Department of Economics
School of Social Sciences

Statistical analysis of EQ-5D profiles: does the use of value sets bias inference?

David Parkin[1]
City Health Economics Centre, City University, London

Nigel Rice[2]
Centre for Health Economics, University of York.

Nancy Devlin[3]
City Health Economics Centre, City University, London

Department of Economics
Discussion Paper Series
No. 08/10

[1] City Health Economics Centre, Economics Department, City University, London EC1V OHB Phone: +44 (0) 20 7040 0171 e-mail: d.parkin@city.ac.uk

[2] Centre for Health Economics, University of York, Heslington York Y01 5DD e-mail: nr5@york.ac.uk

[3] City Health Economics Centre, Economics Department, City University, London EC1V OHB Phone: +44 (0) 20 7040 8518 e-mail: n.j.devlin@city.ac.uk

Statistical analysis of EQ-5D profiles: does the use of value sets bias inference?

David Parkin
City Health Economics Centre, City University, London
Nigel Rice
Centre for Health Economics, University of York.
Nancy Devlin
City Health Economics Centre, City University, London

Summary

Health state profile data, such as those provided by the EQ-5D, are widely collected in clinical trials, population surveys and a growing range of other important health sector applications. However, these profile data are difficult to summarise to give an overall view of the health of a given population that can be analysed for differences between groups or within groups over time. A common way of short-cutting this problem is to transform profiles into a single number, or index, using sets of weights, often elicited from the general public in the form of values. Are there any problems with this procedure? In this paper we demonstrate the underlying effects of the use of value sets as a means of weighting profile data. We show that any set of weights introduces an exogenous source of variance to health profile data. These can distort findings about the significance of changes in health between groups or over time. No set of weights is neutral its effect. If a summary of patient reported outcomes is required, it may be better to use an instrument that yields this directly – such as the EQ VAS – along with the descriptive instrument. If this is not possible, researchers should have a clear rationale for their choice of weights; and be aware that those weighs may exert a non-trivial effect on their analysis. This paper focuses on the EQ-5D, but the arguments and their implications for statistical analysis are relevant to all health state descriptive systems.

Keywords: Utility measurement, Health economics, Quality of Life, EQ-5D, Statistical inference

JEL: C120, C810, D610, I120, I180

Introduction

Health state data, derived from a variety of different measurement instruments, are widely collected in clinical trials, population surveys and a growing range of other important applications in the health sector. Many of these instruments describe health states in terms of multiple attributes. Moving beyond description to statistical analysis of health outcomes imposes requirements on the way that these data are presented – in particular, the need to construct from them a single indicator of the direction and magnitude of differences in health.

An important example of an instrument for measuring health outcomes is the EQ-5D, developed by the EuroQol Group and intended for patient self-completion.[1] The standard *EQ-5D self-report questionnaire* comprises two parts: the *EQ-5D self-classifier*, which uses the *EQ-5D descriptive system*, and the *EQ VAS*, which records an overall rating of health from 0-100 on a visual analogue scale. Together, these give valuable information on the health states of individuals, groups of patients and populations. Common uses of the EQ-5D include comparisons of population health over time and between countries; monitoring the health of patient groups; and gauging the effectiveness and cost effectiveness of treatments. In the United Kingdom (UK), a number of other, more innovative uses of the EQ-5D have recently been proposed. From April 2009, patients' self-reported health improvement on the EQ-5D will be used as a hospital performance indicator in the National Health Service (NHS), used to help patients choose which hospital to be referred to.[2] The EQ-5D has also been advocated for use in measuring the productivity and performance of the NHS.[3]

The self-classifier provides health state *profile* data that categorise the respondent's health state according to the EQ-5D descriptive system, known as *EQ-5D health states*. Unlike the EQ VAS data, these data are not single numbers, but a set of categorical variables recording the respondent's health as one of three levels (essentially no, some or extreme problems) on each of five dimensions (mobility, self-care, usual activities, pain & discomfort and anxiety & depression). Profiles therefore contain rich and detailed information on health states; unfortunately it is hard to summarise them concisely and even harder to analyse them for statistical inference purposes.

Many studies have attempted to overcome this by converting the profile to a single index number, known as an *EQ-5D index*. Such an index is easy to summarise and is easily amenable to statistical inference techniques. It is constructed by applying weights to each of the levels within each dimension and adding these together for particular health states. However, although this is in some cases a valid procedure, in many it is not. The most obvious context in which an index is appropriate is cost-effectiveness analysis.[4] However, there are many uses, both economic and non-economic, in which it is not so obvious.

This paper examines the statistical properties of EQ-5D indexes, exposing the underlying processes and assumptions, and offers recommendations to users of the EQ-5D regarding the collection and analysis of data. In particular, it examines the index in the context of statistical inference. While our focus is on the EQ-5D, the issues we identify are relevant to *any* multi-dimensional health state classification system which seeks to provide an overall summary measure of health.

The uses of profile data

EuroQol Group guidance on analysing and reporting EQ-5D data[5] suggests that profile data should be reported as tables of frequencies and percentages of respondents having particular levels in particular dimensions. In the Group's publication on measuring self-reported population health[6], descriptive data are presented as the percentage reporting each level in each dimension and summary statistics of EQ VAS ratings, with hypothesis testing restricted to the EQ VAS ratings. The EuroQol website gives similar examples on the page 'How to report EQ-5D results'.[7]

The Group gives no guidance on how profile data might be analysed for statistical inference purposes, for example testing for significant differences in health status between patient groups or for significant changes within groups over time. The most obvious descriptive technique to use for such categorical data would be contingency tables and tests of association such as χ^2. However, there are 243 possible EQ-5D health states, and although this is small relative to many other health status instruments, it is large enough to mean that there are severe limitations on how useful these techniques can be. It is possible to examine each of the five dimensions separately for differences in the number of respondents in each level, requiring a 3x3 contingency table. Analysing five such tables is quite feasible, but this does not provide an overall summary of the profile. If a complete profile is looked at, there is potentially a 243x243 contingency table, which is too cumbersome for any practical purposes. Of course, in practice a large number of the potential states will not be found within real samples, but it is nevertheless unlikely to result in a small enough number to be analysable. Furthermore, samples where there are many small or empty cells will require categories to be collapsed to make the data amenable to analysis, resulting in a loss of information.

This problem is less acute if the EQ-5D is used in the way intended by the EuroQol Group, which is to use the complete EQ-5D self-report questionnaire. The EQ VAS provides an overall summary of self-reported health as a single number between 0 and 100, which can readily be presented in the form of summary statistics or graphs and used for undertaking statistical inference. However, in many applications the EQ-5D self-classifier is used as a stand alone measure, without the EQ VAS. Examples include the use of the EQ-5D in official surveys such as, in the UK, the Health Survey for England (HSE)[8] and surveys of NHS hospital inpatients.[9]

Analysing EQ-5D profile data is also not problematic if they are to be used for cost-effectiveness analysis – the context in which health economists are most familiar with the use of the EQ-5D. The EQ-5D is one of the health outcome measures recommended by pharmaceutical reimbursement authorities in The Netherlands, New Zealand, Norway, Italy, Hungary, Poland and Portugal[10]. The UK's National Institute for Health and Clinical Excellence (NICE) states that "To allow comparisons across technologies, the Institute requires that health states should be measured in patients" and that "Currently, the most appropriate choice in the UK appears to be the EQ-5D."[11]. In its application in cost effectiveness analysis, an EQ-5D index is used as the quality of life element in the calculation of Quality Adjusted Life Years (QALYs). The index is constructed by applying to each EQ-5D health state a weight which represents the utility or value of that state, on a scale which has a maximum value of 1, representing full health, an anchor of 0, representing a state equivalent to being dead, and with states regarded as worse than being dead having a value lower than 0. The use of the EQ-5D in economic evaluation is facilitated by the existence of EQ-5D value sets, often called 'tariffs', in many countries[10]. These have been generated by asking members of the general public to consider health states described by the EQ-5D, which they may or may not have experienced, and to value those states using techniques such as a Visual Analogue Scale (VAS) and Time Trade-Off (TTO). In the UK, the most widely used weights are TTO values from a UK population study known as the MVH (Measuring and Valuing Health) study.[12] More recently, a similar set of weights has been produced for the United States.[13,14]

However, what of applications of the EQ-5D other than in economic evaluation? The EQ-5D is now used in a wide range of applications in the health sector other than assessments of value for money – see Table 1.

Table 1: Applications of the EQ-5D in the health sector (*excluding* economic evaluation).

Application	Examples of use
Comparing the health status of populations over time; comparing the health status of local populations with national population health, comparing population health internationally.	Within the UK, the EQ-5D has been used in Health Survey for England (HSE) surveys for a number of years[8] and in NHS surveys of inpatients.[9]
Comparing the health of patient groups with that of comparable members of the general public	Self-reported health on the EQ-5D has been compared with age/sex-adjusted population norms for the UK, to gauge the effects on quality of life associated with type 2 diabetes.[15]
Determining clinical priorities and managing demand for referral from primary to secondary services.	New Zealand's 'points system' for elective surgery utilised specially-designed scoring instruments to determine 'clinical thresholds' and 'financial thresholds'[16]. The EQ-5D could potentially be used in this context.
Routine use of the EQ-5D before and after surgery, as a means of monitoring, managing and reporting the performance of hospitals (or clinical teams) in improving health.	From 2009, NHS hospitals will have a requirement to measure, using the EQ-5D as well as, in each case, a condition-specific instrument, patients' self-reported health before and after surgery, for all patients undergoing four surgical procedures.[2]
Monitoring variations and trends in the health of patients with long term conditions.	Use of the EQ-5D in a daily patient diary for multiple sclerosis patients[17].

Each of these applications confronts the challenge noted above: given that EQ-5D profiles are not readily amenable to statistical analysis, how can overall health and changes in health be summarised and analysed? One solution is to use exactly the same procedure used in economic evaluation and apply a set of weights to create a single index. In principle this could be any set of weights, but in practice the most common approach is to use the published value sets used in economic evaluation. Are there any problems in using this solution?

EuroQol Group guidance to users of EQ-5D value sets[18] warns against using value sets to produce a single index for statistical analysis of profiles that are meant to be purely descriptive. This is because "there is no 'neutral' set of weights that can be used for this purpose" and "No set of weights is objective". It advises that "it may be better *not* to use an index, but to report the EQ-5D profiles themselves in some detail" and "where a single number is required to represent health … it may be more appropriate to focus on the EQ VAS data provided by the relevant patients or populations themselves … rather than applying social value sets to their EQ-5D profiles." [18 p. 40]

However, the Group's guidance is not always followed. The report[19] which informed the UK Department of Health's introduction of routine use of the EQ-5D collected EQ-5D profiles and applied the MVH TTO value set to these to facilitate analysis. Similarly, the EQ-5D data collected by the Health Survey for England appears to be summarised by application of the 'EuroQol tariff'.[20] In both contexts, the data are explicitly meant to represent *patient* reported outcome measures, and there is no intention to interpret the numbers as values. Indeed there are numerous published examples of EQ-5D profile data being converted into EQ Index values in non-economics applications, either instead of or in addition to analysis of the EQ VAS. In most cases, no clear rationale for doing so is provided by the authors. Recent examples include a study of the quality of life of diabetes patients[21]; the relationship between quality of life and alcohol dependency[22]; a longitudinal study of population health in Sweden[23]; and quality of life among stroke survivors[24].

The EuroQol Group's guidance is mainly based on the disputed legitimacy of using value sets applicable to economic evaluation for other purposes. The value sets that are used for economic evaluation have a clear theoretical rationale that underpins the form of the weights, the way that they are derived and their meaning. This rationale may not extend to other uses and indeed may be entirely out of keeping with them. The weights used in economic evaluation are explicitly regarded as 'values' or 'utilities', with a quite narrow definition attached to them. There is a clear meaning attached to the values 1 and 0 and to values less than 0; it is desirable to use a recognised stated preference technique to obtain them; and there is a justification for the use of the general population as a source of weights. The resulting weights should only be used in other applications if the same theoretical rationale also applies.

A contrasting view is that the weights really do not matter; most of the variation in an EQ-5D index is due to differences between respondents rather than the weighting structure and it is unnecessary to use social preference-based values rather than a simple set of weights. This was the argument of Prieto and Sacristán (2004), who compared for a large data set an index weighted using the MVH values with one using equal weights for both levels and dimensions, finding mean values that differed by what they regarded as a negligible amount.[25]

But there is another relevant question about the use of weights: what are the statistical properties of the resulting index? One issue is that although the weights are treated as fixed coefficients, they are in fact themselves estimates derived from a sample, and therefore have a sample distribution which ought to be taken into account in any statistical inference. Since statistical testing of the index tests both the profile data and the weights, we speculate that some account ought to be taken of the variability of the weights. Unfortunately, as this will be derived from a completely different sample, it is not obvious how this might be undertaken. This is a complex question which is not dealt with in this paper. We look at another issue, which applies whether or not weights are fixed - how adding weights to profile data affects statistical inferences made about the resulting index.

A decomposition of the EQ-5D index

It is important to remember that we are interested in the EQ-5D index as a summary of a set of EQ-5D profile data, referring to a particular patient or population group. Any EQ-5D index, whatever the source and structure of its weights, is made up of not only the profile data under analysis but also another data set made up of the weights that are used to convert the profile data into a single number. It is useful to analyse the role of each of these data sets in determining the numbers that are calculated for the index, and we will therefore decompose the index into its separate constituents.

The usual procedure for calculating an EQ-5D index is first to convert the profile data into a set of binary variables. The most important of these binary variables are derived from the categorical variables that describe the levels of each dimension of the EQ-5D. As described, each dimension of the EQ-5D has three levels: ‘no problems’; ‘some problems’; and ‘extreme problems'. Two binary variables can therefore be used for each dimension: the presence or absence of the 'some problems' level, which we will refer to as Level 2; and the presence or absence of the 'extreme problems' level, which we will refer to as Level 3. The absence of both of these indicates 'no problems', or Level 1. As there are five dimensions, there are 5x2=10 binary variables of this kind.

There may also be other binary variables that represent interaction effects between dimensions and levels. Two widely used variables of this kind are one that records the presence of any Level 2 or Level 3 state, referred to as a 'constant' because of its derivation from a regression equation, and another that records the presence of any Level 3 health state, referred to as 'N3', a term used in the MVH study.[12]

Weighting the profile data to create an index therefore in practice means attaching weights to these binary variables. In what follows, it is assumed that for each binary variable, 0 represents absence of a level within a dimension or an interaction and 1 represents its presence.

Let

H_j = Index score[4] for individual j
b_i = Binary variables representing levels within dimensions
w_i = Index weights for binary variable b_i

Then

$$H_j = \sum_{i=1}^{k} w_i b_{ij} \qquad \text{where } \begin{cases} i = 1,\ldots,k \\ j = 1,\ldots n \end{cases} \qquad [1]$$

In this formulation, no constraints are placed on the numeric values given to the index weights or the index scores because, as will be shown, this has no impact on inference issues. Nevertheless, it does raise an important issue, because for most applications using health status information we will be interested in absolute values of the index, of differences in the

[4] For clarity, we describe the index here as an indicator of subtractions from full health. It would be necessary to subtract this value from 1 to give an index where higher values mean better health.

index between different groups and of changes within groups over time. In clinical studies, for example, knowing the absolute value of the effect size is as important as knowing whether or not it is significantly different from zero. For that reason, it is not enough simply to have weights that describe the relative importance of different levels and dimensions. A key point here is the fact that the concept of overall health status has no natural units in which it can be measured, although value sets do in effect have units of measurement, derived from the way that they are constructed and the way that they are used.

In what follows we will examine the problems that arise in statistical testing of the mean of this index, in particular comparing differences in its value for two population or patient groups. However, it should be noted that the same problems will arise in calculating and making inferences about other statistics, for example correlation coefficients, and therefore tests of association as well as tests of difference.

To tests for differences in the value of the index requires us to know its mean and variance. Since the individual values of the index are calculated from the values of the binary variables and the weights, it follows that the mean and variance of the index must also be a function of these two elements. In fact, the mean and variance of H turn out to have fairly obvious relationships with the corresponding statistics for the binary variables.

Mean of the index

Let
$\overline{H}$ =Mean value of H_j
$\overline{b}_i$= Mean value of b_i

It can be shown that

$$\overline{H} = \sum_{i=1}^{k} w_i \overline{b}_i \qquad [2]$$

The mean value of H_j is therefore a weighted sum of the means of the binary variables, using the same weights and weighting structure as for the individual values of H_j.

What does this mean value tell us? If the binary variables are indicators of the presence or absence of a particular level in a particular dimension, then their mean values are simply the proportions of the sample with that level in that dimension. The mean value of H_j is in effect a summary of these proportions over all levels and dimensions, weighting the relative importance attached to them. This seems a reasonable thing to do, if the weights are appropriate to the task. However, it is important to recognise that the mean value is determined both by the data that directly describe the sample – the profile data - and by data that are generated externally to the sample – the weights. It is equally accurate to describe the mean value of the index as a set of constant values (the w_i) that are weighted by the relative proportions observed in a particular sample or population.

The addition of terms such as N3 and the constant complicates this slightly. It is arguably justified to include these if the index is regarded as measuring a concept of health in which the simple binary variable weightings do not fully capture the relative importance of different levels and dimensions. However, users of the index who mean it merely to be a descriptive summary ought to be aware of the weight structure that they are building in, and its origins. In the case of most of the value sets widely used, the interaction terms are included solely to optimise the statistical properties of an equation describing an entirely different type of data – valuations of described health states - taken from a completely separate sample or population.

Variance of the index

Let

σ_H^2 = variance of H_j

σ_i^2 = variance of b_i

σ_{li} = covariance of b_i and b_l, $i < l$

Again, it can be shown that

$$\sigma_H^2 = \sum_{i=1}^{k} w_i^2 \sigma_i^2 + 2\sum_{l<i} w_l w_i \sigma_{li} \qquad [3]$$

The variance of H_j is a weighted sum of both the variance of each binary indicator and the covariance between each binary indicator, again using the index weights. It cannot be expressed as a function simply of the variances.

The interpretation of the term in [3] that includes the variances of the binary variables is quite straightforward. The variance of a binary variable is the proportion of ones in the data multiplied by the proportion of zeroes; since the proportion of ones is the mean, the variance is:

$$\sigma_i^2 = \bar{b}\left(1 - \bar{b}\right)$$

The covariance term is more complicated. The covariance of two binary variables is the proportion of cases in the data where both variables take the value one, which is the mean of their products, minus the product of the proportion of each that takes the value one, which is the product of their means. The covariance is therefore:

$$\sigma_{li} = \overline{b_l b_i} - \bar{b}_l \bar{b}_i$$

This is complicated by the fact that the b_i include indicators representing different levels within a single dimension. Where this is the case, the first term in the covariance definition will be zero and the covariance must therefore be negative.

The covariance term introduces interactions not only between the different levels and dimensions of the data but also between the weights – the $w_i w_l$ terms. In general, the variance for H_j is a complex function not only of variance and covariance terms but also of the weights. Expressed in terms of means of the binary variables, the variance in [3] becomes:

$$\sigma_H^2 = \sum_{i=1}^{k} w_i^2 \left(\bar{b}\right)\left(1 - \bar{b}\right) + 2\sum_{l<i} w_l w_i \left(\overline{b_l b_i} - \bar{b}_l \bar{b}_i\right) \qquad [4]$$

Testing for differences between means

For simplicity, let us assume that we are conducting a test of the differences in H between two groups X and Y with the same sample size, n. The means for populations X and Y are:

$$\overline{H}_X = {}^{1}\!/_{n} \sum_{j=1}^{n} H_{Xj}$$

$$\overline{H}_Y = {}^{1}\!/_{n} \sum_{j=1}^{n} H_{Yj}$$

The difference in means can be shown to be

$$\overline{H}_X - \overline{H}_Y = \sum_{i=1}^{k} w_i \left(\overline{b}_{Xi} - \overline{b}_{Yi}\right) \qquad [5]$$

The difference in means is therefore a linear weighted sum of the difference in means of the binary variables, in other words of the difference between the proportions of the two samples having each level in each dimension. This definition and those of the individual index numbers and their mean value are symmetrical.

The variance of this difference is:

$$Var\left(\overline{H}_X - \overline{H}_Y\right) = {}^{1}\!/_{n}\left(\sum_{i=1}^{k} w_i^2 \left(\sigma_{Xi}^2 + \sigma_{Yi}^2\right) + 2\sum_{l<i} w_l w_i \left(\sigma_{Xli} + \sigma_{Yli}\right)\right) \qquad [6]$$

which can also be expressed as

$$Var\left(\overline{H}_X - \overline{H}_Y\right) = {}^{1}\!/_{n}\begin{pmatrix} \sum_{i=1}^{k} w_i^2 \left(\left(\overline{b}_X\right)\left(1 - \overline{b}_X\right) + \left(\overline{b}_Y\right)\left(1 - \overline{b}_Y\right)\right) \\ + 2\sum_{l<i} w_l w_i \left(\left(\overline{b_{Xl} b_{Xi}} - \overline{b}_{Xl}\overline{b}_{Xl}\right) + \left(\overline{b_{Yl} b_{Yi}} - \overline{b}_{Yl}\overline{b}_{Yl}\right)\right) \end{pmatrix} \qquad [7]$$

The variance of the difference in means is again a non-linear weighted function of the variances and covariances. This definition and that of the variance are symmetrical.

The appropriate test statistic for a normally distributed variable with unknown population standard deviation is the *t* statistic:

$$t = \frac{\overline{H}_X - \overline{H}_Y}{\sqrt{\operatorname{var}\left(\overline{H}_X - \overline{H}_Y\right)}} \qquad [8]$$

If we substitute expressions [5] and [6] into [8], it is apparent that the numerator of [8] is linear in weights, but the denominator is non-linear and complex. The weights do not ‘cancel out’ and they are therefore an important determinant of the value of the *t* statistic. The consequence is that the weights chosen may determine statistical significance; different sets of weights may give different conclusions about whether or not two groups are significantly different to each other. As suggested earlier, the same problem would occur if other statistical procedures are carried out, such as tests of association using correlation coefficients.

This problem is however restricted to weights that differ in their relative values, not their absolute values. To demonstrate this, suppose that we have two sets of weights that give the same relative weights to different levels and dimensions – for example, the weight for Level 2 pain is twice as big as that of Level 3 pain, the weight for Level 2 anxiety is twice that of Level 2 self-care, and so on for all possible dimensions and levels. However one has absolute values twice as big as the other. This can be represented as a scalar λ, in this case 2, applied to each w_i. Examining equations [2] and [5], it is apparent that this will result in $\overline{H}$ becoming $\lambda\overline{H}$ and $(\overline{H}_X - \overline{H}_Y)$ becoming $\lambda(\overline{H}_X - \overline{H}_Y)$. Similarly, from equations [3] and [6], σ_H^2 will become $\lambda^2\sigma_H^2$ and $Var\left(\overline{H}_X - \overline{H}_Y\right)$ will become $\lambda^2 Var\left(\overline{H}_X - \overline{H}_Y\right)$. λ will therefore cancel out in equation [8], so that the value of t is not dependent on the absolute values of the weights.

A simulated empirical example

In order to demonstrate the conclusions of our analysis, a simple simulation was performed. The procedure was as follows:

1 Generate a random sample of 100 from the 243 possible EQ-5D health states, without replication (Group 1).

2 Generate an identical set of data, except for 5 of the health states chosen at random, which are changed to give a one-level improvement in one dimension of the EQ-5D (Group 2).

3 Apply different sets of weights to the resulting data. Each set of weights is applied equally to the two groups to generate an index score for each.

4 Calculate paired comparison *t* tests of the differences in the index scores of Group 1 and Group 2 for each set of weights.

Because of their widespread use in the UK and elsewhere, the MVH weights were chosen as one of the comparators. It is possible to compare these directly with two other published sets of weights, based on data from the Netherlands[26] and Spain[27]. There are several other TTO-based sets available, but these have a slightly different structure (for example, no N3 terms for weights from Denmark, Japan, the USA and Zimbabwe)[10]. Our conclusions apply equally to the use of these sets of weights, and to others based on other valuation methods, but to make a clean comparison we focus on the three sets that are identically structured and derived.

To explore in more detail how different relative weights affect inference, we also examined nine artificial sets of weights. Four of these had equal weights for each of the dimensions, but different weights for the levels within each of them. As explained, the binary variables represent the presence or absence of Level 2 and Level 3 responses within a dimension. A higher weight was given to Level 3 by multiplying the Level 2 by a constant. Four different multipliers were used; 2, 3, 4 and 10. The other five sets had different weights for each level in different dimensions. This was achieved by adding a constant absolute increment to both weights within a dimension, so that, for example, 0.005 is added to the Level 2 and Level 3 weights for mobility to obtain those for self-care, 0.01 is added to obtain those for usual activities, and 0.015 and 0.02 to obtain those for pain/discomfort and anxiety/depression respectively. Using the same increment, two different multipliers, 4 and 5, were used for the relative weights for levels. In each case, the resulting weights were applied first to the dimensions in the order in which they appear in the EQ-5D questionnaire and then to the dimensions in a random order. The final set of weights was for a higher increment, 0.01.

These numbers used as multipliers and increments were chosen to produce mean values comparable to those arising from the MVH weights. That is an arbitrary criterion and it is therefore not meaningful to compare the actual mean values. However, as explained, the absolute values do not affect the results in terms of *t* and *p* values. So, to prevent any misleading comparisons, Table 2 reports the *p* values for a two-tailed paired comparison test for each set of weights, but does not report any of the means and standard deviations.

Table 2: Significance tests of differences between simulated samples of health states according to different sets of weights.

Description	*p* value, two tailed *t*-test
UK weights	0.041
Netherlands weights	0.067
Spanish weights	0.029
Equal weights dimensions; Level3 = 2*Level2	0.025
Equal weights dimensions; Level3 = 3*Level2	0.033
Equal weights dimensions; Level3 = 4*Level2	0.052
Equal weights dimensions; Level3 = 10*Level2	0.160
Unequal weights dimensions; Increment = 0.005, Level3 = 5*Level2	0.095
Unequal weights dimensions; Increment = 0.005, Level3 = 5*Level2 Same weights as previous, but moved between dimensions	0.054
Unequal weights dimensions; Increment = 0.005, Level3 = 4*Level2	0.065
Unequal weights dimensions; Increment = 0.005, Level3 = 4*Level2 Same weights as previous, but moved between dimensions	0.041
Unequal weights dimensions; Increment = 0.01, Level3 = 5*Level2	0.122

The table show that when using the UK and Spanish weights the *t* test detected a significant difference in the means at the 5% level, although with quite different *p* values. Using the Netherlands weights, however, suggests that the means are not significantly different at the 5% level. Where weights were assumed equal for each dimension, the *p* value depended on the relative size of the weights for level 2 and level 3. The greater the difference in the weights between levels, the higher the *p* value. Relative weightings of 3 and below generated a significant difference at the 5% level, those 4 and above an insignificant difference.

Where weights were unequal between dimensions, the same result held and in addition the greater the difference in weights between dimensions the lower the *p* value. The interaction between the weighting structure and the data was also an important determinant of the *p* value. If the weights for different dimensions were switched, the *p* value changed, and in one of the examples this resulted in a change from significance to non-significance.

We therefore find that our theoretical results are both correct and of some practical importance. The *t*-test is sensitive to weights, and therefore a test of differences in the means of the health index between groups tests not only differences that arise from the groups but also differences between the weights. Another way to view this is that there is, for a given sample size, a value of *t* for a given set of weights that is simply modified by the data. There must be some doubt, therefore, about whether the levels of significance that are implied by the test are in fact appropriate for this kind of data.

Conclusions

An alarming conclusion from our empirical analysis is that clinical trials carried out in the UK and the Netherlands that produce exactly the same EQ-5D data might lead researchers to make completely different statistically-based conclusions about the effectiveness of the intervention being studied. Many countries do not have weights of their own and therefore have to use foreign weights; whether or not an intervention is seen as effective in such a country might depend on which other country's weights it chooses.

The problem of using an index created from weighted profile data is that any statistical analysis is affected by information that is not inherent to the sample data; variations in the index reflect variations not only in the sample but also in the weights. Statistical tests of significance also introduce complexities into the weighting system via the variance, in particular interactions between weights, levels and dimensions that are not in the original weighting structure. This may have the uncomfortable implication that conventional significance tests are inappropriate and give misleading levels of significance. This is most obviously an issue where the index is intended as a convenient summary of descriptive data, but the problem will also apply where it is intended as a value or utility, unless the underlying weights can be regarded as fixed. If they are regarded as variable, this casts some doubt on the results from very many published cost-effectiveness studies.

The main conclusion from this analysis is to reinforce the recommendation that applying sets of weights to profile data in order to produce an index should *not* be used as a short-cut method of summarising profile data and of facilitating statistical inference from such data. It is only justifiable to create and make inferences from an index where the weights have specifically been created to produce a meaningful value relevant to the purpose for which the data are being used, which will represent some concept of what health is. This need not be 'value' in the economic sense; it could for example represent a clinical view of the severity or burden of illness. For inference purposes, all that is then needed is weights that represent the relative importance of different dimensions and levels of health, although producing numbers that are meaningful in an absolute sense requires much more. If the weights used in an index have simply been imported and are not relevant to the purpose for which they are used, using that index is at best misguided and at worst misleading.

References

1. Brooks R. EuroQol: the current state of play. Health Policy. 2006; 37(1):53-72.

2. Department of Health. Guidance on the routine collection of Patient Reported Outcome Measures. 2008. London: Department of Health.

http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_081100 (Accessed May 9th 2008).

3. Office for Health Economics. Report of the OHE Commission on NHS outcomes, performance and productivity. 2008; London: Office for Health Economics.

4. Mortimer M, Segal L. Comparing the incomparable? A systematic review of competing technologies for converting descriptive measures of health status into QALY-weights. Med Decis Making. 2008; 28: 66-89.

5. Krabbe PFM, Weijnen T. Guidelines for analysing and reporting EQ-5D outcomes. In: Brooks, R., Rabin, R. and de Charro, F. (Editors) The measurement and valuation of health status using EQ-5D: a European perspective. Kluwer Academic Publishers: Dordrecht, 2003.

6. Szende A, Williams A. (Editors) Measuring self-reported population health: An international perspective based on EQ-5D. SpringMed Publishing Ltd, 2004.

7. EuroQol Group. 2007. http://www.euroqol.org/

8. The Information Centre. Health Survey for England 2006: summary of key findings. NHS: The Information Centre, 2008.
http://www.ic.nhs.uk/webfiles/publications/HSE06/HSE06_Summary.pdf

9. Picker Institute Europe. Inpatient questionnaire. Department of Health. 2002. DH_4076506.pdf.www.dh.gov.uk/prod_consum_dh/idcplg?IdcService=GET_FILE&dID=24955&Rendition=Web (accessed May 8th 2008).

10. Szende A, Oppe M, Devlin N. (Editors) *EQ-5D Value Sets: Inventory, Comparative Review and User Guide*, Springer Verlag, 2007

11. National Institute for Clinical Excellence. Guide to the Methods of Technology Appraisal. National Institute for Clinical Excellence, London, 2004. http://www.nice.org.uk/niceMedia/pdf/TAP_Methods.pdf (accessed June 10th 2008)

12. Dolan P. Modelling valuations for EuroQol health states. Medical Care. 2007; 35(11):1095-108.

13. Shaw JW, Johnson JA, Coons SJ. US valuation of the EQ-5D health states: development and testing of the D1 model. Medical Care 2005; 43:203-220.

14. Sullivan PW, Ghushchyan V. Preference-based EQ-5D index scores for chronic conditions in the United States. Med Decis Making. 2006; 26: 410-420.

15**.** Holmes J, McGill S, Kind P, Bottomley J, Gillam S, Murphy M. Health related quality of life in type 2 diabetes (T^2ARDIS-2). Value in Health. 2000; 3(1):47-51.

16. Derrett S, Devlin N, Hansen P, Herbison P. Prioritising patients for elective surgery: a prospective study of clinical priority assessment criteria in New Zealand. Int J Technol Assess Health Care. 2003; 19(1):91-105.

17. Parkin D, Rice N, Jacoby A, Doughty J. Use of a visual analogue scale in a daily patient diary: modelling cross-sectional time-series data on health-related quality of life. Soc Sci Med. 2004; 59(2):351-60.

18. Devlin N, Parkin, D. Guidance to users of EQ-5D value sets. In: Szende, A., Oppe, M. and Devlin, N. (Editors) EQ-5D Value Sets: Inventory, Comparative Review and User Guide, Springer Verlag, 2007.

19. Browne J, Jamieson L, Lewsey J, van der Meulen J, Black N, Cairns J, Lamping D, Smith S, Copley L, Horrocks J. Patient Reported Outcome Measures (PROMs) in elective surgery. Report to the Department of Health. 12 December 2007. London: London School of Hygiene and Tropical Medicine. www.lshtm.ac.uk/hsru/research/PROMs-Report-12-Dec-07.pdf (accessed June 10th 2008).

20. The Data Archive. UK Data Archive Study number 5809 Health Survey of England 2006: List of Variables. 2008. University of Essex. http://www.data-archive.ac.uk/doc/5809/mrdoc/pdf/5809datadocs.pdf (accessed May 9th 2008).

21. Grandy S, Fox KM. EQ-5D visual analog scale and utility index values in individuals with diabetes and at risk for diabetes: Findings from the Study to Help Improve Early evaluation and management of risk factors Leading to Diabetes (SHIELD). Health Qual Life Outcomes. 2008; 27: 6:18.

22. Günther OH, Roick C, Angermeyer MC, König HH. Responsiveness of EQ-5D utility indices in alcohol-dependent patients. Drug Alcohol Depend. 2008;1; 92(1-3): 291-5.

23. Burström K, Johannesson M, Rehnberg C. Deteriorating health status in Stockholm 1998-2002: results from repeated population surveys using the EQ-5D. Qual Life Res. 2007; 16(9):1547-53.

24. Fischer U, Anca D, Arnold M, Nedeltchev K, Kappeler L, Ballinari P, Schroth G, Mattle HP. Quality of Life in Stroke Survivors after Local Intra-Arterial Thrombolysis. Cerebrovasc Dis. 2008; 16; 25(5): 438-44.

25. Prieto L, Sacristán, J. What is the value of social values? The uselessness of assessing health-related quality of life through preference measures. BMC Medical Research Methodology. 2004; 4: 10.

26. Lamers LM, McDonnell J, Stalmeier PFM, Krabbe PFM, Busschbach JJV. The Dutch tariff: results and arguments for an effective design for national EQ-5D valuation studies. Health Economics. 2006; 15(12): 1121-32.

27. Badia X, Roset R, Herdman M, Kind P. A comparison of GB and Spanish general population time trade off values for EQ-5D health states. Med Decis Making 2001; 2191: 7-16