A Longitudinal Comparison of 5 Preference-Weighted Health State Classification Systems in Persons with Intervertebral Disk Herniation

Christine M. McDonough, Tor D. Tosteson, Anna N. A. Tosteson, Alan M. Jette, Margaret R. Grove and James N. Weinstein

Med Decis Making 2011 31: 270 originally published online 22 November 2010
DOI: 10.1177/0272989X10380924

The online version of this article can be found at:
http://mdm.sagepub.com/content/31/2/270

Published by:

SAGE
http://www.sagepublications.com

On behalf of:

Society for Medical Decision Making

Additional services and information for Medical Decision Making can be found at:

Email Alerts: http://mdm.sagepub.com/cgi/alerts
Subscriptions: http://mdm.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav

>> Version of Record - Mar 10, 2011
OnlineFirst Version of Record - Nov 22, 2010

What is This?
A Longitudinal Comparison of 5 Preference-Weighted Health State Classification Systems in Persons with Intervertebral Disk Herniation

Christine M. McDonough, PhD, PT, Tor D. Tosteson, ScD, Anna N. A. Tosteson, ScD, Alan M. Jette, PhD, MPH, PT, Margaret R. Grove, MS, James N. Weinstein, DO, MS

Objective. To assess the longitudinal validity of widely used preference-weighted measurement systems for economic studies of intervertebral disk herniation (IDH).

Methods. Using data at baseline and 1 year from 1000 Spine Patient Outcomes Research Trial (SPORT) participants with IDH and complete data, the authors considered the EQ-5D with UK and US values (EQ-5D-UK and EQ-5D-US), 2 versions of the Health Utilities Index (HUI3 and HUI2), the SF-6D, and a regression-estimated quality of well-being score (eQWB). Differences in mean change scores (MCS) were assessed using signed rank tests, and Spearman correlations were calculated for change scores by system pairs. Using the Oswestry Disability Index, symptom satisfaction, progress rating, and self-perceived health ratings as criterion measures, the authors tested for trend in MCS across levels of change in criteria. They calculated floor and ceiling effects, effect size (ES), standardized response mean, and minimal important difference estimates.

Results. All systems demonstrated linear trends with external criteria and moderate to strong correlations between systems. However, differences in performance were evident. SF-6D and eQWB were most responsive (ES: 1.9 and 2.3, respectively), whereas EQ-5D-US and EQ-5D-UK were least responsive (ES: 1.23/1.20). Ceiling and floor effects were noted for all systems within key dimensions and for EQ-5D-UK and EQ-5D-US for overall score. MCS ranged from 0.40 (0.38) for EQ-5D-UK to 0.13 (0.09) for eQWB and differed significantly, except between EQ-5D-UK and HUI2.

Conclusions. This research supports the validity of all systems for measuring change in persons with IDH, without finding a clearly superior system. The unique characteristics of each system revealed in this study should guide system choice. Key words: health state preferences, utilities, and valuation; health status indicators; spine diseases; quality of life; economic evaluation; SPORT; scale validation; cost-utility analysis; cost-effectiveness analysis. (Med Decis Making 2011;31:270–280)

There is evidence of increasing use of cost-utility analysis to assess the relative value of alternative treatment interventions when resources are limited.1,2 To estimate quality-adjusted life years (QALYs) for the denominator of the incremental cost-effectiveness ratio (ICER), outcomes of treatment are measured using a single score, anchored at 0 for death and 1 for perfect health, and weighted for the relative desirability of the health state. Standards for economic evaluations recommend using societal values (utilities or preferences).3 The two main approaches to obtaining “societal health state values” include 1) direct measurement of values for health states of a representative sample of the population using methods such as standard gamble, time tradeoff, and visual analog scale (VAS) ratings and 2) indirect measurement using preference-weighted health state classification systems such as the Quality of Well-Being Scale,4 the EuroQoL EQ-5D,5,6 the McMaster Health Utilities Index HUI,7-9 or the SF-6D.10,11 In addition, methods have been developed to estimate health state values from existing health-related QOL (HRQOL) data, for example, using regression models.

Preference-weighted health state classification systems are increasingly used in cost-utility analyses to estimate change in QALYs.12 Furthermore, they are increasingly used as measures of health outcome in clinical trials. Systems vary in their

DOI: 10.1177/0272989X10380924
approaches to the design of each component: the descriptive system, preference measurement method, source of community preferences, and approaches to scoring. Questions remain about the comparability of systems in specific populations and about the extent to which differences in systems could affect the results of cost-utility analyses and, therefore, policy decisions. In choosing among measurement systems, researchers need to know the strengths and weaknesses of alternatives to optimize measurement performance for the particular problem under study, to interpret score changes or differences, and for study planning.

Evidence from cross-sectional comparisons indicates that significant variation exists in mean scores obtained from different systems. However, when the research purpose is to measure change due to treatment, as in cost-utility analysis, longitudinal studies are necessary to evaluate system performance. Longitudinal head-to-head comparisons of preference-weighted systems indicate that change scores vary across systems used to estimate health state values (HSV). In addition, there is some evidence of difficulty measuring change at the best levels (ceiling effects) for SF-derived systems and at the worst levels (floor effects) for SF-derived systems and the question of ability to detect changes in function unrelated to the extremities for the HUI3.

To our knowledge, there are no longitudinal comparisons of these systems in the population of patients with intervertebral disk herniation, a common and costly spine disorder. Therefore, we aimed to conduct a critical comparison of the measurement characteristics of the EQ-5D-UK, EQ-5D-US, HUI3, HUI2, SF-6D, and an algorithm to estimate the quality of well-being (QWB) from SF-36 data (eQWB) among Spine Patient Outcomes Research Trial (SPORT) participants with intervertebral disk herniation (IDH).

METHODS

Sample and Selection

We used baseline and 1-year data from an ongoing prospective study of interventions for symptomatic lumbar spine disorders (SPORT). The design of this study has been previously reported in detail. In brief, SPORT is a multicenter study including 3 randomized trials and 3 observational cohorts. To be eligible for SPORT, participants were 18 years or older and had a diagnosis of IDH, spinal stenosis (SpS), or degenerative spondylolisthesis (DS). Participants were excluded if there was evidence of nonsurgical treatment for fewer than 6 weeks for IDH and 12 weeks for SpS and DS, cauda equina syndrome, contraindications to spine surgery, possible pregnancy, active malignancy, current fracture, infection, or prior lumbar spine surgery.

Measures of Health State Value

The instruments used to characterize health state values are described below.

**EQ-5D.** The EuroQoL EQ-5D includes 5 attributes rated on 3 levels to define 245 health states (when “dead” and “unconscious” are added). Using the same EQ-5D health state classification system with the reference timeframe “today,” we applied EQ-5D-UK preference weights and EQ-5D-US weights. The UK (York) weights were measured using time tradeoff values for a subset of health states from a sample of the UK population. The US weights were measured using time tradeoff in a representative sample of the US population. Both systems use additive models of attribute independence with different adjustments for any health state at the worst possible level.
Health Utilities Index. The McMaster Health Utilities Index has been well described. Using the same health state classification system, SPORT is licensed to apply the Mark 2 (HUI2) and the Mark 3 (HUI3) utility functions. HUI2 represents 7 attributes on 4 or 5 levels and defines 24,000 health states. HUI3 has 5 or 6 levels for each of its 8 attributes and encompasses 972,000 unique health states. The HUI2 and HUI3 use multiplicative multiattribute utility functions based on visual analog and standard gamble scores obtained from community samples in Canada. The reference timeframe for the questionnaire was “the past 4 weeks,” and we did not include the fertility dimension in our survey.

SF-36-derived measures. The SF-6D, version 2, provides a method for deriving a preference score from the SF-36 instrument. It represents 6 attributes on up to 6 levels. An additive model was used and community weights were derived using standard gamble utilities from a UK population for a subset of health states.

Estimated quality of well-being score (eQWB). The QWB Scale is a preference-based health measure that includes 3 additive functional dimensions and a symptom dimension. Community preferences were measured using a category rating for a representative sample of 866 adults in the San Diego area. Scores can range from 0.0 to 1.0, although the lowest score for a health state other than death is 0.32. We previously estimated QWB scores using a regression model based on 5 subscales of the SF-36 reported by the Beaver Dam Health Outcomes Study.

Criterion Measures

Criteria for comparison used for this study included a disease-specific measure, the Oswestry Disability Index (ODI), and patient ratings of satisfaction with symptoms (symptom satisfaction), self-perceived progress (progress rating), and self-perceived health (SPH).

Oswestry Disability Index (AAOS—modified version). The ODI includes 9 items on 6 levels and yields an index score from least to most disability of 0 to 100. For consistency of interpretation, we subtracted ODI scores from 100 so that higher scores indicate better health.

Symptom satisfaction. The participant is asked, “If you had to spend the rest of your life with the symptoms you have now, how would you feel about it?” The response categories are very dissatisfied, somewhat dissatisfied, neutral, somewhat satisfied, and very satisfied.

Progress rating. The participant is asked, “How would you rate your progress with your spine-related problem since you first enrolled in SPORT?” Response categories are major improvement, minor improvement, no change, minor worsening, and major worsening.

Self-perceived health. The first question of the SF-36 asks, “In general would you say your health is excellent, very good, good, fair, or poor?” Participants first completed the ODI, followed by the SF-36; the EQ-5D, including a VAS rating; a symptom satisfaction rating; progress rating; and the HUI.

Statistical Analyses

We summarized participant characteristics according to the 4 criteria: ODI, symptom satisfaction rating, progress rating, and self-perceived health rating. Mean change scores (change scores) were calculated for each system from baseline to 1 year. We described the distribution of change scores using means, standard deviations, and ranges. We summarized the distribution of health state classifications by dimension and level using percents.

We tested differences between change scores using signed rank tests. We assessed longitudinal validity by calculating Spearman correlation coefficients for change scores for system pairs and using tests for trend across changes in levels of each criterion measure. We evaluated floor and ceiling effects for each system by calculating the proportion of participants who received the highest and lowest possible scores at baseline and at 1 year. This analysis was repeated for the key dimensions of pain and physical function.

We calculated responsiveness statistics and 95% confidence intervals at 1 year for each system using distribution- and anchor-based methods.

We calculated distribution-based effect size and standardized response mean estimates as follows:

\[
\text{Effect size} = \frac{\text{Mean (follow-up – baseline)}}{\text{SD}_{\text{baseline}}}
\]

\[
\text{Standardized response mean (SRM)} = \frac{\text{Mean (follow-up – baseline)}}{\text{SD}_{\text{change score}}}
\]
We calculated anchor-based minimal important difference (MID) estimates and 95% confidence intervals according to 4 anchors: ODI, symptom satisfaction, progress rating, and SPH rating. MID was calculated as the mean change for those who reported minimal important change according to each anchor. The scores of those who worsened were multiplied by $-1.47^{48}$ Minimal important change was defined as one level of change from baseline to 1 year for symptom satisfaction and self-perceived health. For progress rating, minimal important change was defined as report of minimal improvement or minimal worsening at 1 year. SPORT sample size calculation was based on a 10-point change in ODI, and consistent with this and other work on important change for ODI, we used a 10- to 19-point change as the definition of minimal important change in this study.43,49

All analyses were undertaken using STATA, version 9 (STATA Corporation, College Station, Texas).

### RESULTS

Data at 1 year were available for 1000 participants whose mean age was $42 \pm 11$ years (Table 1). This was a highly educated population, with the majority classifying their race as white. The majority of participants reported improved health based on all criterion measures.

A summary of mean change in health state values (change scores) is shown in Table 2. The largest mean change score was 0.40 for the EQ-5D-UK, which was 3 times the mean change of 0.13 for the eQWB. Standard deviations of the change scores were largest for the EQ-5D-UK, followed by the HUI3, EQ-5D-US, HUI2, SF-6D, and eQWB. Standard deviations were largest for the change scores, followed by baseline scores, and smallest for the 1-year scores.

Correlation between change scores as measured by Spearman coefficients ranged from 0.55 to 0.99 (Table 3). Not surprisingly, strongest correlations were noted between change scores from related systems, such as the EQ-5D-UK and EQ-5D-US, HUI3 and HUI2, and SF-6D and eQWB. Moderate to strong correlations were noted between change scores of all other systems.

When compared using sign rank tests, all change scores were significantly different from each other except the EQ-5D-US and HUI2. All systems demonstrated linear trends and high correlations between change scores and change in levels of ODI, symptom satisfaction, progress rating, and self-perceived health (Figure 1a-d).

At 1 year, less than 1% of participants received the lowest possible score for each system, and 28%
Figure 1  Trends in mean change in health state value with levels of progress rating and change in Oswestry Disability Index (ODI) score, symptom satisfaction, and self-perceived health rating. (a) Change in the ODI scores are presented in 10-point intervals. For consistency of interpretation, ODI was coded so that negative values represent worsening. Due to the small numbers of participants, intervals at the extremes of the range were collapsed. (b) Symptom satisfaction response levels include very satisfied, somewhat satisfied, neutral, somewhat dissatisfied, and very satisfied. Due to the limited number of participants reporting worsening, levels of worsened symptom satisfaction were combined. (c) Progress rating. Progress compared to baseline was directly rated by participants among these levels: major improvement, minor improvement, no change, minor worsening, and major worsening. (d) Self-perceived health rating. Participants rated current health from among the following categories: excellent, very good, good, fair, and poor. Due to the limited number of participants reporting more than one level of change, categories were combined for improvement or worsening of 2 levels or greater.
received the highest possible score for the EQ-5D-UK and EQ-5D-US (Table 2). In contrast, the HUI3 and HUI2 classified less than 10% at the ceiling, the SF-6D defined 5%, and the eQWB classified less than 1%. At baseline, each system classified a significant proportion of patients at the worst level for usual/physical function and pain. For the pain dimension, percentage at the floor was 40% for the EQ-5D, 29% for the HUI3, 19% for the HUI2, and 28% for the SF-6D. For mobility/physical function, it was 3% for the EQ-5D, 2% for the HUI3 and HUI2, and 14% for the SF-6D. The EQ-5D classified 25% at the floor for usual activities. At 1 year, there were large proportions at the best level (ceiling) for all dimensions. Specifically, in the mobility/physical function dimension, percentage at the ceiling was 73% for the EQ-5D, 82% for the HUI3 and HUI2, and 22% for the SF-6D. The proportions at the ceiling for pain were 33% for the EQ-5D, 20% for the HUI3 and HUI2, and 12% for the SF-6D. All systems classified a smaller proportion at the floor at 1 year. Floor effects were also noted for the EQ-5D in usual activities and pain/discomfort and the SF-6D in role limitations and vitality.

Table 4 summarizes responsiveness statistics. The estimated QWB score was most responsive, followed by the SF-6D. The EQ-5D-UK was consistently the least responsive, although the EQ-5D-US, HUI3, and HUI2 demonstrated similar or slightly less responsiveness. For example, the effect sizes for the EQ-5D-UK and eQWB were 1.2 and 2.3, respectively. The SRMs were 1.1 and 1.4, respectively.

Overall, MIDs were smaller for the eQWB and SF-6D than for the other 4 systems. Values for the EQ-5D-UK were approximately 3 to 5 times larger than those for the eQWB. For example, using ODI as the anchor, the MID estimate for the EQ-5D-UK was 0.12, whereas the estimate for the eQWB was 0.05.

DISCUSSION

Our study is the first large longitudinal comparison of preference-weighted system performance in persons with confirmed diagnosis of IDH. Correlations between systems and tests of trend with external criteria support the notion that all systems were valid measures of HRQOL. Estimates of effect size and standardized response mean indicate that all systems demonstrated the ability to measure change in key dimensions of HRQOL in this population of persons with spine disorders.

Considering the results of our validity tests together with the differences in mean change in health state values across systems, it is clear that there is no one system whose overall performance was superior to others. For example, the superior responsiveness of the eQWB and SF-6D, evidenced by the effect size and SRM estimates found in this study, confers an important advantage by enabling the detection of change with fewer study participants. Similarly, MID estimates indicated that the EQ-5D-UK would require a larger magnitude of change to be considered clinically important compared to the eQWB. However, the limited variation in scores upon which estimates of responsiveness are based has implications for policy applications. The eQWB and, to a lesser extent, the SF-6D did not provide scores across the full range of health state values relative to the anchors of dead and perfect health. Our study findings were consistent with other comparisons that support overall validity of all systems and somewhat better responsiveness for the SF-6D, paired with potential overvaluation of lower health states.32,50 Other studies indicate variations in performance across diagnoses and severity of health states.

### Table 4  Responsiveness Statistics for 1000 Spine Patient Outcomes Research Trial Participants with Intervertebral Disk Herniation

<table>
<thead>
<tr>
<th>System</th>
<th>Effect Size (95% CI)</th>
<th>SRM (95% CI)</th>
<th>ODI (n = 172) (95% CI)</th>
<th>Satisfaction (n = 170) (95% CI)</th>
<th>Progress (n = 194) (95% CI)</th>
<th>Self-Perceived Health (n = 403) (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EQ5D-UK</td>
<td>1.20 (1.13, 1.27)</td>
<td>1.06 (1.00, 1.13)</td>
<td>0.12 (0.09, 0.16)</td>
<td>0.25 (0.20, 0.30)</td>
<td>0.15 (0.10, 0.20)</td>
<td>0.14 (0.09, 0.19)</td>
</tr>
<tr>
<td>EQ-5D-US</td>
<td>1.23 (1.16, 1.30)</td>
<td>1.07 (1.01, 1.13)</td>
<td>0.08 (0.06, 0.11)</td>
<td>0.17 (0.13, 0.20)</td>
<td>0.10 (0.07, 0.13)</td>
<td>0.10 (0.06, 0.13)</td>
</tr>
<tr>
<td>HUI3</td>
<td>1.25 (1.18, 1.32)</td>
<td>1.17 (1.10, 1.23)</td>
<td>0.11 (0.08, 0.15)</td>
<td>0.21 (0.16, 0.25)</td>
<td>0.15 (0.11, 0.19)</td>
<td>0.13 (0.09, 0.18)</td>
</tr>
<tr>
<td>HUI2</td>
<td>1.24 (1.18, 1.31)</td>
<td>1.18 (1.11, 1.24)</td>
<td>0.09 (0.065, 0.12)</td>
<td>0.18 (0.15, 0.21)</td>
<td>0.11 (0.08, 0.14)</td>
<td>0.09 (0.05, 0.12)</td>
</tr>
<tr>
<td>SF-6D</td>
<td>1.89 (1.80, 1.98)</td>
<td>1.34 (1.28, 1.40)</td>
<td>0.09 (0.07, 0.11)</td>
<td>0.11 (0.09, 0.13)</td>
<td>0.08 (0.06, 0.10)</td>
<td>0.08 (0.06, 0.11)</td>
</tr>
<tr>
<td>eQWB</td>
<td>2.32 (2.22, 2.42)</td>
<td>1.42 (1.36, 1.49)</td>
<td>0.05 (0.04, 0.06)</td>
<td>0.06 (0.05, 0.07)</td>
<td>0.03 (0.02, 0.04)</td>
<td>0.05 (0.04, 0.07)</td>
</tr>
</tbody>
</table>

CI, confidence interval; ODI, Oswestry Disability Index; SRM, standardized response mean.
Unique Characteristics of Systems

In the absence of a clearly superior system, the combination of unique strengths and limitations incorporated into preference-weighted health state classification systems presents difficult tradeoffs for researchers considering system choice. The SF-6D demonstrated superior responsiveness and fewer ceiling problems for the pain and mobility dimensions. It is based on the longest of the surveys, with 36 questions, and covers several dimensions particularly relevant to persons with spine problems. Although the SF-36-derived approach may convey advantages in terms of responsiveness and ceiling effects, there is some indication that the SF-6D may provide higher values for more severe health states and therefore may undervalue interventions.

The ease of administration is a key advantage of the EQ-5D-UK and EQ-5D-US. However, floor and ceiling effects in dimensions highly relevant to spine disorders should be weighed against resource efficiency. The EQ-5D-UK does not provide health state values between 0.88 and 1 and has been shown to provide lower mean health state values for health states compared to other systems.14,36

The questionnaires for the HUI2 and HUI3 are of intermediate length compared to the SF-36 and EQ-5D, with 15 questions. The HUI3 incorporates dimensions not likely to be critical in the spine population, such as speech, vision, and hearing, and the dimensions covering mobility are limited to ambulation and dexterity. The HUI3 has potential limitations in characterizing diminished mobility other than ambulation.8 The HUI2 includes mobility and self-care dimensions which are important aspects of HRQOL for persons with spine disorders. Both the HUI3 and HUI2 demonstrated limitations in characterizing change in mobility in our study.

Practical issues should also be considered in choosing a measurement system. The tradeoffs between resources required for survey administration, acceptability to participants, and measurement properties must be considered carefully. The systems considered in this study range from 5 to 36 questions, with various response levels. Depending on the research context, these may represent important differences.

Comparisons with Other Studies

Although all systems demonstrated very similar patterns of psychometric performance across all criteria, some important differences emerged that can be compared to prior research. Although responsiveness statistics indicated acceptable performance for all systems, the eQWB and SF-6D demonstrated the highest responsiveness, as indicated by the larger effect size and SRM and smaller MIDs. The EQ-5D-UK had the lowest responsiveness of the measures. Other studies of the EQ-5D-UK and SF-6D in persons with various conditions have found similar results.34,51–54 Walters and Brazier48 found slightly better responsiveness for the SF-6D than the EQ-5D in the results of combined analysis of data from 11 cohorts. In the same study, the SRM and effect size were larger for the SF-6D than for the EQ-5D among patients with unspecified back pain. Longworth and Bryan50 found that the SF-6D was limited in capturing change for severe health states but was more responsive among better health states compared to the EQ-5D-UK among liver transplant patients. In contrast, Conner-Spady and Suarez-Almazor24 found slightly better responsiveness for the EQ-5D-UK compared to the HUI3 and SF-6D among rheumatology patients stratified by change status.

We found that responsiveness statistics for the HUI3 and HUI2 were similar to or slightly better than those for the EQ-5D-UK and EQ-5D-US and slightly worse than those for the SF-6D and eQWB. Studies conducted among persons with stroke, epilepsy, and heart disease have reported similar patterns.20,21,55 In contrast, Feeny and others22 found better responsiveness for the HUI3 and HUI2 than the SF-6D among patients undergoing hip replacement. The HUI3 demonstrated slight advantages over the HUI2 in responsiveness among patients undergoing breast reduction surgery.56

Responsiveness statistics, including MIDs, are important for study planning and for interpretation of changes in each system among patients with symptoms related to IDH. Estimates of effect size, SRM, and MID were generally larger in magnitude in our study than in other studies.20–22,24,30,34,51–55 This can be explained by the large functional health status changes in our population over the study period. Our MID results for those who reported minimal change were 0.08 for the SF-6D and 0.15 for the EQ-5D-UK, meaning that a mean change of 0.08 in a clinical study using the SF-6D would correspond with the lowest threshold for important change from the patient perspective using progress rating as the criterion. Alternatively, when judging the magnitude of change reported in clinical studies, a mean difference between treatment arms of 0.08 would indicate clinically meaningful difference using the SF-6D. However, using the EQ-5D-UK, the threshold would be 0.15.
Similarly, deficits in coverage for systems indicated by floor or ceiling problems have very important implications for system performance in measuring change over time. No ceiling effect was noted at baseline in our study. Consistent with previous studies conducted using data from persons with health conditions and from general population samples, large ceiling effects were noted in our study for the overall preference score for the EQ-5D-UK and EQ-5D-US at 1-year follow-up. Smaller but potentially significant ceiling effects were noted for the HUI3, HUI2, and SF-6D. These results would indicate that the EQ-5D-UK and EQ-5D-US may have difficulty characterizing change for long-term outcomes compared to other instruments.

Perhaps of greatest significance was the remarkable proportion of patients at the ceiling on all systems for the key dimensions targeted by treatment. All systems classified significant proportions of participants at the ceiling of the pain and mobility dimensions at baseline or follow-up. These dimensions are particularly relevant for measuring effects of treatment for this population. Floor effects were evident at baseline in the pain dimension for the EQ-5D-UK, EQ-5D-US, HUI3, HUI2, and SF-6D. Our review of the literature found that the SF-6D demonstrated floor effects and a limited range of available scores. This is consistent with other studies of the SF-36 in this population.

Floor effects were evident in the mobility dimension at baseline for the EQ-5D-UK, EQ-5D-US, HUI3, and HUI2 and in mobility and pain dimensions for all systems at 1 year. Ceiling effects were greater for the EQ-5D-UK and EQ-5D-US than for other systems in the pain dimension. In contrast, ceiling effects were greater for the HUI3 and HUI2 in the mobility dimension. Feeny and others suggested that the HUI3 may be limited in detecting changes in mobility that do not involve the hands. Our findings are consistent with this concern. However, the HUI2 performed similar to the HUI3 despite describing mobility in broader terms.

Although psychometric evaluations are fundamental in establishing the measurement characteristics of preference-weighted systems, it is critical to assess validity in the context of their application for policy decision making. To address this question, we compared estimates of mean score change, or mean change in health state value, since this estimate is fundamental to QALY calculation. Except for the EQ-5D-US and HUI2, we found that systems produced significantly different estimates of mean change in health state value. The EQ-5D-UK produced the largest estimate of mean change, followed by the HUI3, HUI2 and EQ-5D-US, SF-6D, and finally eQWB. Other studies have found differences in head-to-head longitudinal comparisons. Similarly, these studies reported that EQ-5D-UK estimates were generally largest, followed by HUI3, HUI2, and finally SF-6D. These patterns are generally consistent with the results of cross-sectional comparisons.

Although comparisons of mean health state values were more common, we identified comparisons of QALYs or ICERs obtained using relevant systems in the published literature. Pickard and others found that QALY differences calculated using the EQ-5D-UK or HUI3 were 2 times larger than those obtained using SF-6D or HUI2. Tosteson and others reported ICERs for the EQ-5D-UK and SF-6D in their cost-effectiveness analysis of surgery relative to non-operative treatment for persons with spinal stenosis with and without degenerative spondylolisthesis. The ICER (95% confidence interval [CI]) for spinal stenosis using the EQ-5D-UK was $77,600 ($49,564–$120,042) compared to $93,400 ($59,205–$143,660) using the SF-6D. The ICER for spinal stenosis with degenerative spondylolisthesis was $115,600 ($90,839–$144,863) using the EQ-5D-UK compared to $172,500 ($132,178–$221,930) using the SF-6D. Van den Hout and others reported change in QALYs for the EQ-5D-UK, EQ-5D-US, and SF-6D in their cost-utility analysis of early surgery v. prolonged conservative care among patients with sciatica from IDH. Their results were consistent with our findings: that EQ-5D-UK provided the largest change in health state values and the smallest cost-utility ratio, followed by the EQ-5D-US and SF-6D. The gains in QALYs were as follows: 0.044 (0.005–0.083), 0.032 (0.005–0.059), and 0.024 (0.003–0.046). Another study investigating acupuncture compared to usual care for nonspecific back pain reported very similar ICERs using the EQ-5D-UK and SF-6D. Joore and others found differences in ICERs and the probabilities of acceptability for the ICERs across 5 conditions using the EQ-5D-UK and SF-6D. Specifically, higher probabilities of acceptability were found using the EQ-5D-UK for milder health conditions and using the SF-6D for more severe health conditions. This study highlights the need to assess the performance of preference-weighted health state classification systems for specific conditions.

The results of our study indicate that using a regression model to “map” from SF-36-based health states to the QWB score for persons with IDH may be
a reasonable approach. The eQWB performed well in psychometric validity tests but provided the lowest estimation of mean change in health state value and very little variation in preference estimates. Kaplan and others\textsuperscript{29} reported similar results among patients with arthritis. Although the psychometric properties of the SF-6D and eQWB may be fairly similar, we recognize the additional steps of incorporating direct valuations as an advantage of the SF-6D over the “mapped” estimates produced by the eQWB. Furthermore, our results indicate that the eQWB may be more likely to produce qualitatively different cost-utility results than alternative systems. However, interest in developing methods to estimate health state values from existing HRQOL data appears to be increasing, and their relative performance should be investigated. It should be noted that the performance of the eQWB is dependent on the characteristics of the SF-36 health state classification system, the QWB scale, and the regression model used to link the two. Our results indicate that the eQWB estimates may be used with some caution in this population, but this should be weighed carefully against the advantages of the SF-6D.

Limitations

As with any validation studies of HRQOL instruments, there is no gold standard for the performance of preference-weighted health state classification systems. Although it is possible to generate and test hypotheses about the behavior of the systems under known circumstances, some variation would be expected between the systems under examination and the measure used to test validity.

We used a longitudinal cohort design to calculate responsiveness statistics, including effect size. This is not the same computation as is used to characterize the effect of treatment compared to control. It may be argued that for the purpose of interpretation in clinical trials, estimation of the effect size of treatment is the most relevant calculation. Because all patients in this trial undergo some treatment, either surgical or nonsurgical, and most were expected to improve, we measured observed change in the entire group. However, effect size is commonly used in applications similar to ours to address longitudinal validity.

Since the order of administration of instruments was not counterbalanced, we cannot rule out an order effect. However, because the questions in each instrument are of similar nature, it is doubtful that there would be significant learning or framing effects exhibited in this application.

In summary, this study provided information about the performance and interpretation of several of the most widely used preference-weighted health state classification systems. The evidence supports the notion that all systems are measuring the same construct, but each has unique characteristics that should be considered when choosing a system. We found evidence that all systems demonstrate validity in this population, with some caveats. All systems demonstrated evidence of ceiling or floor effects for key dimensions relevant to spine disorders. In the context of cost-effectiveness analysis, we found that change scores were significantly different except for the EQ-5D-US and HUI2. Change scores were largest for the EQ-5D-UK, followed by the HUI3, HUI2, EQ-5D-US, SF-6D, and finally eQWB. Such differences indicate that care should be taken when interpreting cost-utility analyses from different systems. Researchers choosing a system should carefully consider the characteristics of each system relative to study goals.

REFERENCES


279


