Comparing Cost-Utility Estimates
Does the Choice of EQ-5D or SF-6D Matter?

Tracey H. Sach, PhD,* †‡ Garry R. Barton, PhD, † Claire Jenkinson, BSc, † Michael Doherty, MD, §
Anthony J. Avery, MD, † and Kenneth R. Muir, PhD †

Background: A number of different measures can be used within cost-utility analyses, we compared results according to both the EQ-5D and SF-6D.

Methods: A randomized trial was conducted to compare 4 options for people with knee pain. Over the 2 year trial period, the change in cost to health-service was estimated, and both the EQ-5D and SF-6D were used to estimate the change in quality-adjusted life years (QALYs). Using a complete case analysis, the cost-utility (incremental cost-effectiveness ratio [ICER]) of each option, according to both the EQ-5D and SF-6D, was calculated and assessed in relation to the cost-effectiveness threshold of £20,000 per QALY.

Results: Of the 389 participants, 247 had complete cost, EQ-5D and SF-6D data. According to the EQ-5D, option 1 had an estimated ICER of £10,815 (compared with option 4), option 2 was dominated by option 1, and option 3 was subject to extended dominance. Conversely, according to the SF-6D, option 3 had an ICER of £9999 (compared with option 4), option 2 had an ICER of £36,883 (compared with option 3), and option 1 was subject to extended dominance.

Conclusion: The EQ-5D and SF-6D estimated that different options (1 and 3, respectively) were cost-effective at the £20,000 per QALY threshold, demonstrating that the choice of measure does matter.

Key Words: cost-utility, cost-effectiveness, EQ-5D, SF-6D

As resources are scarce economists have developed a common utility scale (where 0 equates to death and 1 to full health) on which to compare the benefits of many interventions.¹ Most commonly, utility scores are estimated using different health state descriptive measures such as the EQ-5D,² health utilities index,³ or SF-6D.⁴ However, as these measures are based on both different health state descriptions,⁵ and valuation techniques,⁶ they have been shown to produce different utility scores for the same group of patients.⁷–¹¹

Though a number of studies have compared the utility scores produced by different measures, few studies have assessed the impact that the choice of measure has on estimates of cost-utility, and it has been argued that further research is required in this area.¹²,¹³ This article aims to fulfil this remit by comparing cost-utility estimates for both the EQ-5D and SF-6D using data from a randomized trial which compared 4 options for overweight and obese people who had knee pain.

METHODS

Participants
All participants were taking part in the Lifestyle Interventions for Knee Pain (LIKP) study—a randomized controlled trial which was designed to compare the effectiveness and cost-effectiveness of 4 different options (1) diet and strengthening exercise advice, (2) dietary advice, (3) strengthening exercise advice, and (4) leaflet provision (hereafter these options are referred to as 1, 2, 3, and 4 as the main focus of this article is methodological). All registered patients in 5 UK general practices were recruited if they were aged ≥45 years, reported knee pain on most days of the last month, had a body mass index >28.0 kg/m², and gave consent to be randomized to one of the aforementioned 4 options.

Estimating Costs
The methods used to estimate the change in cost to the health service for each participant in the LIKP study have been described elsewhere,¹⁴ briefly they were as follows. Three cost components were monitored—visits by the health care professionals providing each option, provision of exercise bands to participants in either of the strengthening exercise groups (options 1 and 3), and the change in the cost of analgesics. All costs were estimated over the 2 year trial period at 2006/7 price levels, costs incurred in the second year were discounted a rate of 3.5%,¹⁵ and multiple imputation was used where data were missing.¹⁶

Estimating Utility
Participants were asked to complete both the EQ-5D and SF-6D at baseline, 6, 12, and 24 months postrandomiza-
tion. With the EQ-5D the respondent is asked to report the level of problems they have (no problems, some/moderate problems, and severe/extreme problems) with regard to mobility, self-care, usual activities, pain/discomfort, and anxiety/depression.\(^2\) Up to 243 different health state descriptions can be derived from responses to these 5 dimension, and utility scores were assigned to each of these states using the York A1 tariff.\(^17\) SF-6D\(^4\) scores were derived from responses to eleven questions on the SF-36\(^18\) questionnaire. The SF-6D has 6 dimensions (physical functioning, role limitations, social functioning, pain, mental health and vitality) which have between 4 and 6 severity levels. Scores for the 18,000 potential health states were estimated using the consistent version of the SF-6D algorithm.\(^5\)

**Estimating Quality-Adjusted Life Years**

Using the area under the curve method (with adjustment for baseline differences)\(^19\) the EQ-5D and SF-6D scores at baseline, 6, 12, and 24 months postrandomization were used to estimate the QALY gain (or loss), which accrued over the 2 year trial period for each measure (QALY scores in the second year were discounted at 3.5%).

**Analyses**

To compare the EQ-5D and SF-6D results, a complete case analysis approach\(^20\) was used. Thereby, only those participants who had an estimated QALY gain according to both measures were included in all subsequent analyses. Summary statistics were calculated, where these included the mean change in cost and the mean QALY gain for each of the 4 options.

**Cost-Utility**

For each of the 4 options estimates of cost-utility were calculated separately for both the EQ-5D and SF-6D (from the perspective of the health service). Dominated options (those with both a higher mean change in cost and lower mean QALY gain, compared with another option) were excluded, as were options which were subject to extended dominance (as combinations of other options could provide a greater benefit at equivalent cost).\(^21\) Subsequently, the cost-utility of nonexcluded options was calculated by estimating the incremental cost per QALY gain (incremental cost-effectiveness ratio [ICER]) associated with each option, relative to the next best alternative. Finally, in line with guidance by the UK National Institutes of Health and Clinical Excellence,\(^22\) we sought to identify the most cost-effective option at the threshold (\(\lambda\)) of £20,000 per QALY.

**Decision Uncertainty**

To depict the level of uncertainty associated with the decision as to which option was cost-effective, according to each of the EQ-5D and SF-6D, we constructed the cost-effectiveness acceptability curve (CEAC)\(^23,24\) and the expected value of perfect information (EVPI).\(^25,26\) The CEAC depicts the probability that an option is cost-effective at different levels of \(\lambda\) (ie, according to how much one is willing to pay for a QALY gain). It was constructed using the technique of nonparametric bootstrapping,\(^27\) where 10,000 simulations of the (per participant) cost and effect were drawn for each option (with replacement) from the original cost and effect data. The probability of being cost-effective was then equivalent to the proportion of the 10,000 simulations for which each option had the highest net benefit\(^28\) (net monetary benefit \(= E \times \lambda - C\), where \(E\) is the QALY gain and \(C\) is the cost) at different values of \(\lambda\).

When making decisions about resource allocation, in addition to estimating the cost-effectiveness of particular options, it has been argued that one should consider whether further research should be undertaken to reduce the level of uncertainty associated with that decision.\(^29\) The EVPI addresses this by estimating the upper bound of the value of undertaking further research.\(^26\) As outlined previously,\(^23\) we estimated it at a \(\lambda\) of £20,000 per QALY by first calculating the value of perfect information in each of the aforementioned simulations (the difference between the highest net benefit across each of the 4 options and the net benefit of the most cost-effective option). Second, the EVPI (per patient) was calculated by taking the expectation of the value of perfect information across each simulation.

**RESULTS**

**Participants**

After sending out 12,500 questionnaires, and conducting a local media campaign, a total of 389 participants were both eligible and consented to take part in the LIKP study.

**Analyses**

The EQ-5D was completed at baseline, 6, 12, and 24 months postrandomization by 270 participants, 254 completed the SF-6D at each of these time points, and 247 completed both measures. Of these, 67 received option 1, compared with 78 for option 2, 47 for option 3, and 55 for option 4. Summary statistics for each option are provided in Table 1. The estimated mean change in cost over the 2 year trial period (for each option), ranged between £821.93 (the negative value can be explained by the fact that those who received option 4 (leaflet provision) did not receive any related health care visits, but had lower mean analgesic costs at 24 months postrandomization, compared with baseline).

For each option the mean EQ-5D score was higher postrandomization, compared with baseline, where the mean QALY gain over the 2 year trial period ranged between 0.007 (option 3) and 0.150 (option 1) (mentioned in Table 1). The mean SF-6D scores also improved, but to a lesser extent as the mean QALY gain ranged between 0.021 (option 4) and 0.065 (option 2) (mentioned in Table 1). The rank ordering of options (from most to least effective) was however different according to the EQ-5D\(^1,2,4,3\) and SF-6D\(^2,1,3,4\).

**Cost-Utility**

By comparing the mean change in cost and mean QALY gain according to the EQ-5D across each option (mentioned in Table 1) it was apparent that option 2 was dominated by option 1, similarly option 3 was dominated by option 4. Compared with option 4, option 1 had a mean incremental cost of £794.90 and a mean incremental QALY gain of 0.074, giving an ICER of £10,814.74. According to
the SF-6D option 1 was subject to extended dominance as combinations of options 2 and 3 could provide a higher benefit at lower cost, option 3 was estimated to have an ICER of £9999.39, (compared with option 4), and option 2 was estimated to have an ICER of £36,882.58 (compared with option 3) (mentioned in Table 1). Thus, at a λ of £20,000 per QALY option 1 was estimated to be cost-effective according to the EQ-5D, compared with option 3 according to the SF-6D.

### Decision Uncertainty

The CEACs which were based on the EQ-5D are plotted in Figure 1. It can be seen that at low levels of λ (<£5000 per QALY) option 4 was estimated to have the highest probability of being cost-effective. However, for values >£5000 per QALY the probability of each option being cost-effective was <30%. This demonstrates that, according to the EQ-5D, there is much uncertainty associated with the decision as to which option is most cost-effective. The CEACs derived from on the SF-6D are plotted in Figure 2, and portray a similar story. With regard to the EVPI at a λ of £20,000 per QALY it was estimated to be £7195 per patient according to the EQ-5D, and £3219 according to the SF-6D (this result is explained further in the Discussion).

### DISCUSSION

We have shown that, when comparing the cost-effectiveness of 4 different options for people with knee pain, the EQ-5D and SF-6D estimated that different options (1 and 3, respectively) were most cost-effective.

---

**TABLE 1.** Estimates of the Mean QALY Gain, Mean Change in Cost and the Incremental Cost Effectiveness Ratio According to the EQ-5D and SF-6D

<table>
<thead>
<tr>
<th>Option</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EQ-5D</td>
<td>SF-6D</td>
<td>EQ-5D</td>
<td>SF-6D</td>
</tr>
<tr>
<td>Baseline utility score</td>
<td>0.535</td>
<td>0.649</td>
<td>0.544</td>
<td>0.643</td>
</tr>
<tr>
<td>6 mo utility score</td>
<td>0.657</td>
<td>0.699</td>
<td>0.626</td>
<td>0.697</td>
</tr>
<tr>
<td>12 mo utility score</td>
<td>0.599</td>
<td>0.670</td>
<td>0.608</td>
<td>0.680</td>
</tr>
<tr>
<td>24 mo utility score</td>
<td>0.620</td>
<td>0.674</td>
<td>0.602</td>
<td>0.667</td>
</tr>
<tr>
<td>Mean QALY gain</td>
<td>0.150</td>
<td>0.052</td>
<td>0.116</td>
<td>0.065</td>
</tr>
<tr>
<td>Mean change in cost</td>
<td>£776</td>
<td>£822</td>
<td>£268</td>
<td>-£19</td>
</tr>
<tr>
<td>Incremental cost per QALY (ICER)</td>
<td>£10,815</td>
<td>ED</td>
<td>D</td>
<td>£26,883</td>
</tr>
</tbody>
</table>

* ED indicates dominated; ED, subject to extended dominance.
Utility scores on the SF-6D have a narrower range (0.296 to 1.00), compared with the EQ-5D (0.594 to 1.00). Additionally, time trade-off valuation scores (the technique used with the EQ-5D) tend to be higher for milder states, but lower for more severe states, compared with standard gamble scores (which are used with the SF-6D). As a result, it has been proposed that estimated utility gains will tend to be higher according to the EQ-5D, compared with the SF-6D, as has been shown previously.

Our results in Table 1 seem to support this conclusion as the QALY gain associated with each of the 4 options was estimated to be higher according to the EQ-5D. However, the rank ordering of the 4 options (from the least to highest QALY gain) differed between the EQ-5D and SF-6D, as has been shown previously. Our results in Table 1 seem to support this conclusion as the QALY gain associated with each of the 4 options was estimated to be higher according to the EQ-5D. However, the rank ordering of the 4 options (from the least to highest QALY gain) differed between the EQ-5D and SF-6D, as has been shown previously. The main implication of the finding that estimates of cost-effectiveness differ between measures is that further research is needed to determine which measure is the most appropriate in this, and other, clinical areas. Thereby, building upon previous studies, which have assessed how the different health state descriptions of different measures may account for differing utility scores, future studies should be encouraged to include more than 1 measure, and to test the various criteria that have been developed to assess whether instruments are measuring what they purport to for eg, construct validity and responsiveness. Due to the range of issues that need to be considered when choosing between measures there is, however, insufficient space in this article to determine which of the EQ-5D and SF-6D is most appropriate for people with knee pain. Indeed, though the EQ-5D could be argued to be the more sensitive measure in this population group (utility gains were consistently estimated to be higher according to the EQ-5D), this does not necessarily mean that it is the preferred measure as this could represent an overestimation of the true gain in utility associated with the particular change in health. Recently, in the light of these factors, and to ensure a consistent approach is taken in different analyses, the UK National Institute of Health and Clinical Excellence has developed a reference case, wherein certain methods for conducting cost-effectiveness analyses are recommended, including use of the EQ-5D to estimate adult utility scores.

An additional finding was that the EVPI (the potential value of further research) was more than twice as high according to the EQ-5D, compared with the SF-6D. The EVPI is determined by both the probability, and consequences, of making the wrong decision. As the probability that each option was cost-effective (depicted by the CEAC) was similar according to the EQ-5D and SF-6D (mentioned in Figs. 1 and 2, respectively), the different EVPI values result from different estimates of the consequences of making the wrong decision. The latter is dependent upon the net benefit ($\lambda E - C$) of each option. Therefore, as the difference in QALYs (between options) was estimated to be higher according to the EQ-5D, the difference in net benefit (between options) was also estimated to be higher, as was the consequences of making the wrong decision and the EVPI. As has
been argued previously, this demonstrates that both the CEAC and EVPI provide different information about the estimated level of the uncertainty associated with a decision regarding cost-effectiveness, and that both values can be of use to decision makers. With the CEAC uncertainty is characterized by estimating the probability that an option is cost-effective at different levels of the cost-effectiveness threshold, whereas the EVPI estimates the upper limit for the value of undertaking further research.

We are aware of 4 articles which have compared estimates of cost-utility according to different measures. Pickard et al estimated the ICER for both a group of asthma and stroke patients, where these were calculated using 10 different algorithms which have previously enabled scores on both the SF-12 and SF-36 to be converted into utility scores. They found that the ICER for the asthma patients ranged between $30,769 and $63,492, compared with $27,972 and $72,727 for the stroke patients. Moreover, in line with our results, they found that those algorithms which produced more favorable ICERs for the asthma patients did not necessarily produce more favorable results for the stroke patients. Marra et al compared 2 drug options for patients with rheumatoid arthritis where, as previous studies had not included a utility instrument, mapping was used to convert scores on a clinical measure into estimated utility scores. The ICERs for the EQ-5D ($46,322), SF-6D ($69,826), HUI2 ($53,429), and HUI3 ($32,018) covered a range greater than $37,000 per QALY. van den Hout et al compared 2 therapy programs for people with rheumatoid arthritis. The exercise program was found to be dominated according to the EQ-5D, yet had an ICER of £67,000 per QALY according to the SF-6D. Michaels et al compared conservative treatment to surgery for varicose veins, where the ICER for surgery was estimated to be £4417 according to the EQ-5D and £7175 according to the SF-6D. With the possible exception of the last study, it can be seen that the findings of these studies are in line with ours in that different measures tend to produce different ICERs, which can thereby lead to quite different conclusions about whether an intervention is cost-effective or not.

One potential limitation of this article is that a complete case analysis approach was used. This enabled one to compare the EQ-5D and SF-6D scores, and cost-utility estimates, on the same group of patients without introducing other factors, which might account for any differences. However, as these 247 participants may be unrepresentative of the 389 who took part in the LIKP study, by way of sensitivity analysis, we also used multiple imputations to estimate the cost-effectiveness of the different options. Though the actual results were slightly different (eg, according to the EQ-5D, option 1 was estimated to have an ICER of £10,469 when multiple imputation was used, compared with £10,815 with a complete case analysis [as mentioned in Table 1]), the implications were the same. The main strength of the article is that it is thought to be the first to compare (1) estimates of cost-utility, based on different measures, when evaluating more than 2 options and (2) how the level of uncertainty, according to the CEAC and EVPI, also differs across measures.

CONCLUSION

We have estimated the cost-utility of 4 different options according to both the EQ-5D and SF-6D. The estimated utility scores and QALY gain for each of the 4 options differed between measures, as did the rank ordering of options (from least to most effective). Consequently, we have shown that the choice of measure can greatly influence decisions as to which option is most cost-effective, and the value of undertaking further research.


