Measuring Preferences for Cost-Utility Analysis
How Choice of Method May Influence Decision-Making

Christine M. McDonough and Anna N.A. Tosteson
Dartmouth Medical School, Multidisciplinary Clinical Research Center in Musculoskeletal Diseases and the Center for the Evaluative Clinical Sciences, Lebanon, New Hampshire, USA

Abstract

Preferences for health are required when the economic value of healthcare interventions are assessed within the framework of cost-utility analysis. The objective of this paper was to review alternative methods for preference measurement and to evaluate the extent to which the method may affect healthcare decision-making. Two broad approaches to preference measurement that provide societal health state values were considered: (i) direct measurement; and (ii) preference-based health state classification systems.

Among studies that compared alternative preference-based systems, the EQ-5D tended to provide larger change scores and more favourable cost-effectiveness ratios than the Health Utilities Index (HUI)-2 and -3, while the SF-6D provided smaller change scores and less favourable ratios than the other systems. However, these patterns may not hold for all applications. Empirical evidence comparing systems and decision-making impact suggests that preferences will have the greatest impact on economic analyses when chronic conditions or long-term sequelae are involved. At present, there is no clearly superior method, and further study of cost-effectiveness ratios from alternative systems is needed to evaluate system performance.

Although there is some evidence that incremental cost-effectiveness ratio (ICER) thresholds (e.g. SUS50 000 per QALY gained) are used in decision-making, they are not strictly applied. Nonetheless, as ICERs rise, the probability of acceptance of a new therapy is likely to decrease, making the differences in QALYs obtained using alternative methods potentially meaningful.

It is imperative that those conducting cost-utility analyses characterise the impact that uncertainty in health state values has on the economic value of the interventions studied. Consistent reporting of such analyses would provide further insight into the policy implications of preference measurement.
As pressures to contain costs of medical care have escalated, cost-utility analysis (CUA) has received both critical acclaim and scrutiny as a methodology to inform decision-makers regarding the economic value of healthcare interventions. The number of published CUA has grown steadily over the past 20 years, and use by policy-makers appears to be increasing. In some jurisdictions, including the UK, Australia and the Canadian provinces of Ontario and British Columbia, a formal role for CUA in pharmaceutical coverage decisions is mandated. In other jurisdictions, such as the US, explicit use of CUA is more limited.

A key source of initial resistance to CUA was concern about the validity and comparability of results between studies. Issues of comparability have the potential to undermine a fundamental strength of CUA, which is meant to facilitate valid economic comparisons across a wide spectrum of interventions. Studies highlighting discrepancies between methods underscore the importance of understanding the potential impact that methodological differences may have on decision making.

Development of a reference case by the US Panel on Cost-Effectiveness in Health and Medicine provided one methodological standard for CUA. The perspective recommended for the reference case was societal, and the methods for valuing health outcomes included use of a generic health state classification system and community preferences, with sensitivity analysis to include patient preferences for studies of specific conditions. Since publication of the reference case, there is evidence of improving quality in CUA methodology and reporting. Support for use of reference case criteria to improve comparability is indicated by recent recommendations within the US regulatory environment. Formal guidelines in other jurisdictions demonstrate that, although differences exist, there may be emerging consensus on key points noted in the reference case. Despite these positive developments, the extent to which methodological differences in CUA may affect policy decisions remains uncertain. In this paper, we explore how alternative methods that fulfill reference case criteria for preference measurement, thereby yielding ‘societal health state values’, may affect decision making. We describe potential sources of variation in societal health state values, highlight relevant studies and discuss research assessing the decision-making impact.

1. Methods for Estimating Societal Health State Values

The QALY is the most commonly used measure for health in CUA. QALYs combine the attributes of length and quality of life (QOL) into a single measure. The length of time in each health state is weighted according to an associated ‘health state value’, on a scale with 1 representing best imaginable health and 0 representing death. In our discussion, we use ‘preference’ as a general term reflecting the desirability of a health state and ‘health state value’ (HSV) to connote the numerical strength of preference for a health state.

We consider two general approaches to HSV measurement: (i) direct preference elicitation for relevant health states; and (ii) preference-based health state classification systems.

1.1 Direct Preference Measurement

The standard gamble, time trade-off (TTO) and rating scale are commonly used preference elicitation methods with unique characteristics. Differences between preferences obtained by these methods are well documented, with HSVs typically highest for the standard gamble and lowest for the rating scale. One study of dialysis patients provided evidence that patient preferences from the standard gamble resulted in cost-effectiveness ratios that were higher by $US5916 per QALY than when TTO values were used. Although this is
one example of how measurement approach may affect cost-effectiveness results, individual patient rather than societal HSVs were used.[21]

Direct measurement of societal HSVs is a resource-intensive endeavour, requiring development of relevant health state descriptions and access to a representative population sample. Since this is often not feasible, researchers may use condition, age- and gender-specific values from a published source.[39]

1.2 Preference-Based Health State Classification Systems

Preference-based health state classification systems define each respondent’s health state based on a questionnaire, and assign a societal HSV with a scoring algorithm that incorporates preferences from a general population sample. This approach allows researchers to use societal HSVs with minimal resources compared with direct preference measurement.

The most widely used systems[19] include the EQ-5D,[40-42] the Health Utilities Index (HUI)[43,44] the Quality of Well-Being Scale (QWB)[45] and the SF-36-derived SF-6D.[46,47] The three basic steps in developing a system are (i) classifying health; (ii) eliciting population preferences for a subset of health states; and (iii) developing a scoring algorithm to assign values for the full range of health states. We address potential sources of variation and related research for each step.

1.2.1 Descriptive System

A generic self-report health status questionnaire is the basis for most systems. Most systems include attributes for pain, physical function, social or role function and anxiety/depression, but differ in the number of response levels and how these are described and weighted. Some systems include other attributes, such as hearing and vision. Other differences include the perspective used in assessing health status. For example, the HUI questions ask about respondents’ functional capacity,[43,48,49] while the EQ-5D and others ask about actual performance. The reference period also varies among systems, from ‘today’ to ‘over the past 4 weeks’.

Each instrument characterises a unique number of health states based on the numbers and levels of attributes included in the questionnaire. It is unknown how many health states are needed to describe health adequately; however, the EQ-5D has 243 health states, HUI-2 has 24 000 and HUI-3 has 972 000.

Taken together, these aspects of the descriptive systems would have an impact on the psychometric properties of the system, including the ability to measure health status numerically, and to detect meaningful changes in health. Inadequacy in the descriptive systems may result in ceiling and floor effects, inability to measure key attributes of health (validity) and inability to measure important change (responsiveness).

1.2.2 Preference Measurement

In the development of preference-based systems, a subset of the unique system-defined health states are valued by a sample of the population as the basis for estimating values for the full range of health states. Choice of direct preference elicitation method implicitly incorporates differences noted in section 1.1 into the estimation of each system’s societal HSVs. The EQ-5D offers value sets based on both TTO and rating scale measurements.[50] The SF-6D employs the standard gamble[46,47] and the QWB uses category scaling[45,51] methods. The developers of HUI and others, for example, argued that the rating scale is not appropriate for use in CUA. They used the standard gamble and a transformation of rating scale preference measurements for HUI-2 and -3.[30] Debate continues about the merits of various elicitation methods[52] and transformation[53,54] of rating scale values, v, into utilities, u, using a power curve such as $u = 1 - (1 - v)^2$.

Comparing CUA results using alternative preference measurement methods for the same system can
demonstrate the isolated impact of preference elicitation methods. A recent study comparing elicitation methods for the EQ-5D and SF-6D concluded that this component alone can contribute to differences in HSV of up to 0.31, and may impact on cost-effectiveness ratios.\textsuperscript{[55]} Conner-Spady et al.\textsuperscript{[56]} compared HSVs from the EQ-5D using the TTO and visual analogue scale (VAS) in 436 patients with joint replacements, and found a lower baseline mean, wider range, more negative HSVs and larger QALY gains for TTO than VAS-based value sets, indicating a more favourable cost-effectiveness ratio for TTO relative to VAS when using the EQ-5D. The QALY gain reported for TTO-based preference weights was 5.14 versus 3.64 for the VAS using a 10-year time horizon. If the cost of joint replacement were $US7000,\textsuperscript{[57]} the incremental cost-effectiveness ratio (ICER) would vary only slightly across methods (from $US1362 to $US1923 per QALY gained) and would be unlikely to influence decision-making.

### 1.2.3 Source of Community Preferences

The representativeness of population samples varies among systems.\textsuperscript{[58]} Furthermore, eliciting and scoring population preferences for health state classification systems are resource-intensive, and HSVs are not available for every population. Therefore, preferences from a different population from that of interest are sometimes used. The extent to which differences in population preferences would contribute to variation in CUA results has been explored. In a study comparing rating scale valuations of EQ-5D health states by Finnish and US general population samples, small differences were noted that the authors concluded would not impact EQ-5D HSVs in international studies.\textsuperscript{[59]} A study comparing TTO HSVs between general US and UK population samples\textsuperscript{[60]} reported higher HSVs for the US sample. Differences in EQ-5D TTO HSVs between UK and Spanish populations\textsuperscript{[61]} suggest that cultural differences may influence health state valuation.

### 1.2.4 Scoring Methods

Using directly measured preferences from a sample of the general population, a statistical model is fitted to estimate HSVs for the remaining health states. The HUI scoring system is based on multi-attribute utility theory\textsuperscript{[62]} and a multiplicative function that captures interactions among attributes and allows characterisation of single attribute utility functions for levels within each attribute.\textsuperscript{[36,63]} EQ-5D with York preference weights uses an additive model that includes level of severity, movement away from perfect health, and a term (the N3 term) to account for interaction between attributes when any attribute is at the worst level.\textsuperscript{[41,42]} Scoring algorithms have been developed for the EQ-5D for various populations. US TTO-based preference weights are now available for the EQ-5D (EQ-5D-US) using a random effects model.\textsuperscript{[64]} SF-6D scoring is very similar to that of the EQ-5D N3 model, with a smaller decrement when any attribute is at its worst level.\textsuperscript{[65]} The QWB utilises an additive function that does not allow for interaction among the attributes.\textsuperscript{[45]} The range of each scale and whether they include states worse than death (e.g. negative scores) varies.

Few studies compared CUA results using different scoring algorithms for the same system. Conner-Spady et al.\textsuperscript{[56]} compared HSVs and QALYs gains estimated by EQ-5D with and without the N3 term. They reported that the N3 terms resulted in lower baseline mean and effect size for values, and more QALYs gained in a sample of patients with joint-replacements.

### 2. Empirical Evidence

We searched the international published literature (See the Appendix for search terms), and found no studies that directly investigated the impact of preference measurement method on policy decisions. Though a systematic review of policy decisions was beyond the scope of this paper, we includ-
ed all identified studies that provided head-to-head comparisons of the most commonly used preference-based systems or addressed impact of system choice.

2.1 Comparisons of Preference-Based Health State Classification Systems

A review of 23 published cost-utility analyses conducted alongside clinical trials found that 20 studies utilised a preference-based health state classification system to estimate QALYs. The HUI and EQ-5D were the most commonly used systems, with 16 using the EQ-5D. The authors suggested that different systems could qualitatively impact CUA results, and called for greater reporting transparency.

Table I summarises studies that have reported cross-sectional comparisons between systems. These studies found varying differences in mean HSVs across systems and offer insight into system characteristics that may contribute to variation in HSVs. Generally, correlations between values from alternative systems were moderate to strong, indicating that they measure the same construct. The SF-6D demonstrated floor effects and a limited range of available scores. The EQ-5D did not provide HSVs between 0.88 and 1, and provided lower HSVs for similar health states than other systems. The HUI-3 was limited in characterising diminished mobility other than ambulation. These characteristics may be important for instrument choice relative to the condition of interest.

Evidence of differences in HSVs from cross-sectional studies is reason for concern, but longitudinal studies are necessary to understand system performance when measuring change in health. Fewer studies report longitudinal head-to-head comparisons of preference-based systems (table II). EQ-5D estimates were generally largest, followed by HUI-3, -2 and finally, SF-6D. Interactions have been noted between HSVs and level of improvement. Pickard et al. reported that change in SF-6D HSVs correlated with mental status, while change in HUI and EQ-5D related more to daily function and disability. No consistent pattern of correlation was evident in the studies we reviewed.

Comparisons of ICERs obtained using alternative systems were less common. Thomas et al. reported CUA results for chronic low back pain acupuncture treatment using the SF-6D and EQ-5D. The ICER for the SF-6D was £4241 (95% CI 191, 28,026) per QALY compared with £3598 (95% CI 189, 22,035) for the EQ-5D. Neumann et al. compared the results of a cost-effectiveness model for Alzheimer’s drug treatment using the HUI-2 and -3, and found lower mean utility scores using the HUI-3 versus HUI-2, resulting in ICERs of $US9000/QALY for the HUI-3 and $US11 000/QALY for the HUI-2 with a duration of drug effect of 18 months. The authors noted that, while the difference in ICERs was slight for this analysis of drug treatment, it could be substantial for a disease prevention drug.

Overall, the EQ-5D tended to have larger changes in HSVs, which would generally translate to more favourable cost-effectiveness ratios when using the EQ-5D compared with the HUI-2 or -3 (see table II). Similarly, in studies that included the EQ-5D and SF-6D, the EQ-5D had larger changes. Comparing the HUI-2 or -3 and SF-6D, the SF-6D tended to have smaller changes. These patterns are generally consistent with the results of cross-sectional comparisons, yet it is difficult to identify a superior system.

2.2 Potential Impact of System Choice on Decision-Making

Recent commentaries illustrate the concern that both lack of data on preferences and variation in methods will impact healthcare decisions. We found no empirical studies that address this question.
<table>
<thead>
<tr>
<th>Study (population)</th>
<th>Number of subjects</th>
<th>Systems studied</th>
<th>Baseline mean (SD)</th>
<th>Correlation coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Luo et al. [85]</td>
<td>4048</td>
<td>EQ-5D</td>
<td>0.87 (0.13)</td>
<td>0.87&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HUI-2</td>
<td>0.86 (0.32)</td>
<td>0.66&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HUI-3</td>
<td>0.81 (0.38)</td>
<td>0.87&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Petrou and Hockley [70]</td>
<td>14736</td>
<td>EQ-5D</td>
<td>0.84 (0.23)</td>
<td>0.70&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HUI-2</td>
<td>0.80 (0.15)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>HUI-3</td>
<td>0.79 (0.23)</td>
<td></td>
</tr>
<tr>
<td>Barton et al. [72]</td>
<td>915</td>
<td>EQ-5D</td>
<td>0.56 (0.15)</td>
<td>0.44&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SF-6D</td>
<td>0.77 (0.08)</td>
<td>0.50&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HUI-3</td>
<td>0.59 (0.22)</td>
<td>0.41&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>McDonough et al. [74]</td>
<td>2097</td>
<td>EQ-5D</td>
<td>0.39 (0.33)</td>
<td>0.68&lt;sup&gt;d&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HUI-2</td>
<td>0.59 (0.22)</td>
<td>0.78&lt;sup&gt;d&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HUI-3</td>
<td>0.45 (0.27)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SF-6D</td>
<td>0.57 (0.12)</td>
<td></td>
</tr>
<tr>
<td>Espallargues et al. [73]</td>
<td>209</td>
<td>EQ-5D</td>
<td>0.72 (0.22)</td>
<td>NR</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HUI-3</td>
<td>0.34 (0.28)</td>
<td>NR</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SF-6D</td>
<td>0.66 (0.14)</td>
<td>NR</td>
</tr>
<tr>
<td>Maia et al. [79]</td>
<td>313</td>
<td>EQ-5D</td>
<td>0.66 (0.13)</td>
<td>NR</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HUI-2</td>
<td>0.71 (0.19)</td>
<td>NR</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HUI-3</td>
<td>0.53 (0.29)</td>
<td>NR</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SF-6D</td>
<td>0.63 (0.24)</td>
<td>NR</td>
</tr>
<tr>
<td>Feeny et al. [71]</td>
<td>264</td>
<td>HUI-2</td>
<td>0.92 (0.13)</td>
<td>NR</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HUI-3</td>
<td>0.84 (0.16)</td>
<td>NR</td>
</tr>
<tr>
<td>Maddigan et al. [83]</td>
<td>372</td>
<td>HUI-2</td>
<td>0.78 (0.17)</td>
<td>NR</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HUI-3</td>
<td>0.64 (0.29)</td>
<td>NR</td>
</tr>
<tr>
<td>Luo et al. [85]</td>
<td>114</td>
<td>EQ-5D</td>
<td>0.75 (0.21)</td>
<td>0.45&lt;sup&gt;e&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HUI-3</td>
<td>0.76 (0.17)</td>
<td></td>
</tr>
<tr>
<td>O’Brien et al. [39]</td>
<td>246</td>
<td>HUI-2</td>
<td>0.61 (0.12)</td>
<td>0.58&lt;sup&gt;f&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HUI-3</td>
<td>0.68 (0.16)</td>
<td></td>
</tr>
<tr>
<td>Schulz et al. [81]</td>
<td>29</td>
<td>EQ-5D</td>
<td>0.81 (0.12)</td>
<td>0.32&lt;sup&gt;i&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HUI-3</td>
<td>0.79 (0.12)</td>
<td></td>
</tr>
<tr>
<td>Brazier et al. [85]</td>
<td>2605</td>
<td>EQ-5D</td>
<td>0.59 (NR)</td>
<td>0.66&lt;sup&gt;j&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SF-6D</td>
<td>0.63 (NR)</td>
<td></td>
</tr>
<tr>
<td>Stavem et al. [81]</td>
<td>397</td>
<td>EQ-5D</td>
<td>0.81 (NR)</td>
<td>0.78&lt;sup&gt;k&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15D&lt;sup&gt;e&lt;/sup&gt;</td>
<td>0.88 (0.12)</td>
<td></td>
</tr>
<tr>
<td>Belanger et al. [86]</td>
<td>1477</td>
<td>EQ-5D</td>
<td>0.83 (NR)</td>
<td>0.69&lt;sup&gt;l&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HUI-3</td>
<td>0.85 (NR)</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> 15D is a preference-based health state classification system for which data were reported compared with more commonly used systems. [81]

<sup>b</sup> Pearson’s correlation coefficient.

<sup>c</sup> Kendall’s tau for n = 863.

<sup>d</sup> Spearman’s rank correlation coefficient.

<sup>e</sup> This study reported values for patient groups: chronic obstructive pulmonary disease, osteoarthritis, irritable bowel syndrome, low back pain, leg ulcers, menopausal symptoms and aged >65 years.

<sup>f</sup> HUI = Health Utilities Index; NR = not reported.

by directly investigating policy decisions relative to different measurement methods. Studies addressing the potential influence of economic evaluation on policy decisions have considered thresholds for fa-
Measuring Preferences for Cost-Utility Analysis 99

3. Discussion

The paucity of published evidence limits our ability to draw conclusions concerning how preference measurement methods affect policy decisions. The ICERs available for comparison do not represent large enough changes to impact decision-making. However, evidence from comparisons of preference-based systems support a wide range of variation in estimates. Based on the evidence reviewed here, it appears that choice of preference measure may contribute to qualitatively different ICERs under some circumstances. As ICERs rise, the probability of acceptance appears to decrease, making the differences in QALYs obtained using alternative methods potentially meaningful. This is especially important for treatments with long-term consequences and ICERs around common thresholds. In our review, the EQ-5D tended to provide more favourable cost-effectiveness ratios than the HUI, while the SF-6D provided less favourable ratios than the other systems. Whether these patterns will hold for all applications depends on each system’s ability to measure change across the full range of health.

To assess the impact of HSV on CUA, our review focused on societal HSVs, estimated with the most commonly used preference-based health state classification systems. However, evidence shows that in practice, the majority of CUA do not yet meet reference case criteria. Evidence from reviews of published CUA indicates increasing use of preference-based systems and community preferences. Among CUA published between 1998 and 2001, 23% used preference-based systems and 36% used direct preference measurement such as standard gamble, TTO and rating scale. Utilities were community-based in 27% of estimates.

Arguments for using community preferences hold that the population potentially affected by the decision should be polled. From a position of uncertainty about their own future health (i.e., ‘the veil of
Table II. Studies addressing longitudinal comparisons in mean health state values provided by the most commonly used preference-based health state classification systems

<table>
<thead>
<tr>
<th>Study (population)</th>
<th>Subjects (n)</th>
<th>Systems studied</th>
<th>Baseline mean (SD)</th>
<th>Change in health state value mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Langfitt et al. [101] (chronic epilepsy)</td>
<td>64</td>
<td>EQ-5D</td>
<td>0.76 (0.26)</td>
<td>0.11 (0.28)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EQ-5D-US</td>
<td>0.82 (0.18)</td>
<td>0.07 (0.21)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HUI-2</td>
<td>0.78 (0.18)</td>
<td>0.03 (0.15)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HUI-3</td>
<td>0.61 (0.30)</td>
<td>0.08 (0.27)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SF-6D</td>
<td>0.70 (0.14)</td>
<td>0.08 (0.16)</td>
</tr>
<tr>
<td>Kaplan et al. [95] (rheumatoid arthritis)</td>
<td>628</td>
<td>EQ-5D</td>
<td>0.56 (0.25)</td>
<td>0.10 (0.27)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HUI-2</td>
<td>0.64 (0.20)</td>
<td>0.06 (0.18)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HUI-3</td>
<td>0.43 (0.27)</td>
<td>0.09 (0.24)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SF-6D (VAS)</td>
<td>0.43 (0.16)</td>
<td>0.06 (0.16)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SF-6D (SG)</td>
<td>0.81 (0.10)</td>
<td>0.03 (0.09)</td>
</tr>
<tr>
<td>Pickard et al. [87] (stroke)</td>
<td>98</td>
<td>EQ-5D</td>
<td>0.31 (0.38)</td>
<td>0.32 (0.38)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HUI-2</td>
<td>0.51 (0.20)</td>
<td>0.12 (0.23)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HUI-3</td>
<td>0.19 (0.30)</td>
<td>0.25 (0.32)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SF-6D</td>
<td>0.55 (0.09)</td>
<td>0.13 (0.15)</td>
</tr>
<tr>
<td>Slavem et al. [99] (HIV/AIDS)</td>
<td>60</td>
<td>EQ-5D</td>
<td>0.77 (0.26)</td>
<td>0.02 (NR)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SF-6D</td>
<td>0.73 (0.17)</td>
<td>0.03 (NR)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15D</td>
<td>0.86 (0.14)</td>
<td>0.02 (NR)</td>
</tr>
<tr>
<td>Thoma et al. [93] (breast reduction surgery)</td>
<td>41</td>
<td>HUI-2 (NR)</td>
<td>0.62 (0.19)</td>
<td>0.22 (0.07)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HUI-3 (NR)</td>
<td>0.63 (0.29)</td>
<td>0.15 (0.28)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SF-6D</td>
<td>0.67 (0.12)</td>
<td>0.08 (0.13)</td>
</tr>
<tr>
<td>Hatoun et al. [88] (coronary artery disease)</td>
<td>184</td>
<td>HUI-2</td>
<td>0.51 (0.20)</td>
<td>0.12 (0.23)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HUI-3</td>
<td>0.19 (0.30)</td>
<td>0.25 (0.32)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SF-6D</td>
<td>0.55 (0.09)</td>
<td>0.13 (0.15)</td>
</tr>
<tr>
<td>Feeny et al. [89] (total hip arthroplasty)</td>
<td>63</td>
<td>HUI-2</td>
<td>0.62 (0.19)</td>
<td>0.22 (0.07)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HUI-3</td>
<td>0.62 (0.32)</td>
<td>0.23 (0.04)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SF-6D</td>
<td>0.81 (0.10)</td>
<td>0.10 (0.02)</td>
</tr>
<tr>
<td>Holland et al. [98] (elderly patients)</td>
<td>123</td>
<td>EQ-5D</td>
<td>0.61 (0.29)</td>
<td>−0.16 (0.34)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AQoLb</td>
<td>0.45 (0.27)</td>
<td>−0.12 (0.24)</td>
</tr>
<tr>
<td>Longworth and Bryan [96] (liver transplant)</td>
<td>183</td>
<td>EQ-5D</td>
<td>0.52 (0.33)</td>
<td>0.09 (0.37)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SF-6D</td>
<td>0.61 (0.12)</td>
<td>0.01 (0.28)</td>
</tr>
<tr>
<td>Bosch et al. [94] (intermittent claudication)</td>
<td>87</td>
<td>HUI-2</td>
<td>0.70 (0.20)</td>
<td>0.10 (NR)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HUI-3</td>
<td>0.66 (0.20)</td>
<td>0.11 (NR)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SF-6D</td>
<td>0.66 (0.09)</td>
<td>0.08 (NR)</td>
</tr>
<tr>
<td>Bosch and Hunink [92] (intermittent claudication)</td>
<td>88</td>
<td>EQ-5D</td>
<td>0.57 (0.25)</td>
<td>0.22 (NR)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HUI-3</td>
<td>0.66 (0.20)</td>
<td>0.11 (NR)</td>
</tr>
<tr>
<td>Suarez-Almazor et al. [93] (low back pain)</td>
<td>37</td>
<td>EQ-5D</td>
<td>0.38 (0.33)</td>
<td>−4.8 (17.4)c</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HUI-2</td>
<td>0.49 (0.19)</td>
<td>−1.8 (16.1)c</td>
</tr>
</tbody>
</table>

15D is a preference-based health state classification system for which data were reported compared with most commonly used systems.[81]

AQoL is a preference-based health state classification system for which data were compared with most commonly used systems.[76,98]

3-month results, rescaled to 1–100, with 100 corresponding to best health.

AQoL = Assessment of Quality of Life; HUI = Health Utilities Index; NR = not reported; SG = standard gamble; VAS = visual analogue scale.
dence that patients provide higher values than other groups for their health states, and the possibility that this could cause undervaluation of preventive intervention and treatment. In addition, there is evidence that differences in valuations between patients and other groups varies with other factors, such as severity and chronicity of illness.

As a practical matter, the choice of methods may be dictated by resource availability, making off-the-shelf tools attractive. To facilitate CUA when primary data are not available, the Panel on Cost-Effectiveness in Health and Medicine recommended construction of a national catalogue of 'off-the-shelf' community preferences for health states. Work is ongoing in this area and includes population-based HSVs using multiple instruments, mapping of health status data to HSVs and catalogues or registries from the published literature. Several methods have been explored to estimate HSVs from the SF-36, the SF-12, condition-specific tools and national survey instruments. Some studies indicate that estimates from mapping may have important limitations while others emphasise the fairly strong correlations between approaches. A study investigating potential decision-making impact of algorithm choice compared ICERs from various SF-36 and SF-12 algorithms and the SF-6D. They reported ICERs for an asthma cohort ranging from $US30 769 to $US63 492 per QALY, and for a stroke cohort of $US27 972 to $US50 000 per QALY. Many of the reported ICERs did not have overlapping confidence intervals. The growth of methods for estimating preferences using existing data is expanding the range of possible CUA applications, and warrants continued comparative study. In addition, efforts have been made to make published CUA more accessible through registries, some of which include information on methodology. These efforts are important steps in improving availability and transparency of CUAs.

Finally, the impact of preference measurement method should be considered in the context of the role of CUA in policy decisions. Empirical evidence is limited; however, one study indicated that economic evaluation in the US has a larger impact on decisions at the organisational level than population and patient/provider levels. Moreover, it appears that economic evaluation is used to varying degrees in different organisations in conjunction with legal, political and regulatory considerations.

4. Conclusion

It is unknown to what extent the choice of preference measurement method impacts health policy decision-making. The existing evidence points toward potential impact for a subset of situations, most likely for ICERs near established thresholds when chronic diseases and/or long-term health effects are involved. At present, there is no clearly superior method for estimating societal preferences; however, alternative systems produce a wide range of HSVs. The ability to convert HSVs from one system to another would greatly improve comparability of CUAs. A recent publication of ongoing research using MEPS (Medical Expenditure Panel Survey) and National Health Interview Survey data uses linear regression to standardise estimates from various systems, and to convert estimates from one system to those of another. Related research provided population-based norms. Ongoing work will further our understanding of system responsiveness.

Psychometric comparisons between systems were more common than comparisons of impact on ICERs and/or policy decisions. Explicit study of cost-effectiveness ratios obtained from alternative preference-based systems is needed to improve our understanding of potential policy implications. We
encourage researchers who have access to comparative data to publish these estimates. Sensitivity analyses are suggested to address uncertainty associated with estimates of health effect, including variation due to measurement method. An in-depth systematic review of decisions made by policy makers relative to alternative measurement method would be a timely and important contribution to the literature. Research in these areas combined with ongoing dialogue about standardisation and conversion of estimates hold promise for attaining consistency of estimates for CUA.

Acknowledgements

This paper was funded by the National Institute of Arthritis and Musculoskeletal and Skin Diseases (P60-AR048094) and the National Institute on Aging (R01-AG12262). The authors have no conflicts of interest that are directly relevant to the content of this article.

Appendix

1. Literature Search Strategy

The following key words were used in Medline to identify papers addressing the topic of preference measurement method and policy decision-making. In addition, selected authors, references from papers known from previous work and known websites were hand-searched. Searches were limited to English language and carried out between December 2005 and June 2006.

- Cost-utility analysis
- Cost-effectiveness analysis
- Economic evaluation
- Methods
- Attitude to health
- Health status indicators
- Health status
- Quality of life
- Quality-adjusted life years
- QALYs
- Time trade off
- Standard gamble
- Rating scale
- Visual analogue scale
- Utilities
- Health state values
- Values
- Preferences
- Preferences for health states
- Preference classification systems
- Preference-based
- EQ-5D
- HUI
- SF-6D
- SF
- Quality of well-being.

References

1. Luce BR. What will it take to make cost-effectiveness analysis acceptable in the United States? Med Care 2005; 43 (7 Suppl.): I44-8


52. Parkin D, Devlin N. Is there a case for using visual analogue scale valuations in cost-utility analysis? Health Econ 2006; 15: 653-64
54. Stevens KJ, McCabe CJ, Brazier J. Mapping between visual analogue scale and standard gamble data: results from the UK Health Utilities Index 2 valuation survey. Health Econ 2006 May; 15 (5): 527-33
55. Tsuchiya A, Brazier J, Roberts J. Comparison of valuation methods used to generate the EQ-5D and the SF-6D value sets. J Health Econ 2006; 25: 334-46
56. Conner-Spady B, Voaklander DC, Suarez-Almazor ME. The effect of different EuroQol weights on potential QALYs gained in patients with hip and knee replacement. 17th Plenary Meeting of the EuroQol Group; 2000 Sep 28-29; Pamplona
63. Torrance GW. Preferences for health outcomes and cost-utility analysis. Am J Manag Care 1997; 3 Suppl.: S8-20
65. Brazier J, Roberts J, Tsuchiya A. A comparison of the EQ-5D and SF-6D across seven patient groups. Proceedings of the 18th Plenary Meeting of the EuroQol Group; 2001 Sep 6-7; Odense
70. Petrov S, Hockley C. An investigation into the empirical validity of the EQ-5D and SF-6D on hypothetical preferences in a general population. Health Econ 2005; 14: 1169-89
88. Hatoum IT, Brazier JE, Akhras KS. Comparison of the HUI3 with the SF-36 preference based SF-6D in a clinical trial setting. Value Health 2004; 7 (5): 602-9
90. Conner-Spady B, Suarez-Almazor ME. A comparison of preference-based health status tools in patients with musculoskeletal disease. 18th Plenary Meeting of the EuroQol Group; 2001 Sep 6-7; Odense
117. Soucek J, Byrne MM, Kelly PA, et al. Valuation of arthritis health states across ethnic groups and between patients and community members. Med Care 2005; 43 (9): 921-8
120. Sullivan PW, Lawrence WF, Gushchyan V. A national catalogue of preference-based scores for chronic conditions in the United States. Med Care 2005; 43 (7): 736-49
136. Lawless W, Fienberg S. Predicting EQ-5D preference scores from the SF-12: Health Survey in a nationally representative sample. Med Decis Making 2004; 24 (2): 247-54
137. Lawrence WF, Fiehley G. Predicting EuroQol EQ-5D preference scores from the SF-12 Health Survey in a nationally representative sample. Med Decis Making 2004; 24 (2): 160-9
143. Feeny D. As good as it gets but good enough for which applications? Med Decis Making 2006; 26: 307-9
155. Feeny D. As good as it gets but good enough for which applications? Med Decis Making 2006; 26: 307-9