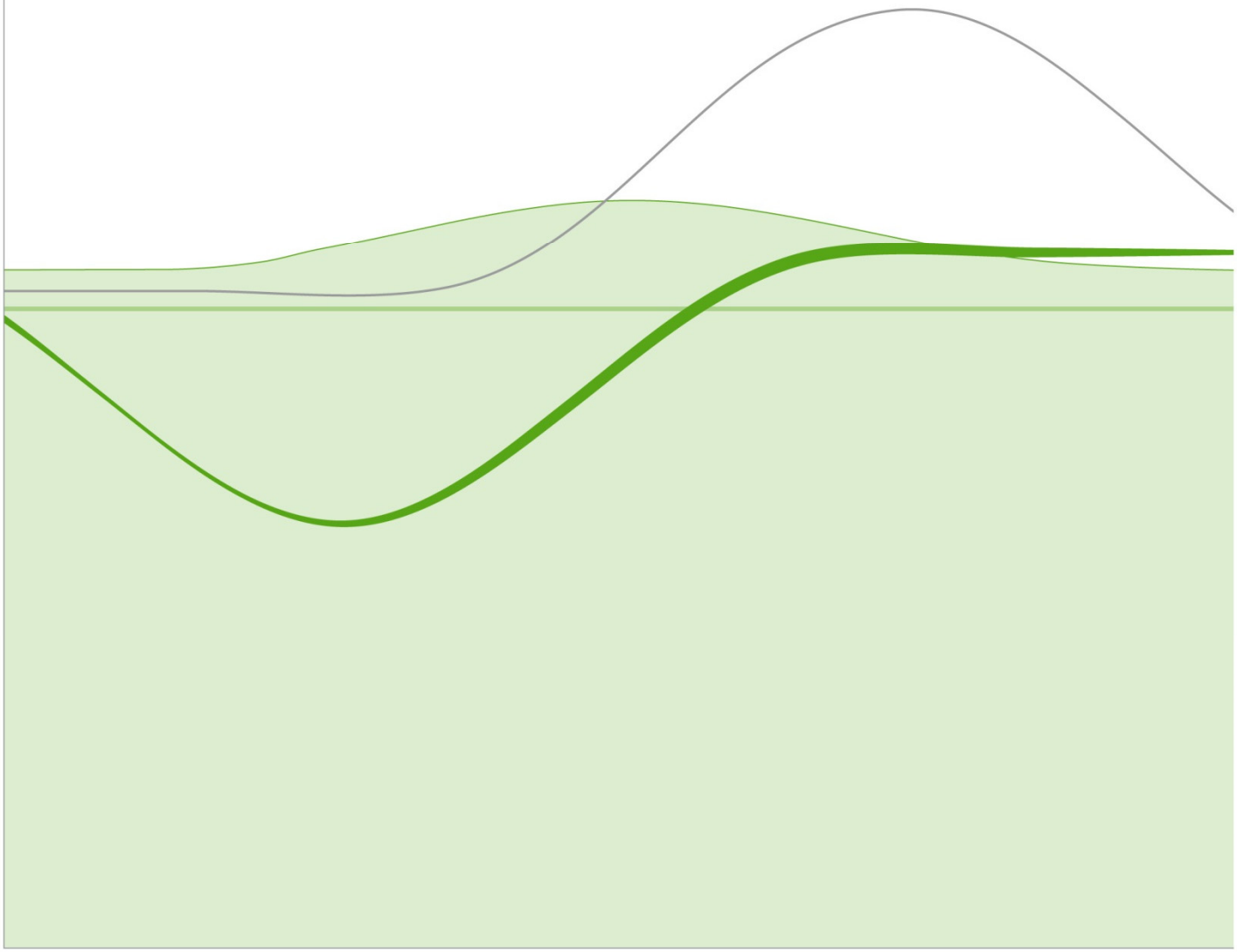
Office of Health Economics Research

Research Paper 14/04

What Determines the Shape of an EQ-5D Index Distribution?

August 2014

David Parkin, Nancy Devlin and Yan Feng



What Determines the Shape of an EQ-5D Index Distribution?

David Parkin[1], Nancy Devlin[2] and Yan Feng[2]

[1]Department of Primary Care & Public Health Sciences, King's College London

[2]Office of Health Economics, London

August 2014

Research Paper 14/04

For further information please contact:
Professor Nancy J. Devlin
ndevlin@ohe.org
Office of Health Economics
Southside, 105 Victoria Street
London SW1E 6QT
United Kingdom
Tel: +44 207 747 8858

©Office of Health Economics

About OHE Research Papers

OHE Research Papers are intended to provide information on and encourage discussion about a topic in advance of formal publication. Once a version of the Research Paper's content is published in a peer reviewed journal, that supersedes the Research Paper and readers are invited to cite the published version in preference to the original version.

Acknowledgements

The authors would like to thank Alastair Gray and Mike Drummond for their helpful comments and suggestions on an earlier draft of this paper.

Funding

The Office of Health Economics receives research and consulting funding from the Association of the British Pharmaceutical Industry (ABPI). No separate funding was received for this paper.

Disclaimer

The views expressed in this publication are those of the authors and do not necessarily represent those of the experts consulted, the ABPI or the OHE.

CONTENTS

ABSTRACT

Background: Distributions of EQ-5D index values in patient and general populations typically have a non-normal distribution, divided into two distinct groups. It is important to understand to what extent this is determined by the way that the EQ-5D index is constructed rather than by the true distribution of ill-health.

Aims: This paper examines the determinants of the 'two groups' distribution typical of EQ-5D index data. We examine the extent to which the distribution is attributable to properties of the EQ-5D classification system used to create health state profiles or to the properties of the weights applied to profiles.

Methods: We analyse data from the English NHS PROMs programme (hip and knee replacements and varicose vein and hernia repairs) and from a study of two chronic conditions (asthma and angina). The distributions of EQ-5D index values are compared with distributions from which weights have been stripped, and profile data decomposed into their constituent dimensions and levels. They are also compared with condition-specific indexes and the distributions of EQ-5D indexes using different country weights, both Time Trade-Off and Visual Analogue Scale based.

Results: The EQ-5D picks out differences between patients in respect of dimensions that are mainly observed at level 2 or 3. The weights commonly used to calculate the index exacerbate this grouping by placing a larger weight on level 3 observations, and generate a noticeable gap in index values between the groups.

Conclusions: It is both important and informative to undertake exploratory data analysis on EQ-5D data. The analytical methods used for this may be simple. Concentrating on the EQ-5D index in effect obscures useful information about health states and may even produce misleading information.

1. INTRODUCTION

The EQ-5D is internationally one of the most widely-used health-related quality of life instruments (Brooks, 1996). It is therefore very important that we obtain a good understanding of the characteristics of EQ-5D data. Distributions of EQ-5D index values in patient and general populations typically have a non-normal distribution divided into two distinct groups. (This is often described as 'bi-modal', but we will argue that this is a misleading label.) In examining the distribution of health states within a particular population, it is useful to know to what extent its shape is determined by the way that the EQ-5D index is constructed rather than by the true distribution of ill-health. In addition, the EQ-5D index is constructed from two separate elements – a classification system used to create health state profiles and a set of weights applied to profiles – and it is useful to know the relative importance of these factors in shaping health state distributions.

This paper examines the determinants of the shape of EQ-5D distributions, and in particular the origins of the 'two groups' distribution. We analyse data from patients undergoing elective surgery (hip and knee replacements and varicose vein and hernia repairs) in the English NHS and from primary care patients in a study of two chronic conditions (asthma and angina). The distributions of EQ-5D index values are compared with distributions from which weights have been stripped, and the profile data are decomposed into their constituent dimensions and levels to see how they influence each of those distributions. The distributions that arise from applying different sources of weights ('value sets') are demonstrated. The distributions of EQ-5D index values are also compared with the distributions of condition-specific indexes assessed for the same patients. We conclude by suggesting practical ways in which researchers should analyse EQ-5D index data to obtain richer results than are conventionally reported.

2. ANALYSING EQ-5D DATA

The EQ-5D is used widely in economic analyses, population health surveys and, more recently, for routine assessment of patients' health, for example the NHS Patient Reported Outcome Measures (PROMs) programme (Devlin, Parkin and Browne, 2010). The EQ-5D instrument comprises two self-report elements: the EQ-5D self-classifier, where respondents tick boxes to indicate which of three levels of problems (no, some, extreme) they have on each of five dimensions (Mobility (MO), Self-care (SC), Usual activities (UA), Pain & discomfort (PD), and Anxiety & depression (AD)) to create a health 'profile'; and a visual analogue scale, the EQ-VAS, on which respondents rate their overall health from 0 (worst health imaginable) to 100 (best health imaginable).

The profile data and the EQ-VAS both provide valuable information about the patient's view of their own health and can, themselves, be the focus of analysis. For example, Devlin, Parkin and Browne (2010) demonstrate a number of ways in which profile data can be analysed, including a Health Profile Grid and a Paretian Classification of Health Change. However, by far the most common way of analysing data from the EQ-5D is to use index values to summarise the profile data. These index values provide, for each of the 243 ($=3^5$) possible states described by the EQ-5D, a value on a scale anchored at 1 (full health) and 0 (a state as bad as being dead), with values < 0 indicating states worse than being dead. Typically, these values are obtained for a sub-set of states from surveys of the general public, using stated preference methods to find out their views about how good or bad the states are in their opinion. The resulting sample data are then used to model the values for all states. The resulting value sets, often called 'tariffs', result from and are influenced by choices about whose values are relevant, for example the general public or patient populations, what methods are used to elicit preferences, for example Time Trade Off, Visual Analogue Scale or Standard Gamble, and how the data are modelled.

There are clear normative grounds for using index values to summarise EQ-5D profile data where the purpose is to estimate Quality Adjusted Life Years in economic evaluation. For example, it is often argued that it is the views of the general public that are relevant in this context, as taxpayers and potential users of the NHS. However, index values are also widely used in other sorts of applications, where that rationale may not be relevant. For example, the Health Survey for England (Craig and Mindell, 2013) reports all EQ-5D data in terms of index values, even though the data are intended as descriptive rather than evaluative. This is probably because converting profile data into a single index number is convenient: single numbers are easier to analyse than profiles comprising multiple dimensions. It may also be that it results from a misunderstanding about what the index values represent; for example, that they are like the 'scoring systems' used in condition specific instruments, which are developed as an integral part of the instrument.

There are some important concerns about the use of index values. Parkin, Devlin and Rice (2010) and Willke et al (2010) show that the use of value sets to summarise EQ-5D

profiles introduces an exogenous source of variance, which can bias statistical inference. For example, conclusions about whether there are statistically significant differences in the health of two regions, or over time, or between two arms of a clinical trial, will be influenced by which value set is used, and what its particular properties are. More generally, there is no such thing as a 'neutral' value set; any way of weighting EQ-5D profile data will exert an influence on results[1].

Given that EQ-5D is such a widely used instrument and that its data are so often summarised using the index, obtaining a good understanding of the characteristics of index-weighted EQ-5D data is very important. One particular issue, which has been widely identified as problematic for the use of the EQ-5D, is that the distribution of values is non-normal and has what is often described as a 'bi-modal' shape. The main concern has been about estimation rather than hypothesis testing, with the suggestion that the unusual shape might result in a non-normal distribution of residuals when the EQ-5D is the dependent variable in regression analyses.

A further problem is that a 'bi-modal' distribution could imply that there are in fact two separable patient populations that should be analysed separately. It might be that many patient populations comprise distinct groups, and the EQ-5D is capable of picking out two from many of them. Alternatively, it might be that the EQ-5D profile tends to divide patients into groups, even though they are from a common distribution, or that this arises from the weighting of EQ-5D profiles. Knowing which of these possibilities is the source of the observed distributions in patient data is crucial to their proper analysis and interpretation.

[1] This point applies equally to the scoring systems of other health measures, both generic and condition specific, and includes measures that simply sum responses with equal weight.

3. THE DISTRIBUTION OF EQ-5D INDEX VALUES

Studies that have identified this issue – which we review below – often suggest that the distribution of EQ-5D index values is bi-modal or tri-modal, but those labels are misleading. A better description is that EQ-5D data appear to fall into two groups with an identifiable gap in index values between them. We will refer to these as the 'high cluster' and the 'low cluster'. The 'tri-modal' label has been used where there are many observations of people who have no problems according to the EQ-5D classification. These form a third group, again with an identifiable gap in values from the high cluster. The size of the gap in value terms between no problems (11111 = 1) and the next best health state, (11211 = 0.883 using the weights most widely used in the UK (Dolan, 1997)) is important, but the reason for the existence of a single value 'group' at 1 is obvious, and its mode is trivially determined. The interesting questions concern the other two groups and the gap between them, which has been identified as being around 0.5 (Versteegh et al, 2010) or 0.45 (Brazier et al, 2004; Hernández Alava, Wailoo and Ara, 2012). The reason why 'bi-modal' is a misleading description for this phenomenon is that the modes of the two groups are not their most interesting feature; the groups do not always have a single local mode; and in practice these modes are never actually identified, reported or analysed.

This feature of the distribution of EQ-5D values has been reported in studies of a diverse range of conditions. There have been many in arthritis (Weijnen, de Wit and de Charro, 1998; Fransen and Edmonds, 1999, Conner-Spady and Suarez-Almazor, 2003; Russell et al, 2003; Marra et al, 2004; Scott et al, 2007; Harrison et al, 2008; Harrison et al, 2009; Lillegraven, Kristiansen and Kvien, 2010; Versteegh et al, 2010 Gaujoux-Viala et al, 2011; Gaujoux-Viala et al, 2012), rheumatic disease (Wolfe and Hawley, 1997; Conner-Spady and Suarez-Almazor, 2001) and in orthopaedic conditions and treatments including herniated lumbar disc (Jansson et al, 2005), common spinal conditions for which surgery is indicated (McDonough and Grove, 2005), total knee replacement (Xie et al, 2007; Dakin, Gray and Murray, 2012), lumbar spinal stenosis (Jansson et al, 2009), all elective orthopaedic operations (Jansson and Granath, 2010; Rolfson et al, 2011) and hip arthroplasty (Paulsen et al, 2012). Studies of other conditions include accident victims with life-threatening injuries (Weijnen, de Wit and de Charro, 1998), breast cancer (Conner-Spady et al, 2001), chronically ill patients undergoing haemodialysis (Gerard et al, 2004), lower back pain, chronic obstructive pulmonary disease, end-stage renal disease (Weijnen, de Wit and de Charro, 1998), irritable bowel syndrome, leg ulcer and osteoporosis (Brazier et al, 2004; Hernández Alava, Wailoo and Ara, 2012), menopausal women and healthy older women (Brazier et al, 2004), pregnant women with chronic energy deficiency (Shaheen and Lindholm, 2006), HIV (Huang et al, 2008), postmenopausal women (Langdahl et al, 2009) multiple myeloma and non-Hodgkin's lymphoma (Versteegh et al, 2010), asthma, chest pain, Clodronate, hormone replacement therapy, leg reconstruction and varicose veins (Hernández Alava, Wailoo and Ara, 2012) and multiple sclerosis (Hawton et al, 2012).

The reason for the gap has often been alleged to be that the decrement in the EQ-5D index between levels 2 and 3 is relatively large compared with that between levels 1 and

2 (Fransen and Edmonds, 1999). More specifically, the so-called 'N3' term (Dolan, 1997) used to calculate the EQ-5D index in the UK and elsewhere has been implicated (Wolfe and Hawley, 1997; Conner-Spady et al, 2001; Brazier et al, 2004; Hernández Alava, Wailoo and Ara, 2012). This gives a large decrement to the index score if there is a level 3 state in any dimension.

The presence or absence of a level 3 observation does, in practice, place the UK index score above or below 0.5. If there is any level 3 in a profile, the index score value is at most (1-[0.081+ 0.269]) = 0.650. The maximum that any profile can take with a level 3 in a particular dimension is therefore 0.336 (Mobility), 0.436 (Self-care), 0.556 (Usual activities), 0.264 (Pain & discomfort) or 0.414 (Anxiety & depression). The smallest level 2 increment for dimensions other than Usual activities is 0.069 (Mobility). Therefore only one of the 211 (= 3^5 – 2^5) profiles that include a level 3 observation also has an index value above 0.5, namely 11311. This is an unusual profile that is very rarely encountered in practice. Conversely, all of the 32 profiles that do not include a level 3 observation take a value above 0.5.

Related to this is evidence that in mapping between the EQ-5D and other health indexes, the EQ-5D may 'overestimate' the scores for more severe EQ-5D health states. Rowen, Brazier and Roberts (2009) hypothesized that predictions are poor for more severe states, defined as EQ-5D index <0.5, because they all have at least one dimension at the most severe level and the EQ-5D model uses an N3 term. They tested the importance of the N3 term by re-estimating the EQ-5D model without it using the same data and methods as the Dolan (1997) original. Although the predictions for more severe health states were better, they still appeared to be overestimated.

Versteegh et al (2010) tested the hypothesis that N3 in itself does not generate a 'bi-modal' distribution, by generating a random set of EQ-5D cases with an equal distribution of answers across the dimensions. They claimed that the resulting index scores were normally distributed (though it would be more accurate simply to say that the distribution did not have the two-groups-and-gap shape) suggesting that N3 is not the sole cause of that shape. They concluded that although N3 is a factor in the 'bi-modal' distribution and 'overestimation' in mapping states whose values are <0.5, these are also due to the fact that there are fewer observed responses at level 3 than at level 1 or 2 and only a few states are observed.

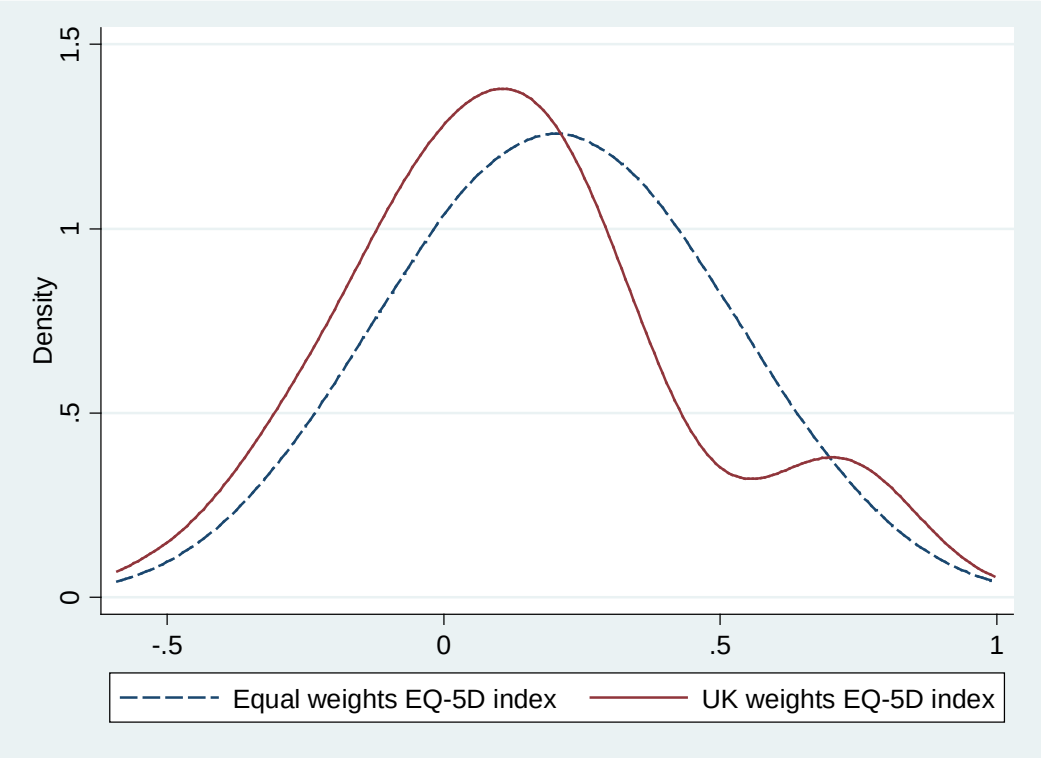
Some studies have examined the issue of whether the existence of two groups is an artefact, or in some cases or in some way do identify different patient groups. Jansson and Granath (2010) repeat an earlier assertion in Jansson et al (2009) that "We strongly believe that it is the structure of the instrument that causes this phenomenon rather than the fact that it appears to highlight 2 sub-groups of patients." One way to examine this is to seek external validation of the groups identified by the EQ-5D distribution. Hawton et al (2012) mapped a condition specific measure, the MSWS-12, with the EQ-5D. To test the specifications of their mapping models they calculated median MSWS-12 scores for two groups defined by a cut-point, which was the EQ-5D score closest to 0.5.

4. EXPLORING EQ-5D DISTRIBUTIONS

4.1 Identifying high and low clusters

The effect of index weighting can readily be shown by contrasting the shape of distributions of health states with and without index weighting. As an illustrative exercise, an artificial data set was created, comprising one observation of each possible health state described by the EQ-5D, from 11111 (no problems in any dimension) to 33333 (severe problems on all dimensions). (This is equivalent to a data set consisting of random health states, each of which has an equal probability of occurring.) Each health state was assigned two index values, one using the UK weights and the other using an 'equally weighted' scoring system, calculated by summing the level numbers (1, 2 or 3) over all dimensions, producing a number from 5 for the best health state to 15 for the worst health state. The equal weighting index was then converted so that it has the same range (-0.594 to 1) as the UK value set, using a simple linear transformation. The two resulting distributions were smoothed using identical kernel density estimation functions. Figure 1 shows the smoothed frequency distributions of the resulting index values.

Figure 1: The distribution of equally and UK weighted scores for all 243 EQ-5D profiles in the EQ-5D descriptive system



Equal weight:
Mean 0.203 Median 0.203 Mode 0.203
Std. Dev. 0.292 Skewness 0.000 Kurtosis 2.700
Normality confirmed by skewness/kurtosis tests
UK weights:
Mean 0.137 Median 0.109 Mode 0.107

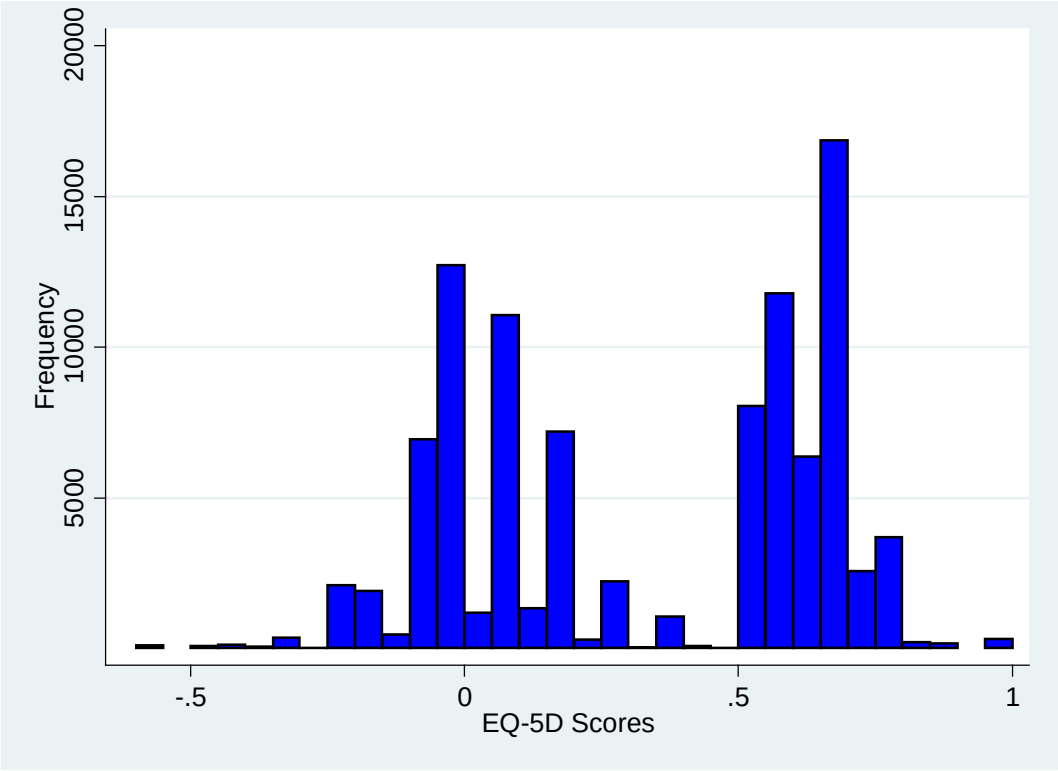
Std. Dev.	0.311	Skewness	0.437	Kurtosis 2.95
Local minimum	0.559	Local mode	0.700	

Non-normality confirmed by skewness/kurtosis tests.

On the face of it, this might suggest that equally weighted data are 'naturally' normally distributed, whereas UK weighted data are 'naturally' skewed and bi-modal. But is this purely a product of index weighting or are there other factors involved?

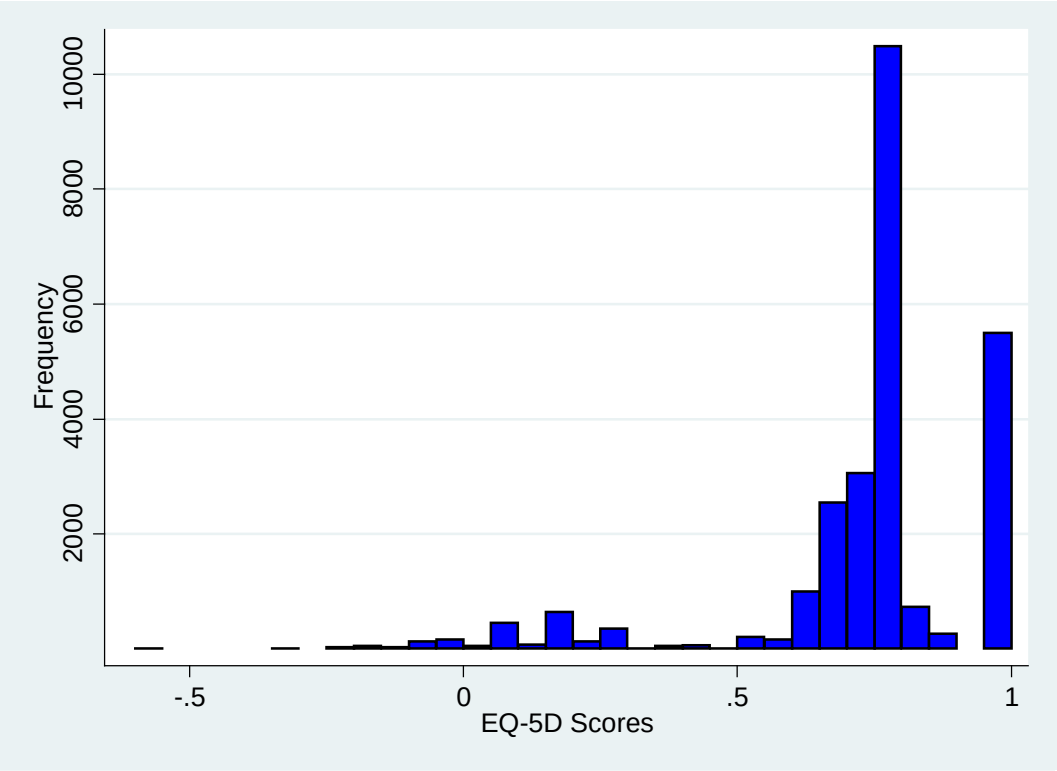
We explored this issue using data on four elective surgical procedures (hip and knee replacements and varicose vein and groin hernia repairs), collected by the NHS PROMs programme (Devlin, Parkin and Browne, 2010) from April 2009 to February 2011, and old trial data on angina and asthma patients from a trial in primary care (Eccles et al, 2002). Some of our analyses focus on hip replacement data, as an example. Figure 2 shows the distribution of UK-weighted EQ-5D index scores for patients prior to hip replacement surgery, which clearly shows a two-groups-with-gaps distribution. As suggested, the upper 'group', consisting of all patients who reported no problems on any EQ-5D dimension, is of less importance for our purposes.

Figure 2: Distribution of pre-surgery EQ-5D index scores for hip replacement patients, using UK weighted index



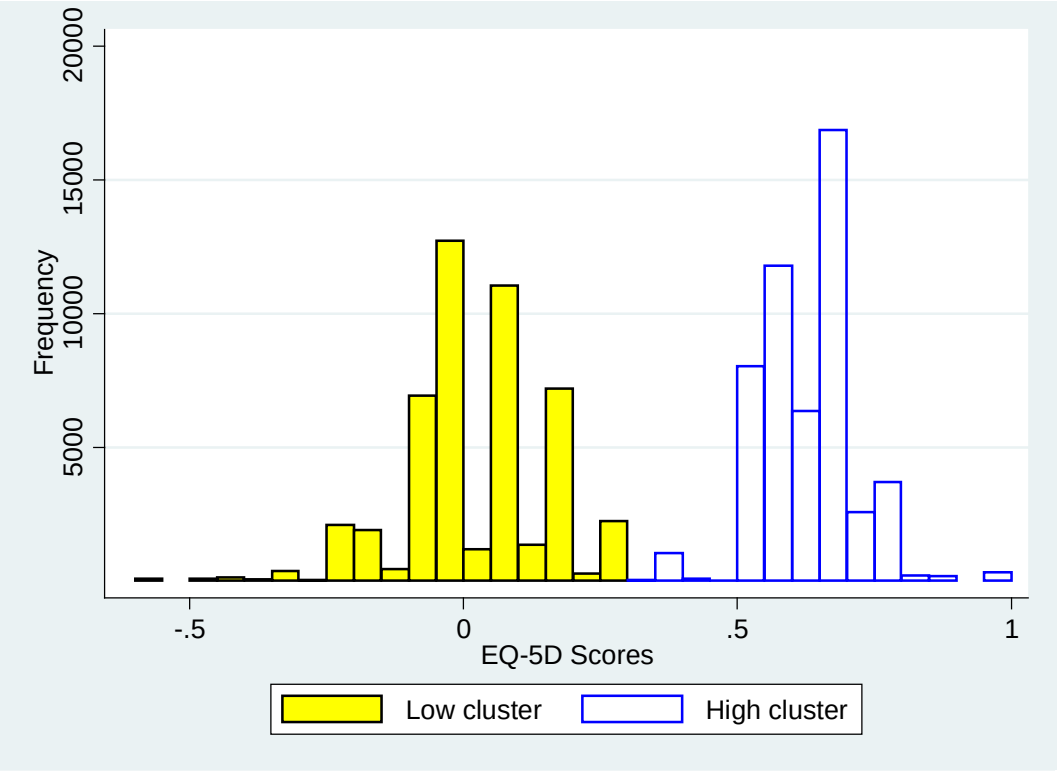
Those undergoing varicose veins surgery have very different underlying health characteristics to those of hip replacement patients. However, as Figure 3 demonstrates, a three-group-and gap distribution is apparent for varicose vein patients.

Figure 3: Distribution of pre-surgery EQ-5D index scores for varicose vein patients, using UK weighted index



Our first task was to identify more rigorously the two groups suggested by the histograms. Simple inspection of the hip replacement histogram suggests that the two could be defined as being above or below the EQ-5D index value 0.5. An alternative way of defining the groups is to use a clustering technique. Using a simple kmeans clustering procedure with two groups identifies a different dividing line, lying between the values 0.313 and 0.329. However, only 1,180 out of 99,447 observations are affected by this, as may be seen from Figure 4, which divides the observations according to the kmeans-derived clusters.

Figure 4: High and low clusters of EQ-5D scores for hip replacement patients, using UK weighted index



4.2 Is clustering a result of the EQ-5D classification system?

The index scores shown in these distributions result from applying a specific set of weights to profiles. We next examine the possibility that a source of the observed clustering is in the profiles themselves. Looking first at the different dimensions separately, Table 1 shows the percentage of responses in each level of each dimension.

Table 1: Percentage of responses in different dimensions and levels of the EQ-5D for pre-surgery hip replacement patients

Level	Mobility	Self-care	Usual activities	Pain & discomfort	Anxiety & depression
1	6.24%	66.17%	8.69%	1.08%	60.15%
2	93.34%	32.84%	76.12%	58.32%	35.23%
3	0.42%	0.98%	15.19%	40.61%	4.62%

These data suggest that in none of the dimensions is there a distribution across all three levels. There is very little difference between NHS patients about to receive hip surgery with respect to Mobility (MO); they almost always record level 2 and very rarely record level 3[2]. In each of the other dimensions, there are two levels that dominate. For Self-

[2] As noted by Oppe *et al* (2011), one reason for this is that level 3 on the EQ-5D is labelled as 'confined to bed', so even patients with very severe limitations on their mobility as a result of their hip problems, if not confined to bed, will not record mobility as a level 3.

care (SC) and Anxiety & depression (AD) these are levels 1 and 2, and for Usual activities (UA) and Pain & discomfort (PD) they are levels 2 and 3. In each case, the less severe of the two levels has the largest numbers.

We can therefore rule out differences in mobility as a cause of the two groups observed in these data from hip surgery patients. To see whether the other dimensions are individually or in combination the cause, we can examine the distribution of profiles, as follows.

Table 2: Distributions of the EQ5D profile between high cluster and low cluster

Profile	Index	Number	Within cluster		Overall	
			%	Cumulative %	%	Cumulative %
High cluster						
21221	0.691	27,412	43.14	43.14	23.95	23.95
21222	0.620	10,040	15.80	58.94	8.77	42.22
22221	0.587	6,613	10.41	69.35	5.78	54.55
22222	0.516	5,525	8.70	78.05	4.83	64.98
Low cluster						
21231	0.159	10,867	21.34	21.34	9.49	33.44
22232	-0.016	7,502	14.73	36.08	6.55	48.77
21232	0.088	6,413	12.59	59.51	5.60	60.15
22231	0.055	5,518	10.84	59.51	4.82	69.80
22332	-0.074	4,177	8.20	67.71	3.65	77.23
22331	-0.003	2,135	4.19	71.90	1.87	83.27
21331	0.101	1,921	3.77	75.68	1.68	84.95
21332	0.030	1,551	3.05	78.72	1.36	86.31

This table shows only the most frequently observed profiles. Between them, these 12 profiles account for 86% of all profiles in this data set. The four within the high cluster account for 78% of profiles in that cluster, and the eight within the low cluster account for 79% of profiles in that cluster.

All twelve profiles have, as suggested by the earlier figures, MO = 2. The four main profiles in the high cluster all have UA = 2 and PD = 2. They are only distinguished by whether they have SC = 1 or 2 and AD = 1 or 2. The eight profiles in the low cluster also appear in the four high cluster profiles, but with PD = 3 and UA = 3.

4.3 Is clustering a result of weighting?

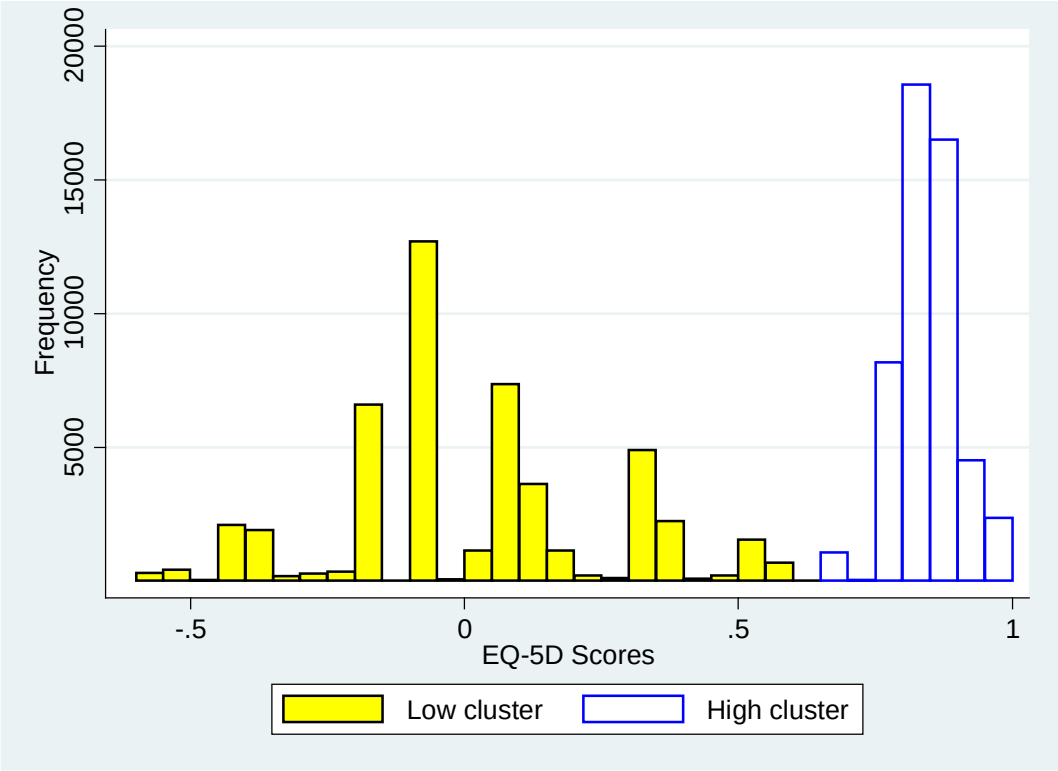
The implication of the analysis in 4.2 is that the difference between the two groups is simply in the dimensions of PD and UA – the low score cluster has people who experience more pain and discomfort and have more restrictions on their usual activities than those in the high score cluster. However, there is a complicating factor, because the difference within these dimensions is between levels 2 and 3 rather than 1 and 2, and the presence of one or more level 3 gives additional decrements in scores within the UK

value set. This is because the differences between levels 2 and 3 in each dimension are greater than those between levels 1 and 2 and also because of the N3 term. The question remains whether it is the difference between the dimensions alone that generates the clusters or the fact that the low score cluster has more level 3 observations.

For varicose veins patients, the difference between the two groups is as clear. Almost all in both groups report SC=1. Almost all patients in the high cluster report PD=2 and do not have a level 3 in any dimension. Almost all of those in the low cluster report PD=3 and a few report a level 3 in dimensions other than SC. Again, the fact that the difference is in level 3 is a complication.

One way to examine this is to change the weights used in calculating the EQ-5D scores from a profile and see if the grouping remains. There is of course no such thing as a truly 'unweighted' score and, as noted earlier, there is no 'neutral' set of weights that can be used for this purpose. As with the early analysis, it is possible to give equal weighting to levels and dimensions, but this over-smooths the data into 15 categories, giving a very weak test of the effect of more specific weights. A better alternative is to convert the set of weights into ranks. This retains the level of detail, in that every profile has an individual score, but removes the impact of size differences in the relative weighting of levels and dimensions, including the level 3 factor. The result of this is shown in Figure 5 for hip replacement patients.

Figure 5: Distribution of pre-surgery EQ-5D rank scores for hip replacement patients by cluster, using transformed ranking



In this figure, the ranks have been transformed into a variable with the same scale as the UK EQ-5D index – this is simply to make direct visual comparisons easier and has no impact on the results. Although the division into two groups is less obvious, because of the wider spread of the data in the low cluster, it is nevertheless there.

A possible conclusion is therefore:

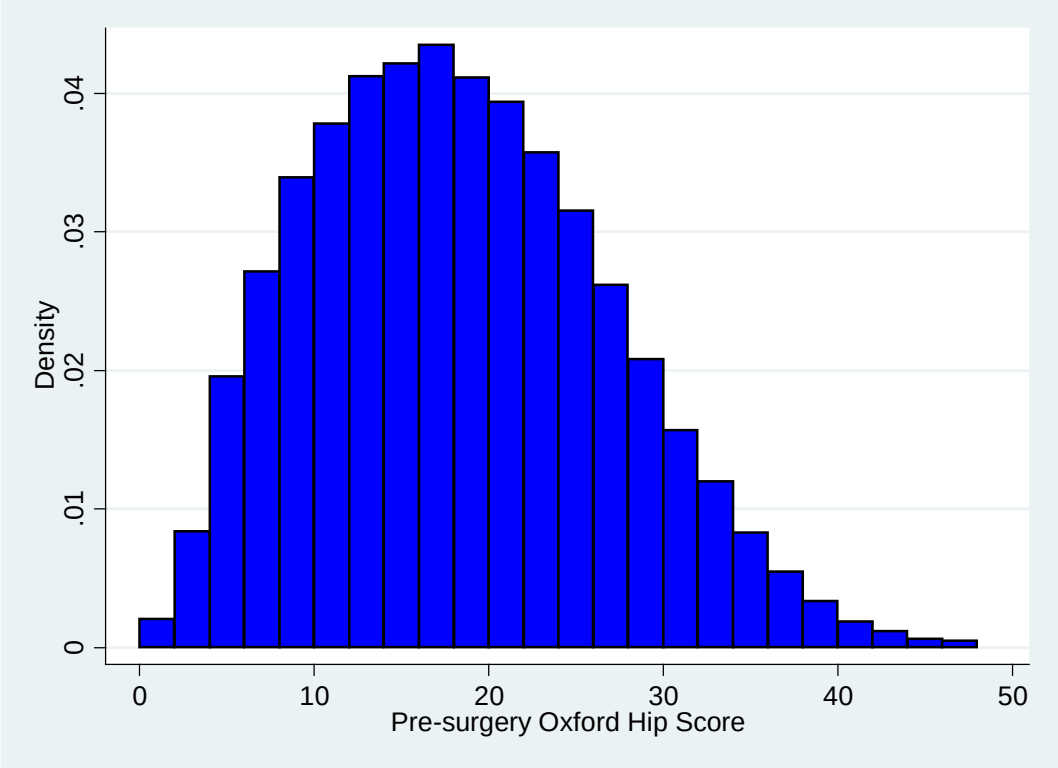
(a) The division into two groups is a result of differences between groups of patients that are identified by the EQ-5D classification system in key dimensions of health.

(b) This distribution is reinforced by the weighting system, which generates the large gap between the two groups in index values.

4.4 Does clustering reflect differences in health within patient populations?

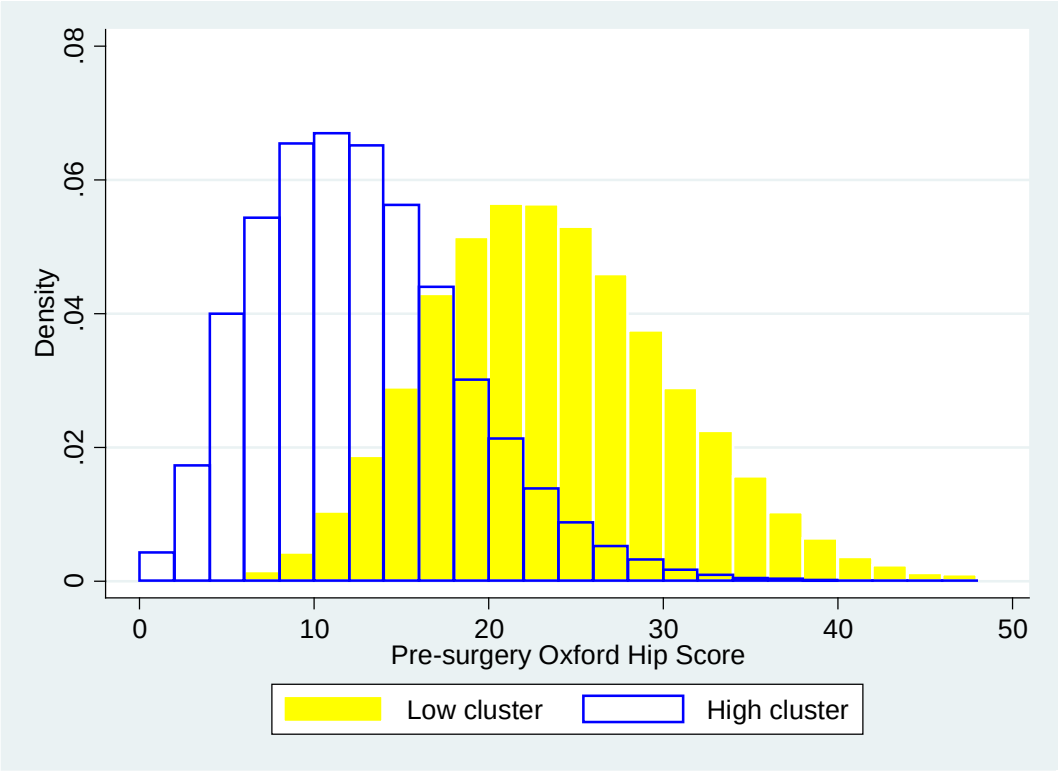
A final question is whether the two-group distribution is reflecting true differences in the underlying health of patient populations or is an artefact of the EQ-5D classification system. It is not possible to answer that question directly, but it is possible in this case to explore it using additional data. The NHS PROMs programme data also include condition specific health state instruments, in this case the Oxford Hip Score (OHS). Is the two-group distribution apparent in those data? Figure 6 suggests not.

Figure 6: Pre-surgery Oxford Hip Score distribution for hip replacement patients



This might suggest that the EQ-5D clusters are indeed an artefact. However, if we apply the EQ-5D clusters to the OHS data, in the same way as Hawton et al (2012), a different picture emerges, as shown in Figure 7.

Figure 7: Pre-surgery Oxford Hip Score distribution for hip replacement patients, divided by EQ-5D clusters



This is consistent with an overall OHS distribution that is actually made up of the overlapping distributions of two groups, one of which has worse health than the other. If so, it suggests that the OHS is not sensitive enough to discriminate between these different groups of patients. It is possible that a proper weighting system, rather than simply adding up ranks within each dimension, might enable the OHS to have more discriminatory power. But it also suggests that the EQ-5D as currently constituted has more discriminatory power.

A conclusion from these analyses is that better descriptions of EQ-5D index distributions are required, which take account of what appears to be a natural tendency for EQ-5D index data to generate a two-groups-and-gap distribution. Table 3 shows descriptive statistics for all four of the PROMs procedures and the two chronic conditions that we analysed.

Table 3: Distributions for six conditions according to high and low clusters

	Low cluster					High cluster				
	Mean	Median	Mode (Profile)	Std. Dev.	Range	Mean	Median	Mode (Profile)	Std. Dev.	Range
Hip	0.019	-0.003	-0.160 (22232)	0.128	-0.594 to 0.313	0.631	0.620	0.691 (21221)	0.087	0.329 to 0.883
Knee	0.052	0.055	0.159 (21231)	0.122	-0.594 to 0.345	0.655	0.691	0.691 (21221)	0.084	0.362 to 0.883
Vein	0.130	0.159	0.088 (21232)	0.140	-0.594 to 0.436	0.755	0.796	0.796 (11121)	0.064	0.487 to 0.883
Hernia	0.183	0.189	0.159 (21231)	0.147	-0.594 to 0.452	0.755	0.796	0.796 (11121)	0.065	0.485 to 0.883
Angina	0.088	0.088	-0.016 (22232)	0.144	-0.594 to 0.383	0.692	0.691	0.620 (21222)	0.095	0.414 to 0.883
Asthma	0.086	0.088	0.088 (21232)	0.159	-0.484 to 0.383	0.722	0.725	0.796 (11121)	0.093	0.414 to 0.883

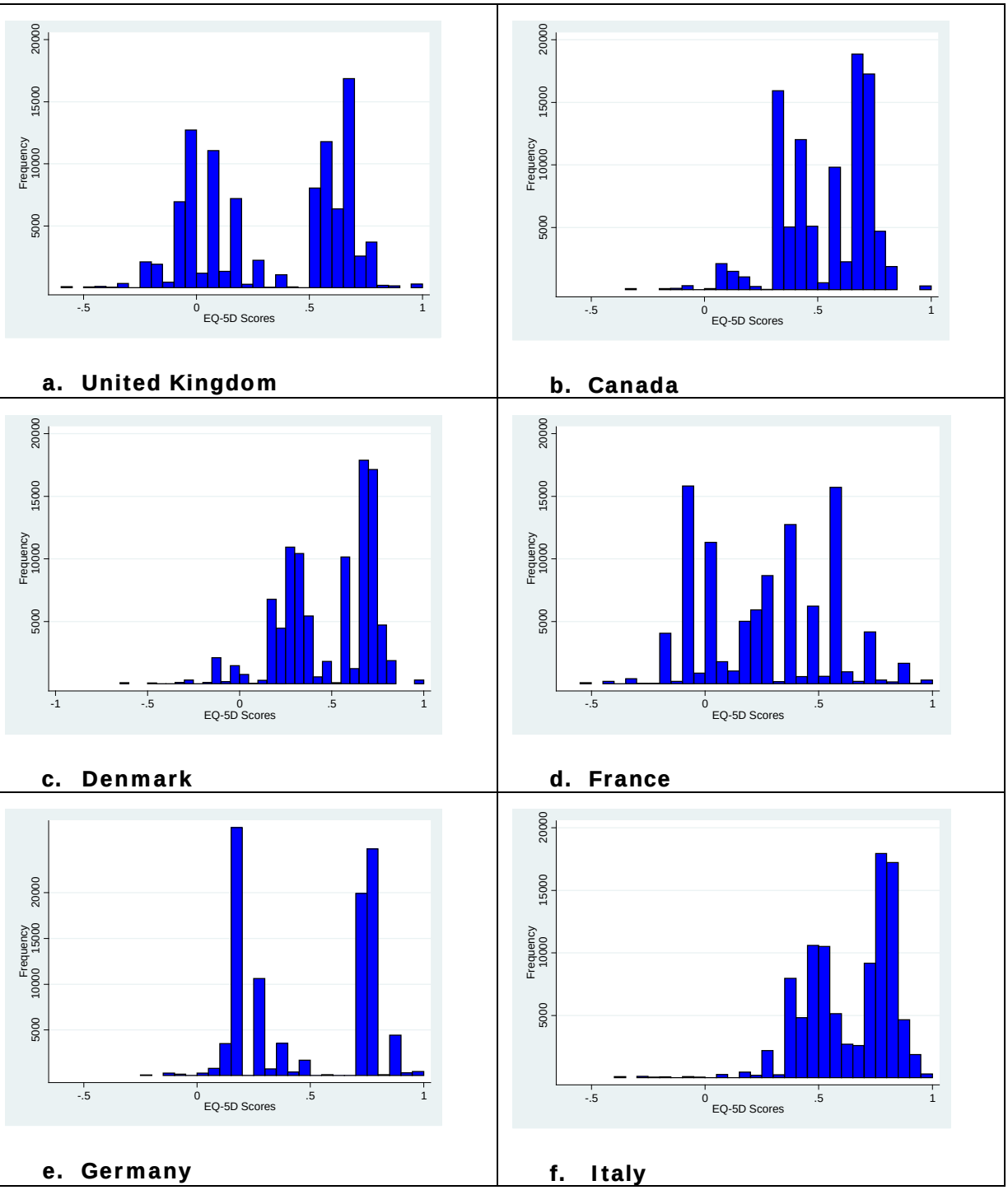
These kind of simple descriptive statistics give a far clearer picture of the distribution of EQ-5D index scores than are usually reported. This table also serves further to emphasise that the mode is not the key feature of the groups or the best descriptor of their distributions, suggesting that the label ‘bi-modal’ should not be used to describe the typical shape of an EQ-5D index distribution.

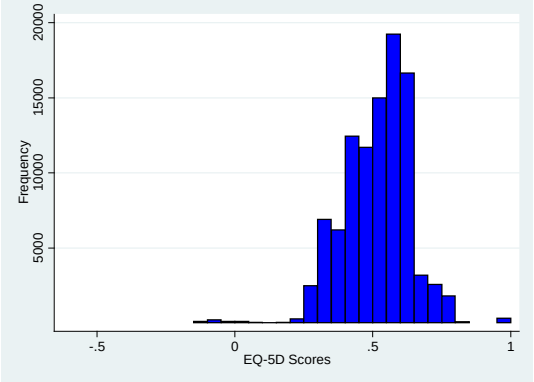
5. EXPLORING THE EFFECT OF USING DIFFERENT VALUE SETS.

Our analyses suggest that the two-group characteristic of distributions of EQ-5D index data are explained by both the nature of the EQ-5D profiles reported by patients and the system of weights used to transform profile data into a single score. In order to further highlight the contribution of the weights, we re-analysed the NHS PROMs data, applying the weights ('value sets') from each of a range of different countries (Szende, Oppe and Devlin, 2007) using a standard algorithm (Ramos-Goñi and Rivero-Arias, 2011). These value sets differ in a number of important respects: for example, some are based on preferences elicited using Time Trade Off (TTO) methods, and others by use of visual analogue scale (VAS) rating; and the way in which data were collected and modelled differed. However, each is intended to capture the preferences of the general public with respect to health as described on the EQ-5D.

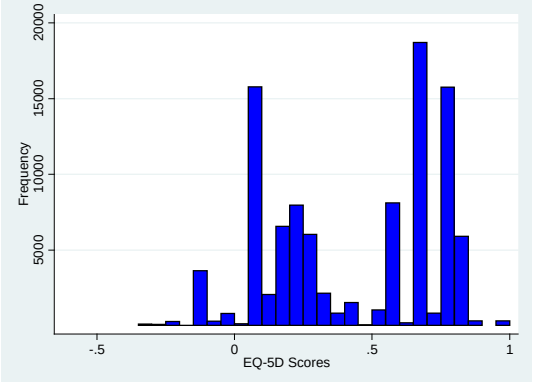
Figure 8 shows the distributions of EQ index data generated from applying each of 12 TTO value sets to the EQ-5D profile data from NHS patients prior to hip replacement surgery. Figure 8a shows the distribution arising from application of the UK value set (Figure 2), and for comparison, from the application of value sets for Canada, Denmark, France, Germany, Italy, Japan, Netherlands, South Korea, Spain, Thailand and USA. Corresponding descriptive statistics for the low and high cluster groups are reported in Table 4.

Figure 8: Distributions of pre-surgical EQ-5D data resulting from application of different value sets, using country specific weighted index

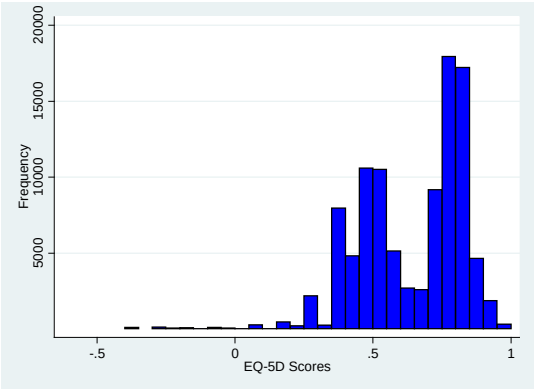




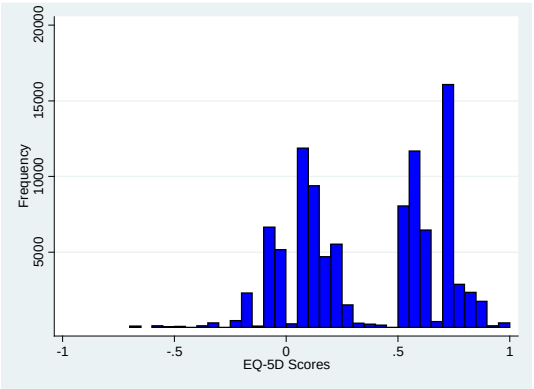
g. Japan



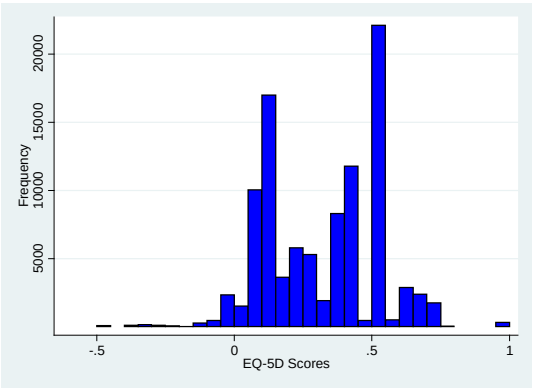
h. Netherlands



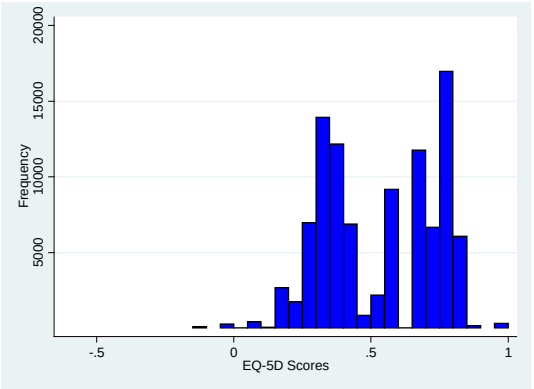
i. South Korea



j. Spain



k. Thailand



l. USA

Table 4: Distributions of pre-surgery hip condition patients, according to 12 country TTO index weights

	Low cluster					High cluster				
	Mean	Median	Mode (Profile)	Std. Dev.	Range	Mean	Median	Mode (Profile)	Std. Dev.	Range
Canada	0.341	0.369	0.339 (22232)	0.113	-0.340 to 0.511	0.681	0.663	0.726 (21221)	0.067	0.514 to 0.844
Denmark	0.231	0.258	0.258 (22232)	0.150	-0.624 to 0.442	0.674	0.660	0.723 (21221)	0.071	0.457 to 0.838
France	0.056	0.038	-0.052 (22232)	0.155	-0.530 to 0.282	0.530	0.577	0.577 (21221)	0.140	0.296 to 0.910
Germany	0.212	0.175	0.175 (22331)	0.091	-0.205 to 0.477	0.763	0.788	0.788 (21221)	0.061	0.490 to 0.999
Italy	0.459	0.482	0.482 (22232)	0.125	-0.380 to 0.614	0.786	0.775	0.827 (21221)	0.058	0.623 to 0.924
Japan	0.401	0.418	0.418 (22232)	0.082	-0.111 to 0.498	0.608	0.595	0.649 (21221)	0.062	0.506 to 0.804
Netherlands	0.143	0.174	0.092 (22232)	0.125	-0.329 to 0.396	0.698	0.693	0.775 (21221)	0.096	0.421 to 0.897
South Korea	0.454	0.513	0.513 (22232)	0.111	-0.171 to 0.591	0.734	0.723	0.766 (21221)	0.068	0.595 to 0.913
Spain	0.064	0.051	0.051 (22232)	0.137	-0.654 to 0.356	0.653	0.648	0.71 (21221)	0.100	0.362 to 0.914
Thailand	0.133	0.117	0.117 (22232)	0.099	-0.452 to 0.316	0.500	0.514	0.546 (21221)	0.091	0.329 to 0.766
UK	0.019	-0.003	-0.016 (22232)	0.128	-0.594 to 0.313	0.631	0.620	0.691 (21221)	0.087	0.329 to 0.883
USA	0.334	0.314	0.307 (22232)	0.091	-0.102 to 0.525	0.716	0.706	0.777 (21221)	0.077	0.527 to 0.860

As they use the same underlying EQ-5D profile data, the differences in the distributions illustrated in Figures 8a - 8l are purely driven by the properties of the weights embodied in each country's value set. Groupings are apparent in all but one of these distributions, although with different characteristics. The exception is the Japanese value set, which yields an almost normal distribution. One explanation for this is that the Japanese value set does not have an 'N3' term and, compared to the UK value set, has much lower utility decrements associated with problems with pain and discomfort.

These analyses show that the 'two-groups' characteristic of EQ-5D index distributions is not uniquely associated with the use of the UK value set. It also serves as a reminder

that, for any given set of patients' EQ-5D data, which value set is used to summarise them will have an important bearing on the results.

Table 5 contrasts the results for each country from two ways of defining high and low clusters. One is where the EQ Index is lower or higher than 0.5 and the other uses kmeans clusters, as described earlier for the UK. In some countries the kmeans clusters divide close to 0.5, for example the division in German weights lies between 0.477 and 0.49, with correspondingly few observations affected, and for Japanese weights no observations are affected. By contrast, the kmeans clustering procedure using South Korean weights identifies a division lying between 0.595 and 0.596, with 24% of observations affected.

Table 5: Results from alternative ways to define high and low clusters

Country	Number less than 0.5	Number in low kmeans cluster	Number affected	Max value of low cluster	Min value of High cluster
Canada	43,792	43,869	77	0.511	0.514
Denmark	46,036	44,223	1,813	0.442	0.457
France	75,279	55,517	19,762	0.282	0.296
Germany	49,288	49,222	66	0.477	0.490
Italy	27,325	45,223	17,898	0.623	0.624
Japan	40,620	40,620	0	0.498	0.506
Netherland	48,292	46,729	1,563	0.421	0.425
South Korea	19,820	43,585	23,765	0.595	0.596
Spain	49,419	49,184	235	0.356	0.362
Thailand	69,458	47,572	21,866	0.316	0.329
UK	49,416	48,236	1,180	0.313	0.329
USA	46,096	47,647	1,551	0.525	0.527

Finally, all of these value sets are derived from studies using Time Trade-Off to derive weights. Table 6 shows the results using value sets from studies that have reported Visual Analogue Scale weights. These also demonstrate clustering, suggesting that the technique used in the population surveys is not a causal factor.

Table 6: Distributions for hip condition according to 9 regional VAS index weights

	Low cluster					High cluster				
	Mean	Median	Mode (Profile)	Std. Dev.	Range	Mean	Median	Mode (Profile)	Std. Dev.	Range
Belgium	0.180	0.186	0.133 (22232)	0.095	-0.158 to 0.386	0.603	0.576	0.659 (21221)	0.081	0.393 to 0.817
Denmark	0.243	0.253	0.226 (22232)	0.082	-0.167 to 0.352	0.473	0.443	0.507 (21221)	0.080	0.361 to 0.711
Europe	0.225	0.210	0.181 (22232)	0.093	-0.074 to 0.422	0.620	0.598	0.687 (21221)	0.087	0.440 to 0.846
Finland	0.327	0.326	0.326 (22232)	0.074	-0.011 to 0.446	0.568	0.542	0.626 (21221)	0.074	0.457 to 0.795
Germany	0.241	0.237	0.237 (22232)	0.086	0.021 to 0.433	0.662	0.613	0.750 (21221)	0.114	0.472 to 0.902
New Zealand	0.211	0.234	0.167 (22232)	0.079	-0.086 to 0.392	0.574	0.556	0.627 (21221)	0.075	0.395 to 0.782
Slovenia	0.287	0.297	0.408 (22221)	0.107	-0.242 to 0.410	0.545	0.501	0.501 (21221)	0.089	0.426 to 0.818
Spain	0.243	0.222	0.222 (22232)	0.079	-0.074 to 0.421	0.601	0.594	0.645 (21221)	0.079	0.427 to 0.799
UK	0.232	0.214	0.201 (22232)	0.080	-0.073 to 0.415	0.611	0.596	0.659 (21221)	0.072	0.422 to 0.814

6. CONCLUSIONS

Non-normal distributions of index weighted EQ-5D data featuring two distinct groups of patients are commonly observed in patient populations. This has implications for statistical analysis and modelling of those data. Our analysis suggests the reasons for this shape are that the EQ-5D classification system picks out differences between patients with the same condition in respect of dimensions that are mainly observed at level 2 or 3. The weights commonly used to calculate the index exacerbate this grouping by placing a larger weight on level 3 observations, and generates a noticeable gap in index values between the groups.

A further factor involved in this is that, in general, only a few of the 243 potential EQ-5D states are observed with any great frequency. Devlin, Parkin and Browne (2010) reported that in a large and diverse data set just 22 of the 243 EQ-5D profiles covered 90% of all health states observed and 161 profiles were not found at all. One reason for this is that profiles that contain very great differences in levels between dimensions are rarely observed, for example profiles having four level 1 dimensions and one level 3. It is therefore reasonable that patients' health states form groupings for particular conditions, since we are unlikely to observe extreme variations from a typical EQ-5D profile for a particular condition.

The analysis that we have carried out is on the original 3 level version of the EQ-5D. It will be interesting to see whether similar issues apply to the five-level EQ-5D-5L (Herdman et al, 2011). We suggest that it will be important to examine EQ-5D-5L data using methods similar to those used here.

One recommendation from this analysis is that it is very important and informative to undertake exploratory data analysis on EQ-5D data, and that the analytical methods used for this may be simple. As we have argued elsewhere (Devlin, Parkin and Browne, 2010; Parkin, Devlin and Rice, 2010), concentrating on the EQ-5D index in effect obscures useful information about health states and may even produce misleading information. We suggest that this exploratory approach will enable us better to analyse EQ-5D data for comparison and inference purposes, and help in developing more accurate mapping between different health measures.

Although we have concentrated on the EQ-5D, our analytical approach also applies to any health status index that uses the weighted profile approach. This includes both generic and condition-specific measures, and also indexes that are calculated without explicit weights, such as the Oxford Hip Score. Indeed, it is arguable that a measure such as the Oxford Hip Score is far more wasteful of useful information than the EQ-5D index and that it positively obscures important differences between patients. There are potentially 244,150,625 (512) different Oxford Hip Score profiles. Its simple scoring system reduces this to 49 categories, involving a huge loss of information. Of course, the vast majority of those profiles would never be observed, but it is likely that far more

would be observed than 49 and that the differences between those profiles will be of relevance in measuring patient health.

REFERENCES

Brazier, J., Roberts, J., Tsuchiya, A. and van Busschbach, J., 2004. A comparison of the EQ-5D and SF-6D across seven patient groups. *Health Economics*, 13, pp.873–884.

Brooks R., 1996. EuroQol: the current state of play. *Health Policy*, 37(1), pp.53-72.

Conner-Spady, B., Cumming, C., Nabholtz, J-M., Jacobs, P. and Stewart, D., 2001. Responsiveness of the EuroQol in breast cancer patients undergoing high dose chemotherapy. *Quality of Life Research*, 10, pp.479-486.

Conner-Spady, B. and Suarez-Almazor, M.E., 2001. A comparison of preference-based health status tools in patients with musculoskeletal disease. Proceedings of the18th Plenary Meeting of the EuroQol Group, pp.235-245.

Conner-Spady, B. and Suarez-Almazor, M.E., 2003. Variation in the estimation of quality-adjusted life-years by different preference-based instruments. *Medical Care*, 41(7), pp791–801.

Craig R, Mindell J. (Eds.), 2013. *Health survey for England 2012*. London: The Health and Social Care Information Centre.

Dakin H, Gray A, and Murray D., 2012. Mapping analyses to estimate EQ-5D utilities and responses based on Oxford Knee Score. *Quality of Life Research*, 22(3), pp.683-694.

Devlin, N, Parkin D, Browne J., 2010. Patient Reported Outcomes in the NHS: new methods for analysing and reporting EQ-5D data. *Health Economics*, 19(8), pp.886-905.

Dolan, P., 1997. Modeling valuations for EuroQol Health States. *Medical Care*, 35(11), pp.1095-1108.

Eccles, M, McColl, E, Steen, N, Rousseau, N, Grimshaw, J, Parkin, D and Purves, I., 2002. Effect of computerised evidence based guidelines on management of asthma and angina in adults in primary care: cluster randomised controlled trial. *British Medical Journal*, 325, pp.941-947.

Fransen, M. and Edmonds, J., 1999. Reliability and validity of the EuroQol in patients with osteoarthritis of the knee. *Rheumatology*, 38(9), pp.807-813.

Gaujoux-Viala, C., Rat, A.C., Guillemin, F., Flipo, R.M., Fardellone, P., Bourgeois, P. and Fautrel, B., 2011. Comparison of the EQ-5D and the SF-6D utility measures in 813 patients with early arthritis: results from the ESPOIR cohort. *Journal of Rheumatology*, 38(8), pp.1576-1584.

Gaujoux-Viala, C., Rat, A.C., Guillemin, F., Flipo, R.M., Fardellone, P., Bourgeois, P. and Fautrel, B., 2012. Responsiveness of EQ-5D and SF-6D in patients with early arthritis: results from the ESPOIR cohort. *Annals of the Rheumatic Diseases*, 71, pp.1478–1483.

Gerard, K., Nicholson, T., Mullee, M., Mehta, R. and Roderick, P., 2004. EQ-5D versus SF-6D in an older, chronically ill patient group. *Applied Health Economics and Health Policy*, 3(2), pp.91-102.

Harrison, M.J., Davies, L.M. Bansback, N.J., Ingram M., Anis, A.H, Symmons, D.P., 2008. The validity and responsiveness of generic utility measures in rheumatoid arthritis: a review. *Journal of Rheumatology*, 35(4), pp.592–602.

Harrison, M.J., Davies, L.M. Bansback, N.J., McCoy M.J., Verstappen, S.M.M., Watson, K. and Symmons, D.P.M., 2009. The comparative responsiveness of the EQ-5D and SF-6D to change in patients with inflammatory arthritis. *Quality of Life Research*, 18, pp.1195–1205.

Hawton, A., Green, C., Telford, C,J., Wright, D.E. and Zajicek, J.P., 2012. The use of multiple sclerosis condition-specific measures to inform health policy decision-making: mapping from the MSWS-12 to the EQ-5D. *Multiple Sclerosis*, 18(6), pp.853-861.

Herdman, M., Gudex, C., Lloyd, A., Janssen, M., Kind, P., Parkin, D., Bonsel, G. and Badia, X., 2011. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Quality of Life Research*. 20(10), pp.1727-1736.

Hernández Alava, M., Wailoo, A. J. and Ara, R., 2012. Tails from the Peak District: adjusted limited dependent variable mixture models of EQ-5D questionnaire health state utility values. *Value in Health*, 15, pp.550–561.

Huang, I-C., Frangakis, C., Atkinson, M.J., Willke, R.J., Leite, W.L., Vogel, W.B. and Wu, A.W., 2008. Addressing ceiling effects in health status measures: a comparison of techniques applied to measures for people with HIV disease. *Health Services Research*, 43(1), pp.327–339.

Jansson, KÅ. and Granath, F., 2010. Health-related quality of life (EQ-5D) before and after orthopedic surgery. *Acta Orthopaedica*, 82(1), pp.82-89.

Jansson KÅ, Németh G, Granath F, Blomqvist P., 2005. Health-related quality of life in patients before and after surgery for a herniated lumbar disc. *Journal of Bone Joint Surgery (Br)*, 87-B, pp.959-964.

Jansson, KÅ., Németh, G., Granath, F., Jönsson, B. and Blomqvist, P., 2009. Health-related quality of life (EQ-5D) before and one year after surgery for lumbar spinal stenosis. *Journal of Bone Joint Surgery (Br)*, 91(2), pp.210-216.

Langdahl, B.L., Rajzbaum, G., Jakob, F., Karras, D., Ljunggren, O., Lems, W.F., Fahrleitner-Pammer, A., Walsh, J.B., Barker, C., Kutahov, A. and Marin, F., 2009. reduction in fracture rate and back pain and increased quality of life in postmenopausal women treated with teriparatide: 18-month data from the European Forsteo Observational Study (EFOS). *Calcified Tissue International*, 85, pp.484–493.

Lillegraven, S., Kristiansen, I.S. and Kvien, T.K., 2010. Comparison of utility measures and their relationship with other health status measures in 1041 patients with rheumatoid arthritis. *Annals of the Rheumatic Diseases*, 69(10), pp.1762-1767.

McDonough, C.M. and Grove, M.R., 2005. Comparison of EQ-5D, HUI, and SF-36-derived societal health state values among Spine Patient Outcomes Research Trial (SPORT) Participants *Quality of Life Research*. 14(5), pp.1321–1332.

Marra, C.A., Esdaile, J.M., Guh, D., Kopec, J.A., Brazier, J.E., Koehler, B.E. et al., 2004.. A comparison of four indirect methods of assessing utility values in rheumatoid arthritis. *Medical Care*, 42(11), pp.1125–1131.

Oppe, M., Devlin, N. and Black, N., 2011. Comparison of the underlying constructs of the EQ-5D and Oxford Hip Score: implications for mapping. *Value in Health*, 14(6), pp. 884-891.

Parkin D., Devlin N., Rice N., 2010. Statistical analysis of EQ-5D profiles: does the use of value sets bias inference? *Medical Decision Making*, 30, pp.556-565.

Paulsen, A., Pedersen, A.B., Overgaard, S. and Roos, E.M., 2012. Feasibility of 4 patient-reported outcome measures in a registry setting. A cross-sectional study of 6,000 patients from the Danish Hip Arthroplasty Registry. *Acta Orthopaedica*, 83(4), pp.321–327.

Ramos-Goñi, J. M. and Rivero-Arias, O., 2011. EQ5D: A command to calculate index values for the EQ-5D quality-of-life instrument, *The Stata Journal*, 11(1), pp.120-125.

Rolfson, O., Kärrholm, J., Dahlberg, L. E. and Garellick, G., 2011. Patient-reported outcomes in the Swedish Hip Arthroplasty Register. *Journal of Bone Joint Surgery (Br)*, 93-B, pp.867-875.

Rowen, D., Brazier, J. and Roberts, J., 2009. Mapping SF-36 onto the EQ-5 D index: how reliable is the relationship? *Health and Quality of Life Outcomes*, 7(27).

Russell, A.S., Conner-Spady, B., Mintz, A., Mallon, C. and Maksymowych, W. P., 2003. The responsiveness of generic health status measures as assessed in patients with rheumatoid arthritis receiving infliximab. *The Journal of Rheumatology*, 30(5), pp.941–947.

Scott, D.L., Khoshaba, B., Choy, E.H. and Kingsley, G.H., 2007. Limited correlation between the Health Assessment Questionnaire (HAQ) and EuroQol in rheumatoid arthritis: questionable validity of deriving quality adjusted life years from HAQ. *Annals of the Rheumatic Diseases*, 66(11), pp.1534-1537.

Shaheen, R. and Lindholm, L., 2006. Quality of life among pregnant women with chronic energy deficiency in rural Bangladesh. *Health Policy*, 78(2–3), pp.128–134.

Szende, A., Oppe, M., Devlin, N. (Eds.), 2007. EQ-5D Value sets: inventory, comparative review and user guide, EuroQol Group monographs, Vol. 2.

Versteegh, M.M., Rowen, D., Brazier, J.E. and Stolk, E., 2010. A Mapping onto EQ-5D for patients in poor health. *Health and Quality of Life Outcomes*, 8, pp.141.

Weijnen T.J., de Wit G.A., de Charro, F., 1998. An assessment of three applications of the EQ-5D in The Netherlands. *In* Greiner, W., Graf vad der Schulenburg, J.M., Piercy, J. eds., EuroQol Plenary Meeting. Hannover: Uni-Verlag Witte, pp.187-193.

Willke, C.T., Pickard, A.S., Walton, S.M., Moock, J., Kohlmann, T. and Lee, T.A., 2010. Statistical implications of utility weighted and equally weighted HRQL measures: an empirical study. *Health Economics*, 19(1), pp.101–110.

Wolfe, F. and Hawley, D.J., 1997. Measurement of the quality of life in rheumatic disorders using the EuroQol. *British Journal of Rheumatology*, 36(7), pp.786–793.

Xie, F., Li, S.C., Luo, N., Lo, N.N.,Yeo, S.J.,Yang, K.Y., Fong, K.Y., Thumboo, J., 2007. Comparison of the EuroQol and short form 6D in Singapore multiethnic Asian knee osteoarthritis patients scheduled for total knee replacement. *Arthritis and Rheumatism*, 57(6), pp.1043-1049.