on pain sensation is used in an asthma clinical trial. The pain items will not shift, but they would be expected to shift if the same scale is used in an arthritis clinical trial. Of course, there will (almost) always be some items in a generic scale that will shift irrespective of the disease, but the proportion of shifting items will typically (but not invariably) be less than in a disease specific scale. Consequently there is a general rule that generic scales are less sensitive to change than are disease specific scales [6] – and which goes some way to explaining the explosion in the number of disease specific scales created over the last 10 years. Generic scales do have another use in clinical trials – their broader spread of items makes them more suited to detecting iatrogenic effects.

An item may be capable of shifting, but not shift because the treatment does not create that kind of improvement. For example a treatment for irritable bowel syndrome (IBS) which reduces diarrhoea will not affect items in the scale that relate to problems arising from constipation (e.g., general malaise and bowel discomfort). Items shift not only as a function of the population, but also as a function of the treatment used. The selection of a QOL scale which is likely to have a good proportion of shifting

Table 1: Properties of QOL scales used for longitudinal and cross-sectional comparison

Likely to be a good Longitudinal questionnaire	Likely to be a good Cross-sectional questionnaire
Short (typically 1 – 40 items)	Long (typically 20 – 100 items)
Multi-response (e.g., 7-response) format item	Simple (e.g., binary or tertiary) response format
Limited severity range: Items describe problems common to most patients, or only in the population to be studied	Items cover the whole severity range of QOL deficit
No items showing floor or ceiling effects (i.e., items where >70% respond at either end of the scale) within target population	Items with floor and ceiling effects should be included
Items must be relevant to most patients	Items need not be relevant to all patients
Items irrelevant to the disease should not be included (unless the scale is used to test for iatrogenic change)	Items irrelevant to the disease should not be included (unless the scale is to be sensitive to co-morbidity)

items therefore involves trying to match between a population and item set, taking into account the kind of improvement that is possible from the treatment.

Because of the need in longitudinally sensitive scales to include only items that are potentially relevant to the population, the item set can be relatively short. Good longitudinal scales are typically not more than about 30–40 items. However, much fewer items can be used, and the shortest scale is the one item global QOL scale [7]. Such short or one item scales can be very sensitive to treatment, but their downside is that they lack the ability to inform *how* QOL is improving [8].

Whether or not an item is capable of shifting is affected by one other factor: the response scale format. Patients may be aware of slight improvement but not substantial improvement. Response scales of up to about 7 points (e.g., the Likert scale format) tend to be more sensitive to change than binary response fomats. A potentially good longitudinal QOL scale is therefore, likely to be quite short, describing commonly experienced problems relevant to the population to be investigated and have a multi-response format. The need for a sensitive multi-response format is particularly relevant where the item number is low or where there is just one item (i.e., the global scale). Single item global scales typically ask patients to choose between 10 and 100 levels of QOL.

Longitudinal comparison for purposes of audit purpose

It is often useful to have a scale that can assess to what extent a particular treatment is successful. Such routine audit allows comparison between different treatment centres as well demonstrating to cost-conscious administrators that the treatment is beneficial. When used as an everyday clinical tool for audit purposes, the QOL scale needs to be short. As indicated above, short scales can be very sensitive to change. In selecting an audit scale, it is important that the scale is sensitive to the particular treatment which is being audited. A good audit tool is not only appropriate for the disease and population, it is also appropriate for the treatment. For example, the short form of the Breathing Problems Questionnaire [9] was designed as a QOL audit scale in pulmonary rehabilitation and consists of items specifically selected on the basis that they shift after rehabilitation. Treatment specific scales would not be appropriate for RCTs. For example, an IBS scale which measured only the QOL deficits of the diarrhoea component of IBS would not be a good scale in an RCT, because this captures only part of the total picture of QOL. When evaluating between to different drugs it is necessary to know the total picture in terms of QOL change. However, when a treatment is audited, then it is appropriate to focus on specifically those aspect of QOL that the treatment can improve.

Cross-sectional comparison for statistical purposes

A scale used for cross-sectional studies needs to provide good discrimination between the severity of QOL deficit between patients. Imagine a QOL scale comprising only one item with three response options. Use of this item enables the researcher to categorise patients on only three levels. Add on another item, and the ability to discriminate between different categories of patients is increased. As more and more items are introduced into the scale, the ability to discriminate between patients becomes yet greater. This example illustrates a general rule: the ability to make fine-grained discriminations between the QOL of different patients increases as the number of items increases.

It is necessary, in the case of longitudinal sensitivity to avoid floor and ceiling effects, but quite the reverse occurs for a scale designed for cross-sectional sensitivity. If a scale is limited to items which show a QOL deficit in the majority of patients, then these items will not be able to discriminate between patients at the severe end of the scale, because at the severe end, all patients will consistently endorse these items. Discrimination occurs only if some of the severe patients, the very severe patients, endorse the

item and the not-so-very-severe ones do not. The same logic applies at the mild end of the continuum, if all patients at the mild end report a problem then there will be no discrimination between mild patients. This example illustrates another rule: a good cross-sectional scale should discriminate between patients over the whole of the severity range, and therefore will include items relevant to all levels of severity. In such a scale, some items will be endorsed by most patients, and some items will be endorsed by very few patients. There are no adverse floor and ceiling effects.

The need to discriminate across the full severity range is particularly important where the scale is used for correlational analysis. The size of a correlation depends on the degree of variation of items in either measure, and if range is attenuated in the questionnaire due to failure to discriminate, then correlations will be reduced. For example, if respiratory function correlates poorly with QOL in the case of severe chronic obstructive pulmonary disease (COPD) patients, it may be that this is caused by lack of variation in that population of severe patients – i.e., they all endorse almost all items as being problematic

Generic QOL scales are sometimes used in cross-sectional studies. Suppose that a generic scale including several pain items is used to assess QOL in chronic obstructive pulmonary disease (COPD) patients. Because they are elderly, it is likely that many patients will have co-morbidity that creates pain. However, this co-morbidity will be due to musculo-skeletal problems, not to poor lung function. The pain items will therefore create variation in the overall score which will not correlate with lung function. Thus, the generic scale would be a poor choice if the aim is to correlate QOL with lung function. On the other hand, the inclusion of the pain items provides a better characterisation of the total impact of disease in this population, so if the aim is compare the overall deficit of a patient population then a generic scale would be better. Generic scales are, like disease specific scales, a conventionally defined shopping basket of deficits from the total supermarket of life experiences. If a generic QOL scale has many pain items but few disturbance sleep items, then other things being equal, asthma will appear to have poorer QOL deficit compared with arthritis. On the other hand, if the generic scale has many sleep disturbance items but no pain items, then asthma patients will, on average, appear to have better QOL than arthritic patients. As with longitudinal comparison, the results are always scale specific.

In the case of scales requiring longitudinal sensitivity it is helpful to have response options that are sensitive to change, for example, by having up to 7 response options. However, the time taken by a patient to complete a 7 response item is longer than that needed for a binary response item – so that a scale of 20 7-response items will take far longer to complete than 20 binary-response. There is therefore a trade-off between the number of items a patient can reasonably be expected to complete and response format. Consequently, because a good cross-sectional scale needs to have a large number of items it may be appropriate to use a simpler, binary response format. The cost of increased number of items is paid for by the simpler response format.

Cross-sectional comparison for clinical purposes

It is sometimes useful to have a QOL scale that provides an overall picture of the patient's QOL and which can then be used for clinical decision making. The characteristics of a good scale for clinical cross-sectional comparison are similar to that for cross-sectional comparison for statistical reasons, but with one important difference. The content of the items in the scale need to be selected on the basis that they inform clinical decision making. For example, the inconvenience or cost of medicine can have an impact on a patient's QOL and this may be particularly relevant for patient management. Other than selecting for the clinical purpose, the general principle of cross-sectional comparison remains, i.e., a number of items are needed that provide discrimination between the mild and most severe patients – or at least that provide discrimination within the population that is clinically relevant. However, because of the time and cost constraints of clinical practice, the scale may need to be shorter than one which can be used in research settings. Where co-morbidity is expected, then a generic scale may be preferable as it provides a more holistic picture of the patient's QOL deficits, but the choice between generic and disease specific is decided by judgements about the clinical usefulness of different scales

Psychometric considerations

Authors of QOL scales normally provide psychometric data, of varying kinds. Factor analysis or item analysis is used to demonstrate the unidimensionality of a scale or subscale (i.e., that the items of the scale can be meaningfully added to form a single score). The reliability of the scale is shown through test-retest correlation or internal consistency (alpha coefficient), and the scale is correlated with validating criteria such as other QOL tests or morbidity. Although all QOL questionnaires should satisfy certain minimum criteria, they do not form an essential part of choosing between scales. For example, a scale that is more unidimensional in the sense of having higher inter-item correlations (or higher factor loadings) is not necessarily better for any of the three purposes above. Reliability is important to the extent that a correlation with test can never by higher than its retest reliability, however, most scales have acceptable levels of reliability above 0.7,

and the majority are above 0.9. I have never come across a QOL scale that is incapable of demonstrating validating correlations with other QOL scales, such as the SF-36. The reason is simply that all self-report measures are strongly correlated with the personality trait of negative affectivity (e.g., neuroticism, depression, anxiety), and so QOL scales inter-correlate amongst themselves. In sum, where scales have adequate psychometric properties, then this should not be an important factor when selecting between them, but it may have an influence, for example, where two scales are very similar but one is more reliable than the other. Where scales do not have adequate psychometric properties, then, perhaps, they should not be used.

Data relating to sensitivity to change, to effect size of treatment, and cross-sectional comparison

A major use of QOL scales is in clinical trials where sensitivity to longitudinal change is an important attribute. Authors often present data demonstrating that their scale is sensitive to change in these circumstances, and the same data can be used for another purpose, to demonstrate effect size of a treatment. A longitudinally sensitive scale should produce a large effect size in a clinical trail. The effect size in a clinical trial depends on the proportion of shifting and shifting items in the QOL scale and that proportion depends, for reasons shown above, on the type of item, the population and treatment. Effect size is always the consequence of interaction between treatment, scale and population – it is not a unique feature of a scale. Neither the scale nor the treatment can be characterised as showing a particular effect size (i.e., having a particular sensitivity to change), because each depends on other factors. Of course, if one compares the effect size of one scale with another over several different studies, including different treatments and populations, then it is possible to draw conclusions about how the two scales perform in general, but comparisons between scales based on only one or two RCTS are unsafe. The same argument applies to the inference of the efficacy of a treatment from its effect size on a QOL scale: the effect size and hence the apparent treatment efficacy will be affected also by the population and the scale.

Discussion

The best way to select a QOL questionnaire is examine the items of the scale carefully, and judge to what extent the set of items – i.e., the shopping bag of experiences – matches the requirements of the research that is to be carried out. This selection does not require a QOL expert: it can be done by anyone with a good understanding of the disease and the research requirements. A QOL scale should have adequate psychometric properties, but beyond that psychometrics seldom plays a crucial role in selecting between scales. In addition, the statistical properties and prior performance of the scale can provide additional information, but it is important not to over generalise from statistical data. The important factor, in the case of longitudinal research, is to choose a scale where items are appropriate for the population and type of improvement predicted from treatment. In the case of cross-sectional research, the important factor is to choose a scale that has a full and varied range of items which apply to the kind discrimination needed. In the case of clinical as opposed to research use, one can focus more on the types of issues that relevant to the clinical setting. Common differences between good longitudinal and good cross-sectional scales are shown in Table 1.

When selecting an instrument from those available it may happen that no one scale is ideal. Under these circumstances, the researcher needs to make a clinical judgement about suitability from those available, compromising between the various attributes described above. QOL scale selection is not an exact science because it is often difficult to predict the performance of a scale in advance. If no scale is suited for a particular purpose, then researchers should consider developing a new one, but there are many currently available [1].

References

1. Garratt A, Schmidt L, Mackintosh A and Fitzpatrick R: **Quality of life measurement: bibliographic study of patient assessed health outcome measures** *BMJ* 2002, **324:**1417-9.
2. Hyland ME: **Recommendations from quality of life scales are not simple** *BMJ* 2002, **325:**599.
3. Guyatt GH, Kirschner B and Jaeschke R: **Measuring health status: what are the necessary measurement properties** *J Clin Epidemiol* 1992, **45:**1341-1345.
4. Hyland ME: **Antiasthma drugs: Quality of life rating scales and sensitivity to longitudinal change** *PharmacoEconomics* 1994, **6:**324-329.
5. Juniper EF, Guyatt GH, Epstein RS, Ferrie PJ, Jaeschke R and Hiller TK: **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.
6. Hyland ME: **The items in quality of life scales: How item selection creates bias and how bias can be prevented** *PharmacoEconomics* 1992, **1:**182-190.
7. Hyland ME and Sodergren SC: **Development of a new type of global quality of life scale, and comparison of performance and preference for 12 global scales** *Quality of Life Research* 1996, **5:**469-480.
8. Singh SJ, Sodergren SC, Hyland ME, Williams J and Morgan MDL: **A comparison of three disease-specific and two generic health-status measures to evaluate the outcome of pulmonary rehabilitation in COPD** *Respiratory Medicine* 2001, **95:**71-77.
9. Hyland ME, Singh SJ, Sodergren SC and Morgan MDL: **Development of a shortened version of the Breathing Problems Questionnaire suitable for use in a pulmonary rehabilitation clinic: a purpose-specific, disease-specific questionnaire** *Quality of Life Research* 1998, **7:**227-233.