Reliability and validity of the EuroQol in patients with osteoarthritis of the knee

M. Fransen and J. Edmonds
St George Hospital/University of NSW, Sydney, Australia

Abstract

Objective. To assess the reliability and validity of the EuroQol (EQ-5D) for osteoarthritis of the knee (OA knee).

Methods. Eighty-two patients with OA knee were asked to complete on two occasions, separated by 1 week, the EQ-5D, the Western Ontario and McMaster Universities (WOMAC) Osteoarthritis Index and the 36-item short form of the Medical Outcomes Study (SF-36).

Results. In this patient population, <10% of the 243 EQ-5D health states were active. The EQ-5D demonstrated a non-Gaussian distribution. Reliability [intraclass correlation coefficient (ICC) = 0.70] is acceptable for aggregate level data. There were significant rank correlations with both the WOMAC and SF-36.

Conclusions. This study provides some evidence of EQ-5D construct validity and reliability. However, the restricted and non-normal distribution of scores, the marked difference between patients' self evaluation and derived societal utility tariffs, as well as the lack of discriminative ability for patients with 'moderate' morbidity within each of the five EQ-5D dimensions, are of concern.

KEY WORDS: EuroQol, OA knee, QALY, Utility, WOMAC, SF-36, Health status measures, Health index.

The diminishing marginal gains in survival due to technological improvements in health care in 'developed' countries have led to the emergence of health-related quality of life as an important issue. It is hoped that health care will not only provide optimal length of life, but also maximize quality of life. Cost per quality-adjusted life year (QALY), analysed from aggregate population-based data, provides a well-developed framework to guide rational and explicit health policy decision making aiming to maximize allocative efficiency of diminishing health care resources [1]. Currently, an abundance of validated disease-specific and generic health status scales and questionnaires can produce a profile of health status with scores on various areas of functioning [2, 3]. For comparative economic analysis, however, it is necessary to obtain a single index of the perceived health-related quality of life which accurately reflects preferences for different health states.

Within a QALY framework, the quality adjustment factor is an estimate of the utility associated with preferences for different health states [4–6]. The strength of an individual's preference or utility is usually elicited using either the standard gamble or the time trade-off method. The standard gamble technique estimates utility under conditions of risk or uncertainty, whereas the time trade-off method incorporates choice under certainty. The numbers or weightings representing the strength of an individual's preferences for an experienced or described health state are scored between 0 (death or worst imaginable health state) and 1 (full health or best imaginable health state). The quality adjustment is then multiplied by the expected life years in the assessed health state to arrive at the number of QALYs achieved. The expected utility associated with a health care intervention is then the sum of the probability of entering a health state multiplied by the utility (QALY) associated with that state.

The aim of this study was to assess the measurement reliability and validity of the EuroQol (EQ-5D), an instrument designed to derive from five dimensions of health (mobility, self-care, usual activities, pain, mood) a single cardinal index for the quality weighting of QALYs. The EQ-5D uses valuations derived with the time trade-off method from a large general population survey to score the five-dimension health profile self-reported, in this study, by subjects with osteoarthritis of the knee (OA knee). The numerous ethical implications of using QALYs as a basis for resource allocation or the valuation validity of the time trade-off method were beyond the scope of this study [7].

Methods

Patients
All patients appearing over a 20 month period on the physiotherapy waiting list of a large public hospital with a diagnosis of OA knee (or knee pain) aged 50 yr and
older were invited to participate in a randomized controlled clinical trial assessing the effectiveness of physiotherapy. More than 95% of those with a confirmed diagnosis of OA knee [8] agreed to participate in the study as participation resulted in a 67% chance of avoiding the usual 6–8 week waiting time for treatment. The first 82 patients (25 males, 57 females), mean age 68.0 yr (range 46–88), mean symptom duration 9.9 yr (range 1–50), were asked to complete three questionnaires: the EuroQol, the Western Ontario and McMaster Universities (WOMAC) Osteoarthritis Index and the 36-item short form of the Medical Outcomes Study (SF-36) health survey questionnaire. Each questionnaire was re-administered to each study participant 1 week later. The questionnaires were presented in the same order on both pre-randomization assessment occasions and self-completed by the subjects. Study participants were not informed at the baseline assessment that they would be completing the same questionnaires at the follow-up assessment, nor were they given access to their baseline responses.

Health status instruments

The EuroQol (index measure). The EuroQol instrument 'has been purposefully developed to generate a cardinal index of health' primarily for evaluative studies and policy research and 'is intended to complement other forms of quality of life measures' [9, 10]. The self-report questionnaire has two sections. The first part (EQ-5D) consists of five questions covering the dimensions of mobility, self-care, usual activities, pain/discomfort and anxiety/depression, each with three levels of response (Appendix 1). The time frame is given as ‘... your own health state today’. The resultant five-part health profile (e.g. 21122, 11223, etc.) is weighted according to EuroQol Group guidelines. The tariffs, or weights, for any of the theoretically possible 245 health profiles have been derived from the latest UK population survey (n = 3337) using the time trade-off valuation method [11, 12], to create a single index of quality of life ranging from 0 to 1. The second part (EQ-VAS) of the EuroQol consists of a 20 cm vertical visual analogue scale (VAS) ranging from 100 (best imaginable health state) to 0 (worst imaginable health state). The EQ-VAS gives a self-assessed measure of overall health state. The EQ-VAS does not express any longevity trade-offs or uncertainty and therefore does not have utility valuation validity, and is not intended as quality adjustments for QALYs.

WOMAC (disease-specific measure). The WOMAC Osteoarthritis Index is a disease-specific self-report questionnaire assessing pain (five questions), stiffness (two questions) and physical disability (17 questions) in patients with OA of either the hip or the knee [13, 14]. The time frame is given as ‘currently experiencing’. Responses to the questions are rated by the subject on 100 mm VAS ranging from 0 (indicating no pain, stiffness or difficulty) to 100 (indicating extreme pain, stiffness or difficulty). The two questions concerning stiffness are currently being replaced by the originator and will not be analysed further in this paper.

SF-36 (generic profile measure). The SF-36 [15, 16] is a widely used self-report generic health status questionnaire measuring eight dimensions of health status: physical functioning (10 questions), role limitations due to physical health problems (four questions), bodily pain (two questions), social functioning (two questions), general mental health (five questions), role limitations due to emotional problems (three questions), vitality (four questions) and general health perceptions (five questions). The time frame is given as ‘during the past 4 weeks’. Each of the eight subscales generates a score from 100 to 0 with 100 indicating best health and 0 indicating worst health. An additional question, not scored, relates to perceived change in health status over the past year. Most of the SF-36 subscales have been validated for diverse patient groups [17–19] and the SF-36 can be completed within 10 min by most people. Recently, the originators of the SF-36 have developed algorithms to calculate two psychometrically based summary measures: the Physical Component Summary Scale Score (PCS) and the Mental Component Summary Scale Score (MCS) [20, 21]. The PCS and MCS are norm-based scores. A linear T-score transformation method is used so that both the PCS and MCS have a mean of 50 and a s.d. deviation of 10 in the general US population. The PCS and the MCS provide greater precision, reduce the number of statistical comparisons needed, and eliminate the floor and ceiling effects noted in several of the subscales [17, 19, 22–24].

Analysis

As the EQ-5D is intended for use in evaluative studies and policy research rather than as a clinical tool, only the reliability of aggregate level data is presented and compared with the WOMAC and SF-36. Systematic error or agreement have been analysed using the difference scores with their associated 95% confidence intervals (CIs). Reliability was analysed using intraclass correlation coefficients (ICCs).

Establishing the criterion or content validity of an instrument claiming to measure health-related quality of life is difficult as there are no established gold standards for comparison. Evidence for construct validity can only be accumulated by ‘a priori’ hypothesized patterns of associations with other validated instruments. In this study, it is hypothesized that associations with the disease-specific WOMAC scores will be weaker than associations with the generic SF-36 scores or the EQ-VAS in this older sample with probable co-morbidities.

To investigate discriminatory validity, an analysis was made to assess whether the three response levels available within each of the five health dimensions of the EQ-5D would accurately discriminate symptom severity or functional disability as rated by comparable generic SF-36 scores.

Data analysis was carried out using SAS Proprietary Software Release 6.12 (SAS Institute Inc., Cary, NC 27511, USA).
Table 1. Intra-subject agreement. Paired t-tests (two-tailed)

<table>
<thead>
<tr>
<th>Questionnaire (n = 82)</th>
<th>Baseline median</th>
<th>Baseline mean</th>
<th>95% CI change score</th>
</tr>
</thead>
<tbody>
<tr>
<td>EQ-5D (0–1)</td>
<td>0.66</td>
<td>0.58</td>
<td>−0.01, 0.03</td>
</tr>
<tr>
<td>EQ-VAS (0–100)</td>
<td>74.0</td>
<td>71.3</td>
<td>−5.30, 0.28</td>
</tr>
<tr>
<td>WOMAC pain (100–0)</td>
<td>36.8</td>
<td>37.2</td>
<td>−1.59, 5.51</td>
</tr>
<tr>
<td>WOMAC function (100–0)</td>
<td>35.8</td>
<td>39.6</td>
<td>−1.32, 4.00</td>
</tr>
<tr>
<td>SF-36 PCS (mean = 50, s.d. = 10)</td>
<td>32.3</td>
<td>34.0</td>
<td>−2.17, 0.21</td>
</tr>
<tr>
<td>SF-36 MCS (mean = 50, s.d. = 10)</td>
<td>43.3</td>
<td>43.6</td>
<td>−0.32, 1.36</td>
</tr>
</tbody>
</table>

Results

Distribution

Less than 10% of the possible 243 EQ-5D health states were active in this sample. Whilst the WOMAC and the SF-36 scores were normally distributed, the EQ-5D score had a marked bimodal distribution (Fig. 1). As the EQ-5D baseline (and follow-up) scores demonstrated a non-normal distribution, non-parametric procedures were often needed. Difference scores between baseline and follow-up scores were, however, all normally distributed, allowing parametric procedures.

Reliability

Paired t-tests were used to detect any systematic error from baseline to follow-up (Table 1). There were no significant systematic changes over the 1 week period at the 5% significance level. The 95% CIs for systematic error were narrow. Analysing each of the five dimensions of the EQ-5D, perfect test–retest agreement on the three response levels ranged from 77% (Usual activities) to 88% (Self-care).

Discriminant validity

Discriminant validity was analysed with what were considered comparable scores on the generic SF-36 questionnaire (Table 4). The EQ-5D does appear to be able to discriminate accurately within each separate health state descriptor.

Discussion

Distribution

In this patient sample, 10 health states (out of a theoretically possible 243) accounted for 82% of the responses with the remaining responses divided over only 13 other
Table 3. Correlation coefficients—Spearman’s rho

<table>
<thead>
<tr>
<th></th>
<th>Baseline EQ-5D (95% CI)</th>
<th>Follow-up EQ-5D (95% CI)</th>
<th>Maximum theoretical correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symptom duration</td>
<td>0.25 (0.04, 0.44)</td>
<td>0.30 (0.09, 0.49)</td>
<td>0.84</td>
</tr>
<tr>
<td>EQ-VAS</td>
<td>0.42 (0.22, 0.58)</td>
<td>0.43 (0.23, 0.59)</td>
<td>0.71</td>
</tr>
<tr>
<td>WOMAC pain</td>
<td>−0.30 (−0.09, −0.49)</td>
<td>−0.34 (−0.13, −0.52)</td>
<td>−0.65</td>
</tr>
<tr>
<td>WOMAC function</td>
<td>−0.33 (−0.13, −0.51)</td>
<td>−0.37 (−0.16, −0.54)</td>
<td>−0.75</td>
</tr>
<tr>
<td>SF-36 PCS</td>
<td>0.20 (0.00, 0.40)</td>
<td>0.32 (0.11, 0.50)</td>
<td>0.72</td>
</tr>
<tr>
<td>SF-36 MCS</td>
<td>0.39 (0.19, 0.56)</td>
<td>0.60 (0.44, 0.72)</td>
<td>0.77</td>
</tr>
</tbody>
</table>

Table 4. Discriminant validity

<table>
<thead>
<tr>
<th>EuroQol</th>
<th>n</th>
<th>Median SF-36</th>
<th>Interquartile range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mobility (SF-36: 3g.h.i)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No problems walking about</td>
<td>13</td>
<td>2.67</td>
<td>2.67–3.00</td>
</tr>
<tr>
<td>Some problems walking about</td>
<td>69</td>
<td>2.00</td>
<td>1.67–2.67</td>
</tr>
<tr>
<td>Confined to bed</td>
<td>0</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Self-care (SF-36: 3j.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No problems</td>
<td>70</td>
<td>3.00</td>
<td>3.00–3.00</td>
</tr>
<tr>
<td>Some problems</td>
<td>12</td>
<td>2.00</td>
<td>2.00–2.55</td>
</tr>
<tr>
<td>Unable</td>
<td>0</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Usual Activities (SF-36: PCS)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No problems</td>
<td>23</td>
<td>40.4</td>
<td>35.93–46.02</td>
</tr>
<tr>
<td>Some problems</td>
<td>59</td>
<td>29.7</td>
<td>26.23–35.79</td>
</tr>
<tr>
<td>Unable</td>
<td>0</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Pain/Discomfort (SF-36: 7.8.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No pain</td>
<td>2</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Moderate pain/discomfort</td>
<td>68</td>
<td>41.0</td>
<td>31.0–61.0</td>
</tr>
<tr>
<td>Extreme pain/discomfort</td>
<td>12</td>
<td>22.0</td>
<td>22.0–31.0</td>
</tr>
<tr>
<td>Anxiety/Depression (SF-36: MCS)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not anxious/depressed</td>
<td>47</td>
<td>47.70</td>
<td>44.11–50.78</td>
</tr>
<tr>
<td>Moderately anxious/depressed</td>
<td>32</td>
<td>37.90</td>
<td>34.04–43.90</td>
</tr>
<tr>
<td>Extremely anxious/depressed</td>
<td>3</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

*SF-36: scores 1, 2, 3 (1 = limited a lot, 2 = limited a little, 3 = not limited).
*SF-36 PCS/MCS: scores normal population (mean = 50, s.d. = 10).
*SF-36 Pain: scores 100–0 (100 = no pain/work interference, 0 = very severe pain/work interference).

health states. This limited distribution of scores is not restricted to patient samples. In a large general population survey, Brazier et al. [23] found that only 10 health states (out of a theoretically possible 216 in a previous version of the EuroQol) accounted for 95% of the responses. In our non-institutionalized sample, in fact, the choice was essentially limited to two response levels at each of the five dimensions of health (Table 4), resulting in <10% of the EuroQol health states being active in this out-patient sample and restricting the potential responsiveness of the instrument in this population.

Whilst only one of the EQ-5D dimensions demonstrated a marked ceiling effect in this sample with ‘moderate’ morbidity (Table 4), the crudeness of the levels means that someone responding with moderate levels of morbidity ‘is not permitted by the descriptive system to describe a further deterioration in health that is anything less than a major collapse’ [25]. Similarly, the EQ-5D descriptive system allows patients with ‘moderate’ morbidity only to describe a full recovery. A large study involving patients with rheumatoid arthritis (RA) did claim to demonstrate that the EQ-5D was ‘as responsive to self-reported clinical change … as many of the condition-specific instruments’ [26]. The RA patients were selected to ‘obtain a broad cross-section of disease severity’; however, approximately half of the sample had severe morbidity with a baseline median EQ-5D score of $\leq 0.12$.

The bimodal distribution of the EQ-5D scores reflects the relatively large decrement in societal valuations from level 2 to level 3 compared with decrements from level 1 to level 2. For example, a fall in Pain/discomfort only from level 1 to level 2 decreases the derived societal index by 0.15, whilst a fall from level 2 to level 3 decreases the valuation by 0.45, or nearly half the total score range. Decrements from level 2 to level 3 in the other four dimensions range from 0.25 to 0.43 for Mobility. The population tariffs therefore reveal that there is a huge difference in health state utility between ‘moderate pain or discomfort’ and ‘extreme pain or discomfort’. Similarly for mobility, there is a large unassessed area between ‘some problems walking about’ and ‘confined to bed’. Whilst the EuroQol instrument
makes no claims that each of the five dimensions has interval properties [9, 27], the disadvantages of the EuroQol descriptive system for other patients with moderate levels of morbidity were well demonstrated in a recent randomized controlled trial involving patients with benign prostatic hypertrophy [25]. An additional level between level 2 and level 3 would potentially overcome the limited and abnormal distribution of scores in patient populations with 'moderate' morbidity. It is suggested that this could be achieved without the EQ-5D become burdensome and losing its aim of being suitable as an 'add-on to other data collection' [27].

**Reliability**

The ICCs of the EQ-5D (Table 2) could be considered modest considering the short test–retest interval of 1 week and may not be adequate for studies using a similar, relatively small sample. For aggregate analysis, however, it is claimed that a reliability of >0.50 may be acceptable with a large sample size [2]. However, ICC analysis does result in lower values compared with the Pearson equivalent as ICCs account for any additive or multiplicative element [3]. Furthermore, the EQ-5D (and the WOMAC pain) are each derived from only five scores, whilst the WOMAC physical function score is a mean of 17 VAS ratings, and the SF-36 PCS and SF-36 MCS are each derived from 35 questions. The inclusion of a greater number of subscales necessarily results in a more stable score, less susceptible to measurement error. This possibly partly explains the superior reliability demonstrated by the WOMAC physical function score and the SF-36 scores.

Modest reliability may, however, just reflect fluctuating symptomatology characteristic of OA knee as both the EuroQol and the WOMAC ask the responders to consider their 'health state today' or 'recently experienced'. In contrast, the SF-36 asks responders to consider 'the past 4 weeks', potentially allowing greater 1 week test–retest reliability compared with the EuroQol and the WOMAC.

**Validity**

Construct validity cannot be proved definitively, but it is a continuing process of accumulating evidence [28]. This study provides some evidence of concurrent validity with significant and mostly stable associations between the EQ-5D and the other health status questionnaires on both assessment occasions (Table 3).

The EQ-5D was easily able to discriminate symptom severity as scored by the chosen SF-36 scores, but discriminative ability was crude as it was essentially limited to two levels within each of the five health dimensions for this sample (Table 4). Although mental health is an important health determinant, for a disease such as OA knee it is noteworthy that the EQ-5D appears to demonstrate a stronger association with a patient’s emotional state (SF-36 MCS) than with a patient’s physical state (SF-36 PCS) (Table 3). In fact, it has been claimed that ‘the attenuation of validity coefficients due to unreliability implies that a correlation of 0.60 between two measures represents an extremely strong association’ ([2], p. 34). More than 40% of this sample, referred for physiotherapy, reported having at least moderate levels of anxiety/depression on the EQ-5D (Table 4). Demonstrated mental distress may have been the stimulus for physicians to refer particularly these patients with OA knee for active physical intervention. Therefore, although exclusion co-morbidities [neurological disorders affecting gait, joint arthroplasty in the lower limbs or unstable cardiac conditions that preclude aerobic exercise at a very moderate level (50–60% maximum heart rate)] were limited, this sample may not be truly representative of community-dwelling people with OA knee. Hurst et al. [29] found in a pilot study of 55 patients with RA that ‘change in mood’ was the strongest predictor of change in the EQ-5D. It may be that a patient’s emotional state is a good index of disease severity in OA knee. If, however, a patient’s emotional state markedly influences their physical health status perception, the resultant random measurement error would restrict the validity of the EQ-5D or other self-report questionnaires to only relatively large studies.

**General points**

Our subjects experienced two main problems with the wording of the EuroQol. Many subjects found the grouping together of the words pain and discomfort confusing. Discomfort was perceived as a mild pain; therefore, extreme discomfort was equated by many as moderate pain. There were also some interpretation difficulties concerning the term ‘walking about’—how far? include stairs? This may explain the low test–retest association for Mobility (Spearman’s rho = 0.29) compared with the other four dimensions (Spearman’s rho range 0.50–0.60).

In agreement with other studies [26, 30], it would appear that patients in this study placed a higher valuation on their own health state than that given by the general population tariffs. If the two EuroQol scores are made mathematically comparable by multiplying the EQ-5D score by 100, the poor agreement between the patient’s rating (median score 74.0) of their own general health state and the valuation given by the general population survey (median score 66.0) is highlighted. It could easily be argued that agreement between the scores is not an issue as there is a fundamental difference between a simple ‘risk-free’ valuation of a health state (EQ-VAS) and a utility measure of preferences involving a longevity trade-off (EQ-5D). Furthermore, it has been well documented that comparable time frame, context and prognosis information are crucial when evaluating a health state [7]. Responders in the population survey were told that each health state, derived from the EuroQol descriptive system (Appendix 1), was to be regarded as lasting for 10 yr without change, followed by death, and were given no information regarding specific disease context [11, 12]. In contrast, the study participants have context knowledge, but uncertain and differing opinions on their probable disease progression rate and life
expectancy. It is probable, however, that the study participant’s median assessment of 74.0 would be even higher if longevity trade-offs were involved in the valuation process. The validity and wider ethical implications of using non-contextual health state descriptors to elicit quality of life valuations in the general population, with little direct experience of various levels of morbidity and little opportunity for reflection, have frequently been questioned [7, 31, 32].

OA knee has been identified as one of the five diseases responsible for the greatest proportion of physical disability in non-institutionalized elderly men and women [33]. ‘Moderate pain/discomfort’, ‘some problems walking about’ and ‘some problems with usual activities’ cover an enormous spectrum of pain and disability. In fact, ‘some problems walking about’ in the EQ-5D covers the entire range of gait disability as it is the only level between ‘no problems’ and ‘confined to bed’! Similarly, ‘some problems with performing my usual activities’ spans the entire spectrum of activities of daily life dysfunction from people able to live entirely independently, and even provide support for others, to those people needing maximum support of carers, equipment and/or finances. The probable limited responsiveness of the EQ-5D to clinically significant changes will be a major problem for patients with diseases of mostly ‘moderate’ chronic morbidity, such as OA knee. QALYs already discriminate against the elderly as, with their decreased life expectancy, they have decreased potential to accrue QALYs. An unresponsive measure of quality of life for chronic ‘moderate’ morbidity will further disadvantage patients with OA knee in resource allocation decisions.

Conclusions

The EQ-5D is a short, easily completed questionnaire feasible for use as an add-on data collection in large clinical studies. The study demonstrated that the EQ-5D has reliability sufficient for aggregate level data and comparable with the WOMAC and the SF-36. The study also provides some evidence of moderate convergent validity. However, the coarseness of the levels within each of the five EQ-5D health dimensions is of concern as it results in <10% of the health states being active in this population despite their chronic ‘moderate’ morbidity, thereby permitting only crude discriminative ability and restricting responsiveness. The described restricted and demonstrated non-normal distribution of the EQ-5D scores, and the incomparability between patient’s valuation of their own health state and the derived societal utility tariffs, should be further investigated before the EuroQol is considered a valid instrument for evaluative studies and policy research in this population. If health care resource allocation becomes purely EQ-5D QALY based, it is highly conceivable that these older patients with ‘moderate’ levels of chronic morbidity will be severely disadvantaged.

References


Appendix 1

EQ-5D

By placing a tick (thus √) in one box in each group below, please indicate which statements best describe your own health today.

Mobility
I have no problems in walking about
I have some problems in walking about
I am confined to bed

Self-Care
I have no problems with self-care
I have some problems washing or dressing myself
I am unable to wash or dress myself

Usual Activities (e.g. work, study, housework, family or leisure activities)
I have no problems with performing my usual activities
I have some problems with performing my usual activities
I am unable to perform my usual activities

Pain/Discomfort
I have no pain or discomfort
I have moderate pain or discomfort
I have extreme pain or discomfort

Anxiety/Depression
I am not anxious or depressed
I am moderately anxious or depressed
I am extremely anxious or depressed

Compared with my general level of health over the past 12 months, my health state today is:
Better
Much the same
Worse