Methods Used to Summarise Health State Preferences

Louise Longworth[1]

Katharine Johnston[2]

Andrew Kennedy[1]

[1] Health Economics Research Group, Brunel University, Middlesex

[2] Health Economics Research Centre, University of Oxford, Institute of Health Sciences, Headington, Oxford, OX3 7LF

Paper presented at the Health Economists' Study Group, University of Birmingham, January 1999

1. Introduction

Information on the costs and benefits of health care technologies is central to the process of allocating health care resources. The need to measure the benefits of health care has led to the development of preference based measures of health. Health state preferences are then combined with life years gained in order to assess the social value of health care technologies by means of cost utility analysis. In addition to informing resource allocation decision making at a macro level, health state preferences can also be used at a more micro level to inform the individual's treatment decisions.

In general, there are two different approaches to measuring health state preferences: direct and indirect. The direct approach involves eliciting preferences directly from patients. The main direct measurement techniques are standard gamble, time trade off and visual analogue (Torrance, 1986). These techniques require individuals to place values upon either described (hypothetical) health states or upon their own current health state. Direct approaches are usually administered by interview. Direct approaches valuing described health states require a decision about whose preferences should be used. Patients, professionals and the general public have been used to generate preferences.

The indirect approach is comprised of two stages. Firstly, patients classify their current health state according to a given descriptive classification system, defined in several dimensions and levels of severity within dimensions. This part of the indirect method is administered by self-complete questionnaire. Secondly, the patient's health state is translated into a preference measure by assigning it with a predetermined score. These scores for health states, also referred to as weights or tariffs, are themselves derived using direct methods from a different population. Preference scores for some of the potential health states within a classification system. Are derived directly and the scores for the remaining health states are then interpolated using modelling techniques. Indirect approaches require selecting a sample to generate the health states to be valued. The population on whom the tariffs or weights for health states are based is usually a representative sample of the general population.

There are several indirect methods, including the EuroQol 5D (EQ-5D)[i] and the Health Utilities Index (HUI). Their descriptive classification systems are comprised of five and eight dimensions respectively. The EQ-5D classification system assesses preferences in the following dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression (with three levels in each dimension). The latest version of the HUI classification system (HUI:3) assesses preferences in eight dimensions: vision, hearing, speech, ambulation, dexterity, emotion, cognition, self-care and pain (with five or six levels within each dimension). The tariffs for the EQ-5D were published in 1995 (Dolan *et al.* 1995) and since then they have increasingly been used as an indirect measure, particularly in clinical trials. The weights for the HUI:3 system have yet to be published.

There are counterbalancing advantages and disadvantages with either approach. The advantage of the indirect approach is that it is easier to administer on a non-interview basis. A further advantage of the indirect method is that when making comparisons between groups of patients, it can provide disaggregated evidence on any change in dimensions. The disadvantages of the indirect approach include that it relies heavily on the modelling techniques used which may, themselves be open to criticism. An additional feature of the indirect approach is that the tariff, or weight, attached to any single health state is a mean value around which there is measure of variability (or dispersion). This is an issue that is explored further in this paper.

As well as being used to evaluate health states and treatments, preference scores may be aggregated to form a 'social welfare function' from which decisions may be made systematically from a societal perspective. The use of a state as a decision maker implies that some individuals' views will be overridden in order to make decisions at the societal level, except in the unlikely event of complete unanimity. When a state is based upon collective choice, the problem arises of how to make decisions that reflect the collective preferences. This is often done through a system of voting, whereby individuals cast votes for issues or for people to act as representatives to themselves on issues.

The ability to formulate a social welfare function depends on the data type of the preference scores. Arrow's Impossibility Theorem proves that there is no satisfactory procedure for aggregating individual preferences to form a social preference from ordinal data[ii]. The condition that must necessarily be sacrificed in Arrow's theorem is that of non-dictatorship, which states that, one person's preferences must not take precedence over all others in all circumstances. For states based upon a notion of democracy, the aggregation of ordinal preferences for societal decision making can not therefore be reconciled with the ethos of the state. Perhaps more critically, it also implies that interpersonal comparisons of utility are not possible if measured as ordinal data. Thus, the use of these preference scores in formulating social welfare functions to aid societal decision making depends upon the nature of the scores, which may in turn depend upon the way in which they were elicited. The difficulties associated with making interpersonal comparisons are well recognised in the literature. In aggregating health state preferences, the approach used is to establish two clearly defined endpoints (0 as dead and 1 as good health) and use them as anchors on a scale. The basis for this aggregation is that the difference in utility between these two outcomes is set equal across individuals (Torrance and Feeny, 1989).

Once the preference scores have been derived, using either the direct or the indirect method, there are a range of issues that must be addressed in their aggregation, analysis and presentation. These issues include choosing a measure of central tendency and an associated measure of dispersion. Despite the extensive use of health state preferences in health economics, there appears to be no consensus as to the best way to handle these issues. By implication, conclusions may be drawn according to which summary measures have been used, which may in turn affect the decisions based upon them. Health state preferences form only a part of cost-utility analyses, and the issues surrounding the choice of summary measures for health state preferences also have relevance for other aspects. For example, the choice of whether to choose the mean or the median is also an issue when calculating Quality Adjusted Life Years (QALYs). This paper uses health state preferences as a starting point to explore these issues.

This paper has five related aims. Firstly, to identify the economic and statistical issues involved in the choice of methods used to summarise health state preferences, with particular emphasis placed upon the measures of central tendency and dispersion. Secondly, to identify the current practice in methods used to summarise health state preference aggregation and presentation from a selective review of the literature. Thirdly, to outline the implications of high levels of variability in preference scores. Fourthly, to illustrate these issues using an empirical demonstration. Finally, it also aims to explore whether the issues involved in choosing summary measures are different for direct and indirect methods of eliciting health state preferences.

Following this introduction, section 2 discusses measures of central tendency and section 3 examines measures of dispersion. Sections 4 and 5 address the implications of dispersion

and the distribution of health state preferences respectively. These issues are examined with respect to both the theoretical literature and current practice (as identified from the literature review). Section 6 presents an empirical analysis of the issues raised. Finally, section 7 discusses the issues raised and their implications. The appendix details the methods of the literature review.

2. Issues of Central tendency

There are a number of different measures of central tendency or location; the most commonly used are the arithmetic mean and the median. There are a number of issues that must be considered when choosing an appropriate measure of central tendency for a given set of data. These include the economic interpretation of the results of the different summary measures, and statistical considerations, such as the robustness of the measure and its capacity for further statistical manipulation. The measure of central tendency is often interpreted as an estimation of a 'true value' from a set of data.

2.1 Economic arguments for measures of central tendency

2.1.1 Mean

The mean has strong links with welfare economic theory as it encompasses the intensity of preference into its calculation. In welfare economics the notion of (Potential) Pareto Improvements is prominent when assessing the relative net benefit of projects (Per-Olov Johansson 1991). This theory guides the introduction of projects such that a move from one state to another should occur if the gainers from the move could (potentially) fully compensate the losers from the transition and still be better off. The compensating variation and equivalent variation methods of calculating welfare changes also involve intensity of preference (Per-Olov Johansson 1991). They calculate the minimum (maximum) amount of money that must be taken and given respectively, in order to leave a household at the same utility level before and after an increase (fall) in prices. Similarly to these theories, the mean aggregates the preferences of all involved and treats each individual's preference anonymously, thus embodying the intensity of preference in its measure.

2.1.2 Trimmed mean

Trimmed means eliminate a set proportion of the highest and lowest values in order to make the summary measure more robust. One implication of using a trimmed mean is that it disregards some individuals' preferences when calculating summary scores. This implies that those respondents with 'extreme' views do not have a valid opinion and should not have their preferences taken into account. Thus the trimmed mean cannot be a compatible measure with a decision making process that is based upon democracy with universal suffrage. This method of eliminating extreme values is generally not reciprocated in economic theory, as theories such as compensating and equivalent variation, or the Pareto principle, do not omit extreme values when examining welfare changes.

2.1.3 Median

The use of the median as a measure of central tendency also has economic implications stemming from its theoretical basis. The median is defined as the fiftieth percentile when all observations are ranked according to their magnitude. Thus each score is treated equally regardless of intensity of preference and is more aligned to a voting system based upon one-person one-vote. The median voter theory is usually expressed with reference to societal decision making to choose the most appropriate level of public expenditure. It assumes that political parties seek only to maximise votes and have no other form of self-interest such as ideology, and that preferences are transitive. Thus, in an electoral system based upon majority voting, each party will suggest a level of public expenditure that will

capture the vote of the median voter in order to maximise the number of votes cast in their favour. Where taxation is proportional the result is that policy is aimed towards the median voter who will be a member of the middle classes, thus the middle classes dictate policy towards the lower and upper classes. However, when the median voter theorem is applied to health state preference scores it is unclear as to who the median voter will be. There is no theoretical justification that the respondent with the median preference will necessarily be a member of the middle classes, and it may be that the class characteristics of the median respondent differs across different health states.

2.2 Statistical arguments for measures of central tendency

Robustness of a measure is desirable from a statistical perspective. Robustness implies that the measure will remain stable when different samples are drawn from the population of interest, thus producing a more accurate representation of the population's central tendency. However, there is no definitive test of robustness so it should be borne in mind that means are generally less affected than medians by many small fluctuations, and medians are robust when faced with extreme outlier observations. The median is also more robust when data is skewed.

Another desirable characteristic of a summary statistic is a clear algebraic definition. This enables further statistical treatment and to its expression with respect to other statistical identities. The mean has a clearer algebraic definition than the median. The mean also has a particular advantage over the median as the means of more than one sample population can be combined to form an overall mean of the merged populations using summary data alone.

From a statistical perspective the omission of some data from a sample, such as required by the trimmed mean, may be seen as an inefficient use of data. In contrast to the trimmed mean, neither the mean nor the median exclude preference scores in their calculation.

2.3 Current practice

From the selective literature review, the mean was identified as the most commonly used measure of central tendency. However, the median was also used, and both the mean and the median were presented together. The selective literature review also identified a study that used five per cent trimmed means in order to eliminate extreme outliers.

2.4 Summary

There are a number of issues surrounding the choice of measure of central tendency. The mean has strong links with economic theory, although the median can also be linked to the median voter theorem. The median is a more robust measure when data is skewed. However the mean has advantages when there is a need to aggregate and disaggregate data as its algebraic properties enable the means of more than one sample to be combined using summary data alone. The final choice of a measure should be a value judgement as to how intensity of preference should be accounted for.

When undertaking economic evaluations the mean is usually the most appropriate measure for cost data because of the direct link between mean average cost and total cost. Thus reimbursement for a good or service may be based upon average cost and related directly to actual expenditure as expressed as total cost. If median costs were to be used as an alternative, skewed data would result in unbalanced budgets. However, the widespread use of the mean as a summary measure in this area has developed because of the mean' algebraic properties rather than its economic interpretation.

Tariffs have been created in order to provide indirect ways of measuring health state preferences, some are based upon mean preference scores, but one uses median scores. Thus the issue of whether to use a mean or a median is also relevant when choosing which tariff to use. Dolan states that the choice between the mean and the median based tariffs for indirect utility measurement "... should ultimately be based upon a prior philosophical position on how preferences should be aggregated rather than on an intuition about which set of valuations seems to produce better answers."(Dolan 1997) However, Nord claims that the choice of tariff should be based upon how well the tariff scores "fit with the preferences that individuals or society as a whole express when asked directly"(Nord 1997).

3. Effect of Distribution

The distribution is the pattern that the ordered observations fall into. As stated previously, the distribution of observations will affect the robustness of the summary measures. The distribution of a sample may not be identical to the distribution of the population from which it came, however, as the sample size increases the distribution of the sample will tend to the distribution of the population.

The Normal, or Gaussian, distribution is a smooth, single humped distribution, which is symmetrical about the mean and continuous over the entire real line. Many natural occurrences follow a Normal distribution, particularly biological variables.[iii] Owing to the complete symmetry about the mean, in a Normal distribution all three measures of central tendency (mean, median and mode) produce identical results. Normally distributed data also has advantages in that it facilitates further statistical procedures such as sample size calculations.

A skewed distribution is not symmetrical as the frequency of observations decreases more quickly on one side of the distribution than the other. Skewness in a distribution shifts the value of central tendency but does not affect the measure of dispersion. Unlike a symmetrical distribution where the mean, median and mode are all equal, in a positively skewed distribution the mean is greater than the median, and the median is greater than the mode. However, if the distribution is negatively skewed the mode is greater than the median, and the median is greater then mean. If the distribution lies over a large interval it implies that there is a lack of consensus, and if there are many peaks to a distribution this again may imply that there is a lack of consensus as to how people value different health states.

Charts such as histograms or box plots can be presented in order to indicate the distribution of data at a glance. Where graphical representations are not available, summary tables can be used in order to infer the nature of the distribution. There are a number of measures that can be included in a summary table, some easier to interpret than others. For example, presenting the mean and median, and standard deviation or interquartile range gives more information as to the Normality of the data than the mean, standard error and standard deviation[iv]. In order to infer whether the distribution of the data is skewed from a summary table the reader could compare the magnitude of the mean and the median, or look at the position of the mean within the IQR. If the standard deviation is over twice the value of the mean it is also an indication of skewness in the data[v].

If the author wants to conduct hypothesis tests where the distribution of scores is not Normal, they are left with choice of using non-parametric tests that do not assume Normality, or transforming the data into a Normal distribution. Different transformations are used for different distributions where the sample data are skewed in different ways. However, transforming data alters the interpretation of scores, which may result in very

abstract concepts with little or no economic meaning. In a study by Patrick *et al*, utility scores were transformed to reduce the skew of the distribution. However, because of the transformation the authors asserted that "the transformed values should no longer be interpreted as utilities"(Patrick *et al* 1994). Logarithms are commonly used as the inverse of the transformation can be easily calculated, which avoids problems encountered when interpreting the results of analyses conducted on transformed data. Where the distribution is bi-modal it may not be possible to find a transformation capable of normalising the data (Buckingham *et al.*, 1996).

In the selective literature review, the studies that indicated the distributions of health state preferences found that the distributions were not Normally distributed. Most of these studies reported that the distributions were negatively skewed, however, a couple reported that severe health states were positively skewed. One study compared various instruments used for eliciting health state preferences, including the EuroQol EQ-5D and SF-36 (Essink –Bot *et al* 1997) It found that the distributions were skewed for all instruments, but that the scores produced by the EQ-5D indirect approach were the most heavily skewed. Of the studies that did give an indication of the distributions in the literature review, some provided a brief discussion of the distribution, and others used histograms or summary tables. However, most studies in the selective literature review did not discuss the distributions of health state preference scores.

4. Dispersion

The measure of central tendency only reflects the value an observation drawn from a population would be expected to take 'on average'. It gives no information about the range of values that an observation can take, thus it is useful to measure the level of dispersion of observations. There are different ways of measuring the dispersion (or variability) within a sample. The choice of measure of central tendency should dictate the choice of measure of dispersion used, such that the standard deviation should be used alongside the mean and the interquartile range alongside the median. The range may also be of interest when the most extreme values are of interest.

4.1 Standard deviation

The standard deviation is commonly used as it is a precise measure and can directly compare variability across different samples. It also exhibits certain statistical properties. One standard deviation either side of the mean encompasses 68 per cent of observations in the sample, 95 per cent by two standard deviations and over 99 percent in three standard deviations if the distribution is Normal (Munroe, BH 1997). If the distribution is not Normal, Chebyshev's theorem states that 75 per cent of observations will fall within two standard deviations of the mean.[vi] There are distinct advantages in using the standard deviation as a measure of dispersion. Like the mean, it can be expressed algebraically and thus its relationship with other measures can be expressed, for example with the mean or the root-mean-squared deviation.[vii] As with the mean, the robustness of this measure depends upon the distribution of the observations.

4.2 Interquartile range

The interquartile range (IQR) is a measure of variability around the median and is defined using percentiles. The IQR is defined as the seventy-fifth percentile minus the twenty-fifth percentile or (Q_3 - Q_1), and thus encompasses fifty percent of all observations. However, unlike the standard deviation, the IQR is not readily expressed in a pliable algebraic format. Also, the seventy-fifth and twenty-fifth percentiles are used only through convention, and there is no theoretical basis to this choice of percentiles. The extent of the robustness of the measures will depend upon the distribution of the sample.

4.3 Range

As the name suggests, the crude range, or range, is a very basic measure of dispersion and is found by subtracting the lowest value from the highest in a sample of observations. Its advantages are its speed and ease of calculation. However, it is likely to be very unstable when drawing different samples of data from the same population as the observations at either extreme of the distribution are likely to vary between samples. It does not take the distribution of the population into account, nor is it expressed in pliable algebraic form. The crude range is sometimes presented by stating the lowest and highest values of the data set. It is useful when the best and worst outcomes are of major interest, for example when the limits of possible values need to be known.

4.4 Current practice

The selective literature review identified the standard deviation as the most commonly used measure of dispersion, but the IQR and range were also used. It was also evident that some studies did not report a measure of dispersion, and some used the standard error in place of a measure of dispersion. The issues related to the standard error will be addressed in section four. In summary, the choice of measure of central tendency should dictate the choice of measure of dispersion used, so that the standard deviation is used with the mean, and the IQR with the median.

4.5 Variability as consensus

Dispersion is often thought of as the variability around a 'true value' of the population, estimated by a measure of central tendency. This 'true value' can be seen to represent the true value of a health state with deviations around it stemming from differences in expression or movement. However, it is also possible that there is no 'true value' when measuring health state preferences and that the dispersion or variability reflects the degree of consensus that exists about the severity of a health state. When the level of dispersion is high (low), there is a low (high) degree of consensus about that health state, implying that there is an array of valid scores that the question at issue may take. A lack of consensus, or variability within the data, may stem from people placing different utilities upon different aspects of health, or from asking a mixture of people who do and do not have experience of the health state being valued.

If this explanation of the variability is the case, and a natural variability exists within health state preference scores, there will be implications for interpersonal comparisons of health state preferences. If health states are valued differently by different people they cannot be represented by a single summary measure unless these differences can be controlled for. The measure of dispersion may then be deemed even more important than the measure of central tendency as it represents the spread of scores rather than a fictitious 'true value'.

When using indirect methods of eliciting health state preferences other issues arise. There will be variability within the tariffs themselves, as these consist of scores that have been reduced from a range of values into a single summary measure for each health state. There will also be variability within values obtained from respondents with regard to the health state that they are assessing, thus a 'double variability' may occur when using indirect methods of eliciting health state preferences. When a single summary measure is necessary, for example in QALY calculations, it would be advisable to conduct a sensitivity analysis in accordance with the level of dispersion to ensure that conclusions based upon such calculations are robust to the variability in health state preference values.

5. Implications of dispersion

Traditional statistical theory suggests that an increase in sample size would be expected to lead to a decreased level of dispersion, a view asserted by Torrance (Torrance 1986). If there is natural variability within a data set, increasing the sample size will not necessarily reduce the level of dispersion, and it is possible that, in non-Normal distributions, small samples may in fact conceal variability thus produce lower levels of dispersion. This was found in a recent study by Dolan *et al.* (1996), which formed the basis of the EuroQol tariffs, which found high levels of dispersion despite having a very large sample.

The dispersion within samples will have implications for other statistical issues, such as the choice of sample size. Collecting health state preferences from the whole population will usually be prohibitively expensive, therefore in most economic evaluations information is collected from samples. Samples are unlikely to be perfect replicas of the population and may be affected by measurement errors, such as selection bias in the sample. In order to eliminate bias in a sample and ensure that the sample is representative of the population, observations should be randomly selected. Larger samples may eliminate bias as increasing the size of a sample will tend its distribution towards that of the population. Although more likely to eliminate bias, large samples are costly. Therefore it is paramount that the sample size is kept to a minimum level for an efficient use of resources, yet be large enough to produce reliable and accurate results. A sample size calculation can be performed to find the minimum acceptable sample size to detect an economically significant difference in health state preferences between the treatments being compared. However in order to calculate it information on the variance must be known. This creates problems for research in new grounds as the variance will not be known until the study is complete. If a natural variability exists within the data on health state preferences, sample size calculations will result in large samples being necessary in order to establish statistically significant results. When hypothesis testing, larger samples are required in order to detect smaller differences between groups for a chosen outcome. Increasing the sample size will also lead to an increase in statistical power for detecting the same level of difference.

As stated above, sample statistics are subject to some degree of error owing to differences between samples and populations. The standard error (SE) measures the precision of the sample mean as an estimator of the population mean in random samples. The standard error is not a measure of dispersion of observations about the sample mean.[viii] By definition, the standard error and standard deviation are closely related. The standard error will always be smaller than the standard deviation and will decrease in size as the sample size increases by definition.

A few studies reported that they had conducted sample size calculations, with degrees of power and sizes of difference, however, the majority did not. It was evident from the selective literature review that some studies use the standard error in place of a measure of dispersion. However, other studies reported the standard error for its true purpose as they were comparing the reliability of different methods of eliciting health state preferences.

6. Empirical analysis

In order to demonstrate the issues that have been raised, a data set from an evaluation of orthopaedic services was interrogated. This data is from a randomised trial assessing the cost-effectiveness of an orthopaedic management service for the management of 'non-surgical' orthopaedic outpatients[ix]. The study included indirect EuroQol scores for self-

reported current health as an outcome measure collected at baseline, three months after treatment and twelve months after treatment. The data presented in tables below summarise the baseline scores purely as an example, and it must be noted that baseline data would not usually be compared between the two treatment groups.

Table 1: Summary of frequency statistics

Sample size:	1435		
Mean:	0.5265	Standard Deviation:	0.3092
Median:	0.6890	Interquartile range:	0.2280 to 0.7600
Trimmed Mean (5%):	0.5380	Range:	-0.349 to 1.000
		Standard Error:	0.0082

The summary statistics in table 1 show that the median is considerably higher than the mean as a measure of central tendency indicating negative skew (see figures below). The trimmed mean is only slightly higher than the mean which indicates that there are few outlier scores in the data.

Table 1 also shows that the data is negatively skewed as the mean is smaller than twice the standard deviation. By comparing the mean and the median, or either measure with the interquartile range, we can infer that the data is negatively skewed, because the mean is smaller than median, and both the mean and the median are close to the higher values of the IQR.

Histograms and box plots enable the distributions of the health state scores to be assessed at a glance. The box plot of EuroQol scores in figure 1 shows that the data is negatively skewed as the line representing the median is towards the top of the shaded box, and the tail line is longer towards the negative values. However, the box plot does not show any outliers or extreme values, which may indicate that the endmost negative values are not a result of idiosyncrasies in the data.

Figure 1: Box Plot of the EuroQol scores

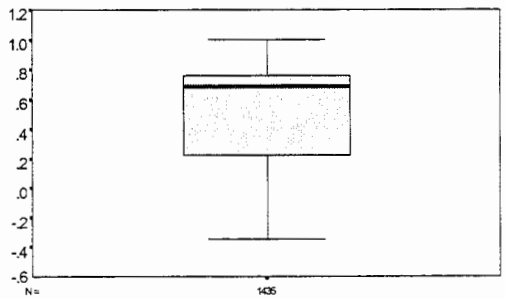
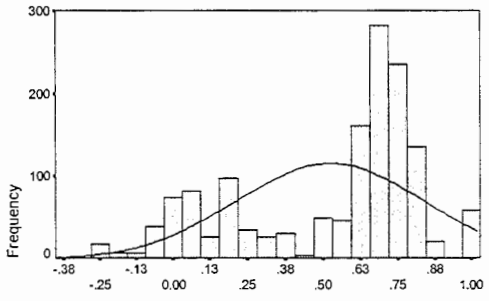


Figure 2: Histogram of the EuroQol scores



Despite the large sample size, the standard deviation and IQR are large indicating a lack of consensus in the data. Figure 2 shows a histogram of the EuroQol scores. The normal curve estimated over the histogram indicates that the data is negatively skewed. However, the bars representing the frequencies of scores show the data to be two humped or 'bimodal'.

In order to further examine the bimodal nature of the data, the data set was split in two at the point where the two distributions appeared to differ in the histogram (0.575). The two groups have been labelled moderate state and severe state. The main results are shown in table 2.

Table 2: Results from splitting the data to examine its potentially bimodal nature

	Sample size	Mean (SD)	Median (IQR)	Range
Severe state	495	0.431 (0.188)	0.124 (0.030 to 0.260)	-0.349 to 0.552
Moderate state	940	0.729 (0.95)	0.691 (0.689 to 0.760)	0.585 to 1.000

Histograms of the distribution of both sets of data are shown in figures 3 and 4. From the histograms it appears that both sets are Normally distributed and not heavily skewed, thus it is possible that this data set is bimodal. There are two possible explanations for this. Either two health states are being valued by one group of respondents with similar preferences, or there are two types of respondent that value aspects of health differently valuing the same health state.

Figure 3 : Histogram showing scores for severe state

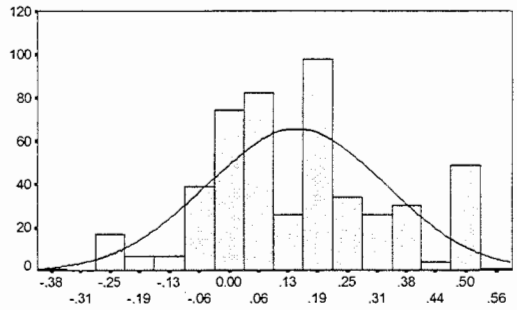
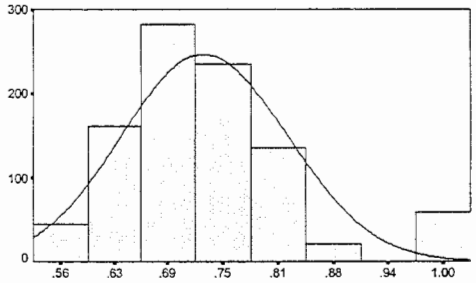


Figure 4: Histogram showing scores for moderate state



Of the possible methods of indicating the distribution of health state preferences, only the histogram picked up the possible bimodal nature of the data. The standard error of 0.0082 is small due to the large sample size of the study. This implies that the sample mean will not differ greatly from the population mean, but gives no indication of the dispersion of the sample.

As previously stated, health state preference scores are sometimes transformed to fit a Normal distribution. Logarithmic transformations are often favoured as the transformed scores are easily transformed back to their original state. The data was subjected to four transformations in accordance with recommendations in the literature for transforming negatively skewed data[x]. However, none of the transformations resulted in a distribution that appeared to be Normal.

The orthopaedic review data set was based upon a large sample size. For further analysis a random sample of the full data set was taken to assess the effect of sample size. The data set was reduced by randomly selecting approximately ten percent (142 cases) of the original sample using a random number generator. This size of sample is more in line with the

results of the literature review whose median (IQR) sample size was 104 (59 to 175), and mean (SD) size was 316 (691).

Table 3: Summary of results from a 10% random sample of the data set

Sample size:	142		
Mean:	0.5783	Standard Deviation:	0.2936
Median:	0.6910	Interquartile range:	0.2930 to 0.7600
Trimmed Mean (5%):	0. 5898	Range:	-0.2390 to 1.000
		Standard Error:	0. 0246

Reducing the sample size led to a very slight change in the median (from 0.689 to 0.691), but a larger change in the mean (from 0.527 to 0.578). This demonstrates that the median is more robust than the mean for skewed data. The trimmed mean also varied more than the median (from 0.538 to 0. 590) indicating that it is not more robust than the median in this case.

Standard statistical theory predicts that as the sample size is reduced the level of dispersion should increase in the data. Contrary to this, when comparing the standard deviation in the original sample (0.3092) to that of the reduced sample result (0.2936) it is evident that it has changed very little, in fact it has slightly decreased in the smaller sample. The IQR has decreased by a larger amount, from 0.2280 to 0.7600 in the original sample, to 0.2930 to 0.7600 in the ten per cent reduced sample. This may be due to the low degree of consensus in preferences towards the health state being valued. It is possible that an increase in the sample size does not necessarily lead to a decrease in the level of dispersion. Finally, in contrast to the measures of dispersion, the standard error has increased as the sample size has decreased. This is expected by definition of the standard error. The distribution of the data did not alter greatly as the sample size was reduced.

The empirical demonstration has shown that more information can be obtained by illustrating the distribution of scores than by presenting summary measures of central tendency and dispersion alone. If the data is has a bimodal nature it may raise issues as to the definition of the health state being valued or to the sample of respondents valuing the health state. It has also demonstrated that the various measures of central tendency can give different impressions of the data, shown by the difference in magnitude of the mean and the median. Of the different measures of dispersion only the range gave a clear indication that negative scores were contained in the data. The standard error was very small in accordance with the large sample size. Thus using the standard error as a measure of dispersion might lead to misleading interpretation of the expected range of values in the data.

In this instance, reducing the sample size led to a very slight change in the median, but a larger change in the mean. Both the standard deviation and interquartile range decreased when the sample size was reduced. This may be because the sample was based upon a broad range of values, and thus exhibited a lack of consensus.

The distribution of the scores was severely negatively skewed. This was confirmed by summary statistics, a histogram and a box plot. The data was subjected to a number of transformations in accordance with statistical convention, but none of these made the data conform to a more Normal distribution. The histogram indicated that the data might be bimodal, a proposition further substantiated after splitting the data for further analysis.

7. Discussion

This paper has demonstrated that there are wide differences in methods of summary and presentation of health state preferences. Inappropriate uses of summary measures can affect the conclusions drawn from the data. There are both economic and statistical issues involved in how health state preferences are aggregated and presented. Economic issues tend to favour the mean because of its links with welfare economic theory. However, with regard to statistical issues, the superiority is not so clear. The mean has advantages in terms of its algebraic properties, but the median is more robust when data is skewed, as it appears health state preference data often is. If valid conclusions are to be drawn from studies using quantitative data, the statistical methods used must be valid. Statistical methods used should generate data that is useful in an economic context and therefore appropriate to the decision making process. For informed decisions to be made consistently, a standardisation of methods used must occur.

If measures of central tendency and dispersion are not fully reported, the values placed upon health states may be under or overstated. The impact of this potential under or overstatement depends upon how the information is to be used. For example, preference scores may be attached to life years gained to estimate QALYs. If a mean or median value is presented alone, without any information on dispersion, it conceals the range of values and this will ultimately have implications for the magnitude of QALY gain. The effect of dispersion in health state values could, and should, be addressed by using the measure of dispersion in a sensitivity analysis and re-estimating the QALY gain. Where dispersion of values is high, the cost per QALY gained, and thus the value for money of a health care technology, may vary considerably.

In order to give an accurate presentation of results, both the mean and the median should be presented along with their associated measures of dispersion. This enables the reader to choose the measures that best meet their purpose and conduct sensitivity analyses where appropriate. If the most extreme values placed upon health states are of interest to the reader, the range may also be reported.

An indication of the distribution of values should also be presented so that the appropriate measure of location, dispersion, and method of statistical analysis can be chosen. The skew of data also has implications for the range of values used in the sensitivity analysis. Histograms impart a lot of information on the distribution of values at a glance, and are able to pick up more aspects of the data than summary tables and box plots, such as bimodality. However, owing to space restrictions in journals it may not be possible to include histograms for each state valued. If so, a brief discussion of the distribution or a summary table of results should be presented.

Sample size calculations should be conducted, where possible, to ensure the efficient use of data, however, if there is high variability in the data this will lead to larger sample sizes being necessary to achieve statistically significant conclusions. Standard errors should not be reported in place of measures of dispersion, but are of value if used to demonstrate the precision of the sample mean as an estimate of the population mean. If authors have reported standard errors and the sample size is known, the standard deviation can be calculated manually. As standard errors are inversely related to the sample size, small standard errors in large studies are to be expected. The effect of sample size upon the measures of dispersion is not so predictable. Statistical theory suggests that the level of dispersion should decrease as the sample size increases, but the results of studies in the selective literature review and empirical analysis do not necessarily support this assertion.

Variability within health state preference data may be due to random error around a true central value, or it may be due to natural differences about the value of a health state occurring. Natural differences about the values of health states have implications for the choice of whose values should be taken, and raise problems of making interpersonal comparisons. High levels of variability within samples of health state preference scores will have implications for other statistical procedures. Larger sample sizes will be necessary in order to detect statistically significant differences where there is a high variability within the data. The distribution of the data may impart more information on the nature of variability within samples. The empirical analysis demonstrated that the preferences in the data set had a bimodal characteristic, which may be due to two types of respondent valuing one health state, or because two types of health were being valued.

From the literature review it appears that most scores were negatively skewed, but some severe health states were positively skewed. Further research is required in order to assess whether there is a typical distribution of values for health states in relation to the severity, regardless of whose values are used and the methods used to elicit them. This may then impart information as to whether there is a 'true value' of preferences for all or some health states, or if there is a low degree of consensus. The differences in variability in health state preference scores derived from direct and indirect methods could also be examined further. The two-stage estimation process required by the indirect approach may result in increased variability and the nature of this variability, for example whether it is additive or multiplicative, could be examined. We are just starting to investigate the differences in the issues between direct and indirect methods of eliciting health state preferences, and it may be of interest to explore the variability in the tariffs used by examining the standard error and confidence intervals around the coefficients. Differences between the mean and the median values for health states could also be examined in order to establish if there is a systematic difference between them. The impact of using different summary measures in QALY gain calculations could be reviewed in order to establish how far they affect conclusions when preferences are aggregated in this way.

It is clear that there is a wide variation in the presentation of summarised or aggregated health state preferences in published studies, which may affect the conclusions drawn by the authors and readers. Also, it appears that some studies do not adequately present the results of health state preference valuations, by not indicating the distribution of scores, or by using the standard error in place of measures of dispersion rather than to indicate the precision of the mean. A set of guidelines is necessary if there is to be consistency in publications and allow cross comparison across studies. These guidelines should encourage the use of the measures outlined above, particularly with regard to the presentation of the mean and the median with their associated measures of dispersion, and an indication of the distributions of scores.

There is therefore a need for a uniform approach to the methods used to summarise health state preferences in order for consistency in health care decision making.

NOTES

i The EQ-5D can also be used to estimate preferences directly using a visual analogue scale but it is used and referred to in this paper in its indirect form.

ii KJ Arrow, 'A Difficulty in the Concept of Social Welfare', from 'Readings in Welfare Economics', Arrow KJ & Scitovsky T eds. 1969 Allen & Unwin Ltd.

iii JM Bland & DG Altman, 'Transforming Data", BMJ 312. 1996.

iv Altman DG & Bland JM, "Detecting Skewness from Summary Information", BMJ 313. 1996

v JM Bland & DG Altman, 'Transforming Data", BMJ 312. 1996

vi Chebyshev's theorem states, "For any set of data (population or sample) and any constant k greater than 1, at least 1 - 1/k2 of the data must lie within k standard deviations on either side of the mean" as quoted in Freund, JE, 'Modern Elementary Statistics', p.47.

vii For a demonstration of this relationship see Yule & Kendall, 'Introduction to the Theory of Statistics', pp.127

viii It is expressed as, SE = $\sqrt{}$(SD2 / N), where SD is the standard deviation and N is the sample size.

ix Leigh Brown AP, Kennedy ADM, Torgerson DJ, Campbell J, Webb JAG, Grant A. 'A randomised trial to assess the cost effectiveness of a orthopaedic medicine service for the management of "non-surgical" orthopaedic outpatients'. Journal of Epidemiology and Community Health (Abstract) 1997;51:603.

x Transformations included squaring and cubing as recommended by JM Bland & DG Altman, 'Transforming Data", BMJ 312. 1996, transformations of [-1 ÷ $\sqrt{}$X] and [-1 ÷ Log(X)] as recommended by Norman GR & Streiner DL, 'Biostatistics: The Bare Essentials'. Mosby 1994. pp208, and taking logarithms, natural logarithms and square roots

REFERENCES

Altman DG & Bland JM. 'Detecting Skewness from Summary Information'. BMJ 313. 1996

Arrow KJ & Scitovsky T eds., 'Readings in Welfare Economics', 1969. Allen & Unwin Limited

Bland JM & DG Altman. 'Transforming Data'. BMJ 312. 1996.

Buckingham JK, Birdsall B, Douglas JG. Comparing three versions of the time tradeoff: Time for a change? Medical Decision Making 1996; 16: 335-347.

Dolan P, 'Aggregating health state valuations' 1997, Journal of Health Services Research and Policy 2:3.

Dolan P, Gudex C, Kind P, Williams A. A social tariff for EuroQol: Results from a UK General Population Survey. Centre for Health Economics Discussion Paper 138, University of York, York, 1995.

Dolan P, Gudex C, Kind P, Williams. 'The Time Trade-Off Method: Results from a General Population Study', Health Economics 5. 1996

Freund JE. 'Modern Elementary Statistics', 1979. Englewood Cliffs, NJ Prentice Hall

Leigh Brown AP, Kennedy ADM, Torgerson DJ, Campbell J, Webb JAG, Grant A. 'A randomised trial to assess the cost effectiveness of a orthopaedic medicine service for the management of "non-surgical" orthopaedic outpatients'. Journal of Epidemiology and Community Health (Abstract) 1997;51:603.

Munro, BH. 'Statistical Methods for Health Care Research'. Lippincott-Raven Publications 1997.

Nord E. 'Comment: Aggregating health state valuations', 1997. Journal of Health Services Research and Policy 2:3

Patrick DL, Starks HE, Cain KC, Uhlmann RK, Pearlman RA. 'Measuring Preferences for Health States Worse than Death', Medical Decision Making 14:1. 1994.

Per-Olov Johansson. 'An Introduction to Modern Welfare Economics'. Cambridge University Press 1991.

Torrance GW. 'Measurement of health state utilities for economic appraisal' Health Economics 5 1986

Torrance GW , Feeny D. Utilities and quality-adjusted life years. International Journal of Technology Assessment in Health Care 1989 :5; 559-575

Yule G & Kendall M. 'Introduction to the theory of statistics', 1964. Charles Griffin & Company Limited

APPENDIX

Methods of literature review

A selective review of the literature was designed to investigate the current practice in presenting results from the elicitation of health state preferences. This aimed to identify the current practice in the use of different methods to summarise and present the data. The review was conducted using Medline 1995 and a hand search of recently published articles. The review included articles published between 1 January 1995 to 1 September 1998 in the following journals: Health Economics; International Journal of Technology Assessment in Health Care; Medical Care; Medical Decision Making; Social Science and Medicine. These journals were chosen as they are among the leading journals publishing studies measuring health preferences in health related quality of life.

The articles selected produced quantitative results from eliciting health state preferences from individuals using either a well established direct valuation method (standard gamble, time trade-off, visual analogue) or indirect method (EQ-5D, Health Utilities Index, Quality of WellBeing Scale).

The search produced thirty articles that met all the inclusion criteria. The studies had different objectives; valuing health states was a main priority for some but not others. Some studies elicited scores for the purpose of cost-utility analyses, others for resolving methodological issues. The majority of the studies elicited preferences directly rather than using a tariff for indirect methods. This may have been due to a large number of studies being primarily interested in the methodological issues surrounding the elicitation of health state preferences.

Summary of results of selective literature review

Table 1: Form of measurement of health status scores

	Number
Direct	25 (83.3%)
Indirect	2 (6.7%)
Both	3 (10%)
Total	30 (100%)

Table 2: Measures of central tendency

	Number
Mean	18 (60%)
Median	1 (3.3%)
Both	10 (33.3%)
None	1 (3.3%)
Total	30 (100%)

Table 3: Measures of dispersion

	Number
Standard deviation	16 (53.3%)
Standard deviation & IQR	2 (6.7%)
Standard deviation & Range	2 (6.7%)
Interquartile Range	2 (6.7%)
Range	2 (6.7%)
None	6 (20%)
Total	30 (100%)

Table 4: Summary of findings from the review

	Performed sample size calculation	Discussed the power of the study	Presented standard errors	Presented confidence intervals	Discussed the distribution(s)	Presented summary tables	Presented graphs
Yes	3 (10%)	4 (13.3%)	6 (20%)	3 (10%)	9 (30%)	5 (16.7%)	1 (3.3%)
No	27 (90%)	26 (86.7%)	24 (80%)	27 (90%)	21 (70%)	25 (83.3%)	29 (96.7%)
Total	30 (100%)	30 (100%)	30 (100%)	30 (100%)	30 (100%)	30 (100%)	30 (100%)

The main figures show the total number of studies reporting each item, and the figure in parenthesis shows the number of studies as percentage of the total.

Articles included in the literature review

Bleichrodt H, Johannesson M, An experience test of a theoretical foundation for rating-scale valuations., 1997, Medical Decision Making, 17 (2)

Bosch JL, Hunink MG, The relationship between descriptive and valuational quality-of-life measures in patients with intermittent claudication, 1996, Medical Decision Making, 16 (3)

Bult JR, Bosch JL, Hunink MG, Heterogeneity in the relationship between the standard-gamble utility measure and health status dimensions, 1996, Medical Decision Making, 16 (3)

Cairns J, Shackley P, Hundley V, Decision making with respect to diagnostic testing: a method of valuing the benefits of antenatal screening, 1996, Medical Decision Making, 16 (2)

Chapman GB, Elstein AS, Kuzel TM, Sharifi R, Nadler RB, Andrews A, Bennett CL, Prostate cancer Patients' Utilities for Health States: How it Looks Depends on Where You Stand, 1998, Medical Decision Making, 18 (3)

Dolan P, Gudex C., Time preference, duration and health state valuations., 1995, Health Economics, 4 (4)

Dolan P, Gudex C, Kind P, Williams A, The time trade-off method: results from a general population study, 1996, Health Economics, 5 (2)

Elvik R, The validity of using health state indexes in measuring the consequences of traffic injury for public health, 1995, Social Science and Medicine, 40 (10)

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

Goossens ME, Rutten-Van Molken MP, Kole-Snijders AM, Vlaeyen JW, Van Breukelen G, Leidl R., Health economic assessment of behavioural rehabilitation in chronic low back pain: a randomised clinical trial. 1998, Health Economics, 7 (1)

Jenkinson C, Gray A, Doll H, Lawrence K, Keoghane S, Layte R., Evaluation of index and profile measures of health status in a randomised controlled trial. Comparison of the Medical Outcomes study 36-Item Short Form Health Survey, EuroQol and disease specific measures, 1997, Medical Care, 35 (11)

Johnston K, Brown J, Gerard K, O'Hanlon M, Morton A, Valuing Temporary and Chronic Health States Associated with Breast Screening, 1998, Social Science and Medicine, 47 (2)

Krabbe PF, Essink-Bot ML, Bonsel GJ, The comparability and reliability of five health state valuation methods, 1997, Social Science and Medicine, 45 (11)

Krabbe PF, Essink-Bot ML, Bonsel GJ, On the equivalence of collectively and individually collected responses: standard gamble and time trade-off judgements of health states, 1996, Medical Decision Making, 16 (2)

Kuppermann M, Shiboski S, Feeny D, Elkin EP, Washington AE, Can preference scores for discrete states be used to derive preference scores for an entire path of events? An application to prenatal diagnosis. 1997, Medical Decision Making, 17 (1)

Lenert LA, Morss S, Goldstein MK, Bergen MR, Faustman WO, Garber AM. Measurement of the validity of utility elicitations performed by computerized interview, 1997, Medical Care, 35 (9)

Lenert LA, Cher DJ, Goldstein MK, Bergen MR, Garber A., The effect of search procedures on utility elicitations, 1998, Medical Decision Making, 18 (1)

Llewellyn-Thomas HA, Williams JI, Levy L, Naylor CD, Using a trade-off technique to assess patients' treatment preferences for benign prostatic hyperplasia, 1996, Medical Decision Making, 16 (3)

Nichol G, Llewellyn-Thomas HA, Thiel EC, Naylor CD, The relationship between cardiac functional capacity and patients' symptom-specific utilities for angina: some findings and methodological lessons, 1996, Medical Decision Making, 16 (1)

Patrick DL, Mathias SD, Elkin EP, Fifer SK, Buesching DP, Health State Preferences of Persons with Anxiety, 1998, Int. Journal of Tech. Health Care assessment, 14 (2)

Pinto Prades JL. 'Is the person trade-off a valid method for allocating health care resources?' 1997, Health Economics, 6 (1)

Revicki DA, Wu AW, Murray MI, Change in clinical status, health status, and health utility outcomes in HIV-infected patients., 1995, Medical Care, 33 (4 Suppl)

Rutten-Van Molken MP, Bakker CH, van Doorslaer EK, van der Linden S., 'Methodological issues of patient utility measurement. Experience from two clinical trials'. 1995, Medical Care, 33 (9)

Sculpher M, A cost-utility analysis of abdominal hysterectomy versus transcervical endometrial resection for the surgical treatment of menorrhagia, 1998, Int. Journal of Tech. Health Care assessment, 14 (2)

Stiggelbout AM, Eijkemans MJ, Kiebert GM, Kievit J, Leer JW, De Haes HJ. 'The 'utility' of the visual analog scale in medical decision making and technology assessment. Is it an alternative to the time trade-off?' 1996, Int. Journal of Tech. Health Care assessment, 12 (2)

Torrance GW, Feeny DH, Furlong WJ, Barr RD, Zhang Y, Wang Q., Multiattribute utility function for a comprehensive health status classification system. Health Utilities Index Mark 2., 1996, Medical Care, 34 (7)

Tsevat J, Solzan JG, Kuntz KM, Ragland J, Currier JS, sell RL, Weinstein MC, Health values of patients infected with human immunodeficiency virus. Relationship to mental health and physical functioning., 1996, Medical Care, 34 (1)

Ubel PA, Loewenstein G, Scanlon D, Kamlet M, Individual utilities are inconsistent with rationing choices: A partial explanation of why Oregon's cost-effective list failed., 1996, Medical Decision Making, 16 (2)

Unic I, Stalmeier PFM, Verhoef LCG, Van Daal WAJ, Assessment of the Time-trade-off Values for Prophylactic Mastectomy of Women with a Suspected Genetic Predisposition to Breast Cancer, 1998, Medical Decision Making, 18 (3)

Zethraeus N, Willingness to pay for hormone replacement therapy, 1998, Health Economics, 7 (1)