Comparison of EQ-5D, HUI, and SF-36-derived societal health state values among Spine Patient Outcomes Research Trial (SPORT) participants

Christine M. McDonough[1], Margaret R. Grove[1], Tor D. Tosteson[1], Jon D. Lurie[1], Alan S. Hilibrand[2], & Anna N.A. Tosteson[1]
Dartmouth Medical School, Community & Family Medicine, [1]Multidisciplinary Clinical Research Center in Musculoskeletal Diseases and the Center for Evaluative Clinical Sciences, Dartmouth Medical School (E-mail: Anna.N.A.Tosteson@dartmouth.edu) [2]Rothman Institute, Thomas Jefferson University, USA

Abstract

Purpose: To compare societal values across health-state classification systems and to describe the performance of these systems at baseline in a large population of persons with confirmed diagnosis of intervertebral disc herniation (IDH), spinal stenosis (SpS), or degenerative spondylolisthesis (DS). *Methods*: We compared values for EQ-5D (York weights), HUI (Mark 2 and 3), SF-6D, and the SF-36-derived estimate of the Quality of Well Being (eQWB) score using signed rank tests. We tested each instrument's ability to discriminate between health categories and level of symptom satisfaction. Correlations were assessed with Spearman rank correlations. We evaluated ceiling and floor effects by comparing the proportion at the highest and the lowest possible score for each tool. In addition, we compared proportions at the highest and lowest levels by dimension. The number of unique health states assigned was compared across instruments. We calculated the difference between those who were very dissatisfied and all others. *Results*: Mean values ranged from 0.39 to 0.63 among 2097 participants ages 18–93 (mean age 53, 47% female) with significant differences in pair-wise comparisons noted for all systems. Correlations ranged from 0.30 to 0.78. Although all systems showed statistically significant differences in health state values when baseline comparisons were made between those who were very dissatisfied with their symptoms and those who were not, the magnitude of this difference ranged widely across systems. Mean differences (95% CI) between those very dissatisfied and all others were 0.30 (0.269, 0.329) for EQ-5D, 0.22 (0.190, 0.241) for HUI(3), 0.18 (0.161, 0.201) for HUI(2), 0.11 (0.095, 0.117) for SF-6D, 0.04 (0.039, 0.049) for eQWB, and 0.07 (0.056, 0.077) for VAS (with transformation applied to group means). *Conclusion*: Differences in preference-weighted health state classification systems are evident at baseline in a population with confirmed IDH, SpS, and DS. Caution should be used when comparing health state values derived from various systems.

Key words: Economic evaluation, Health status indicators, Quality of Life, Spinal Diseases, SPORT

Introduction

Surgery accounts for a large proportion of the health care costs associated with spinal disorders [1]. Persons with spinal disorders have worse self-reported SF-36 physical health outcomes than those with cancer, congestive heart failure, and other chronic illnesses [2]. This combined with wide variation and increasing rates of spinal surgery [3–5] make economic evaluation of interventions for spinal disorders an important area of research. Societal health values for musculoskeletal problems have not been well studied. Unfortunately, studies of the important world of spine problems have been limited by lack of clear delineation of associated signs and symptoms into specific diagnostic groups. Cost-effectiveness studies for surgical interventions have relied on

utilities derived from small populations within the spine literature or general population studies [6, 7]. These utilities may not apply to persons with specific diagnoses across a more generalizable population. The Spine Patient Outcomes Research Trial (SPORT) is the first multi-center trial to study a clearly defined population of the three most common reasons for which spine surgery is performed; intervertebral disc herniation (IDH), spinal stenosis (SpS), and degenerative spondylolisthesis (DS). The primary aim of SPORT is to compare surgical and non-surgical treatment for persons with common spinal disorders for whom surgery is an option.

SPORT's secondary objective is to conduct a cost-effectiveness analysis of surgical vs. non-surgical treatments for these three diagnoses using QALYs as the effectiveness measure [8]. Quality-adjusted life years (QALYs), which account for both length and quality of life, are the recommended health outcome for use in economic evaluation when the objective is to inform public policy makers on allocation of limited resources. There is general consensus that for public policy applications the health state values used to estimate QALYs should be derived from a societal perspective [9]. As a result, several generic preference-weighted health state classification systems, which combine a health status measure with a societal value or utility for each health state [10, 11], have become widely used in economic studies. The most commonly used systems include the Health Utilities Index (HUI) [11–13], the EuroQoL EQ-5D [14–16], and the Quality of Well-Being Scale [17]. In addition, preference classification systems have been derived from the SF-36 health status measure [18–20].

Differences between systems exist in health state descriptions and the methods used to infer societal values. Recent studies comparing systems have described important differences, but have not suggested one superior tool for use in cost-effectiveness analyses [20–28]. No study has compared these systems for persons with confirmed diagnosis of IDH, SpS, or DS. Given that several equally supportable options exist, and that each instrument has unique properties, it is important to consider the public policy implications of using health state values derived from different systems. Therefore, the objective of this study is to compare the health state values derived from EQ-5D [15, 16]; HUI [11–13]; and two of the SF-36-derived preference classification systems–the SF-6D [29, 30] and estimated Quality of Well-Being score (eQWB) [19]–in a population of persons with these lumbar spine disorders, and to describe system performance at baseline.

Methods

Participants

Cross-sectional baseline data from an ongoing prospective study of interventions for back or leg symptoms associated with lumbar spine disorders (SPORT) are presented. The design of this study has been previously reported in detail [8]. In brief, SPORT is a multi-centre study including three randomized controlled trials and three observational cohorts. To be eligible for SPORT, participants must be 18 years or older and have a diagnosis of IDH, SpS, or DS. Participants are excluded if there is evidence of non-surgical 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 [8].

SPORT's intent is to characterize the impact of surgery on health-related quality of life using a generic health status instrument, the SF-36, [31, 32] and a disease-specific instrument, the Oswestry Disability Index (ODI) [33–35]. The SF-36 was developed in the Medical Outcomes Study for the measurement of health status. It measures eight dimensions on up to six levels. SPORT uses version 1 of the SF-36, with the reference time frame for health status of 'the past four weeks'. The ODI, an outcome measure of pain and physical function related to spinal disorders which includes ten items on six levels, yields an index score from least to most disability of 0–100. Participants first complete the ODI, followed by SF-36; EQ-5D (including a visual analogue scale (VAS) health rating); a symptom satisfaction rating (very dissatisfied, somewhat dissatisfied, neutral, somewhat satisfied, very satisfied); and HUI. Outcomes are measured at baseline, 6 weeks, and 3, 6, 12, and 24 months after initiation of treatment. Two

QALY estimates are derived from the SF-36: an estimated Quality of Well-Being score and the SF-6D.

Baseline measures of health state value

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

EQ-5D
The EuroQoL EQ-5D includes five attributes rated on three levels to define 245 health states (when 'dead' and 'unconscious' are added), and uses an additive model of attribute independence [15, 16]. Preferences were measured using time-tradeoff values for a subset of health states from a sample of the UK population [14]. The reference time frame for health status in the questionnaire was 'today'.

Health Utilities Index
The McMaster Health Utilities Index has been well described [11, 12, 36–38]. Permission is required for use of this instrument. SPORT is licensed to use HUI (Mark 2) and HUI (Mark 3). The reference time frame for the questionnaire was 'the past four weeks' and we did not include the fertility dimension in our survey. HUI(2) represents seven attributes on four or five levels and defines 24,000 health states. HUI(3) has five or six levels for each of its eight attributes and encompasses 972,000 unique health states. The HUI(2) and HUI(3) use multiplicative multi-attribute utility functions based on visual analogue and standard gamble scores obtained from community samples in Canada [11, 37, 38].

SF-36-derived measures
The SF-6D, version 2, provides a method for deriving a preference score from the SF-36 instrument [29, 30]. It represents six attributes on up to six 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. The Quality of Well-Being scale (QWB) is a preference-based health measure that includes three additive functional dimensions and a symptom dimension [17]. We estimated the QWB score using five subscales of the SF-36 based on regression analysis reported by the Beaver Dam

Health Outcomes Study [19] and denote it as eQWB.

Visual analogue scale
Subjects are asked to rate their current health on a thermometer-type scale from 0, representing worst possible health, to 100, representing best possible health. The distance from 0 to the subject's mark is measured and used as the value estimate [9]. We used a transformation of VAS as described by Torrance et al. [39], which was applied to group means: $u = 1-(1-v)^{2.3}$

Statistical analyses

Demographic characteristics of study participants were described using means, standard deviations and percents. Summary statistics were estimated for each preference-weighted health state classification system. Comparisons between the distributions for each system pair were made using Wilcoxon signed rank tests. Construct validity for each system was assessed with a test for trend across self-perceived health categories (i.e., excellent, very good, good, fair, poor), quintiles of ODI, and levels of dissatisfaction. We hypothesized that values would decrease monotonically between health categories and ODI quintiles, and that the instruments would discriminate between respondents who were very dissatisfied with symptoms and those who were not. In addition, we tested each instrument's ability to discriminate between health categories and levels of symptom satisfaction (very dissatisfied vs. all others), and present 95% confidence intervals for differences between groups. Spearman correlation coefficients were used to describe the strength of association between instruments. We evaluated ceiling and floor effects for each measure by comparing the proportion of participants who attained the highest and lowest possible scores for each tool. We also compared the proportion of participants at the highest and lowest levels for the key dimensions of pain and physical function. To characterize the range of health states utilized by our study population, the number of unique health states assigned using each of the instruments were compared. Although analyses were conducted separately for each disease group, we present analyses for all disease groups pooled due to the

consistency of findings across disease groups. All analyses were undertaken using STATA, version 7 (STATA Corporation, College Station, Texas).

Results

The 2097 participants ranged in age from 18 to 93 years old, with an average age of 53 years. Overall, they were a highly educated group that comprised of both men and women, with the majority having a diagnosis of intervertebral disc herniation as the underlying cause for their back or leg symptoms (Table 1).

Mean preference scores ranged from 0.39 for EQ-5D to 0.63 for VAS (Table 2). Significant differences between baseline mean health state values were noted for all pair-wise system comparisons. The distribution of preference scores obtained with each instrument is provided in Figure 1. In contrast to all other instruments, the EQ-5D appeared bimodal. EQ-5D and HUI(3) and HUI(2) covered a larger range than the two SF-36-derived tools.

Each instrument demonstrated significant linear downward trends in health state values with poorer self-rated health (Figure 2a) and higher disability as measured by ODI score in quintiles (Figure 2b). Mean health state values increased as symptom satisfaction increased (Figure 2c). EQ-5D, HUI(3), HUI(2) and SF-6D could not discriminate between those reporting excellent and very good health while eQWB and VAS could. All instruments except eQWB could discriminate between fair and poor health. All instruments showed significant differences in means between those who were very dissatisfied with their current symptoms relative to all others. Mean differences between those very dissatisfied and all others were: 0.30 (0.269, 0.329) for EQ-5D; 0.22 (0.190, 0.241) for HUI(3); 0.18 (0.161, 0.201) for HUI(2); 0.11 (0.095, 0.117) for SF-6D; 0.04 (0.039, 0.049) for eQWB; and 0.07 (0.056, 0.077) for VAS (with transformation applied to group means).

Spearman correlations between instruments ranged from 0.30 to 0.78 (Table 3). The highest correlations were among EQ-5D, HUI(3), HUI(2) and SF-6D. Moderate correlations were found between these four tools and eQWB. Modest correlations were noted between VAS and all other instruments.

The total number of unique health states assigned using EQ-5D, HUI(3), HUI(2) and SF-6D were 72, 674, 451, and 1012, respectively. Fewer than 1% of participants scored at the floor or the ceiling values for all tools. Table 4 describes the proportion of participants by level and attribute for the tools that can be decomposed. For EQ-5D 45.5% of participants responded at the most severe level on at least one dimension. In the pain dimension, EQ-5D had 1.9% at the least severe level while the other instruments had less than one percent. EQ-5D had 38.0% at the most severe level

Table 1. Summary of participant characteristics at baseline

	All participants n = 2097 (%)
Age	
<50 years	43.5
50–64 years	27.4
65 years and over	29.1
Female Sex	47.5
Body Mass Index ≥29	41.0
Disease Group	
Intervertebral Disc Herniation	51.0
Spinal Stenosis	26.2
Degenerative Spondylolisthesis	22.8
Enrolled in RCT	34.1
Ethnicity: Hispanic	2.0
Race	
Black/African American	8.3
White	84.7
Other (including unknown)	7.0
Education: College graduate	42.7
Married or living with a significant other	68.8
Insurance coverage[a]	97.8
Employer or union	44.6
Former employer	11.3
Spouse's insurance	23.8
Medicare	28.4
Medicaid	2.6
Other/private purchase	16.1
Satisfaction with current symptoms	
Very dissatisfied	75.2
Not very dissatisfied	24.8
ODI (Mean ± SD)	45.8 ± 20.4

[a]Among those who had insurance, 74.2% had one source of insurance coverage, 24.8% had two sources and 1.0% had three.

Table 2. Summary statistics for each instrument at baseline

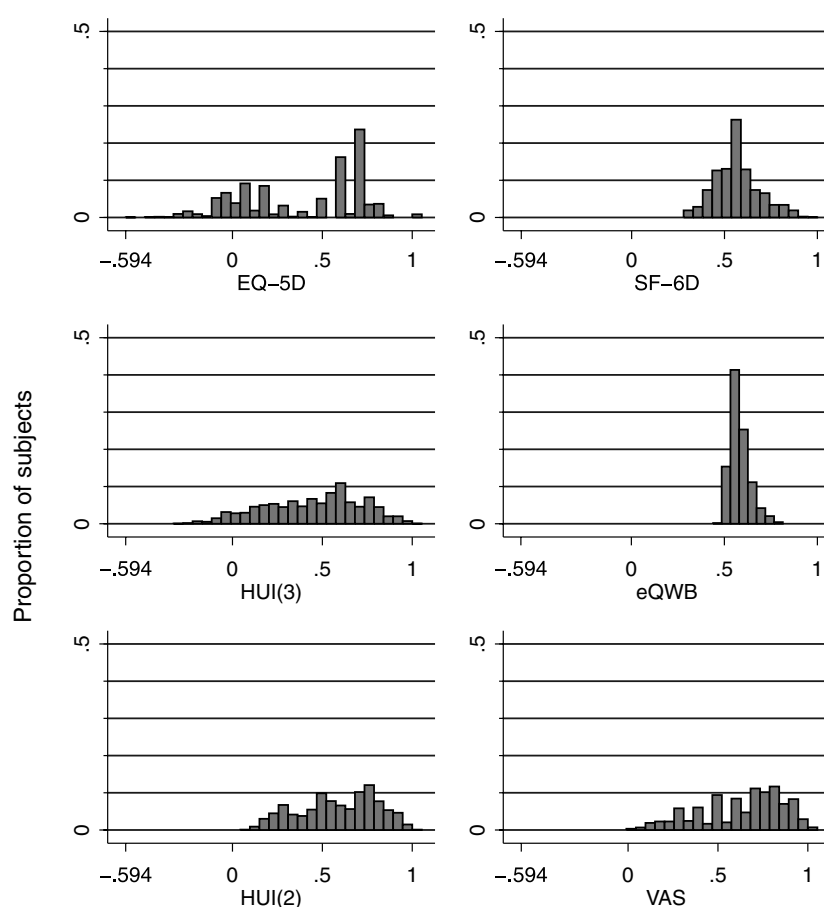|	EQ-5D n = 2097	HUI(3) n = 2097	HUI(2) n = 2097	SF-6D n = 2097	eQWB n = 2097	VAS[a] n = 2097
Mean ± SD	0.39 ± 0.33	0.45 ± 0.27	0.59 ± 0.22	0.57 ± 0.12	0.58 ± 0.05	0.63 ± 0.22
Median	0.52	0.48	0.60	0.56	0.57	0.70
Interquartile Range (25%, 75%)	(0.08, 0.69)	(0.23, 0.66)	(0.43, 0.77)	(0.49, 0.63)	(0.54, 0.61)	(0.50, 0.80)
Range	−0.594, 1	−0.31, 1	0.07, 1	0.296, 0.948	0.480, 0.814	0.01, 1

[a] Untransformed VAS.

Figure 1. Frequency distributions for instruments at baseline among 2097 SPORT participants. VAS distribution represents untransformed baseline values.

compared to 24.2% for HUI(3), 15.9% for HUI(2) and 22.4% for SF-6D. In the area of physical function, EQ-5D had 14.6% at the least severe level, HUI(3) and HUI(2) had 25.1%, and SF-6D had 1.1%. For the most severe level, EQ-5D had 2%, HUI(3) had 1.2%, HUI(2) had 0%, and SF-6D had 11%.

Discussion

Our analysis of baseline health state values among more than 2000 individuals with lumbar spine disorders indicated a high degree of correlation between preference-weighted health state classification systems. Nonetheless, many differences in
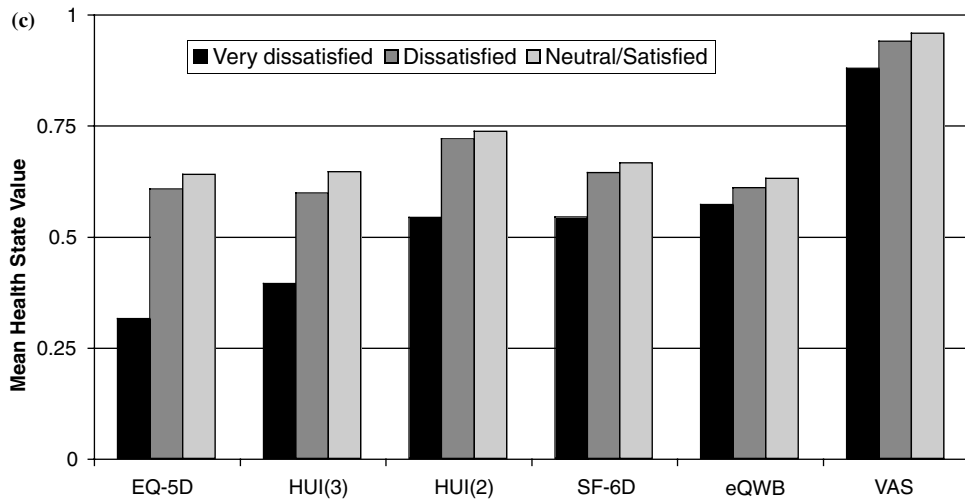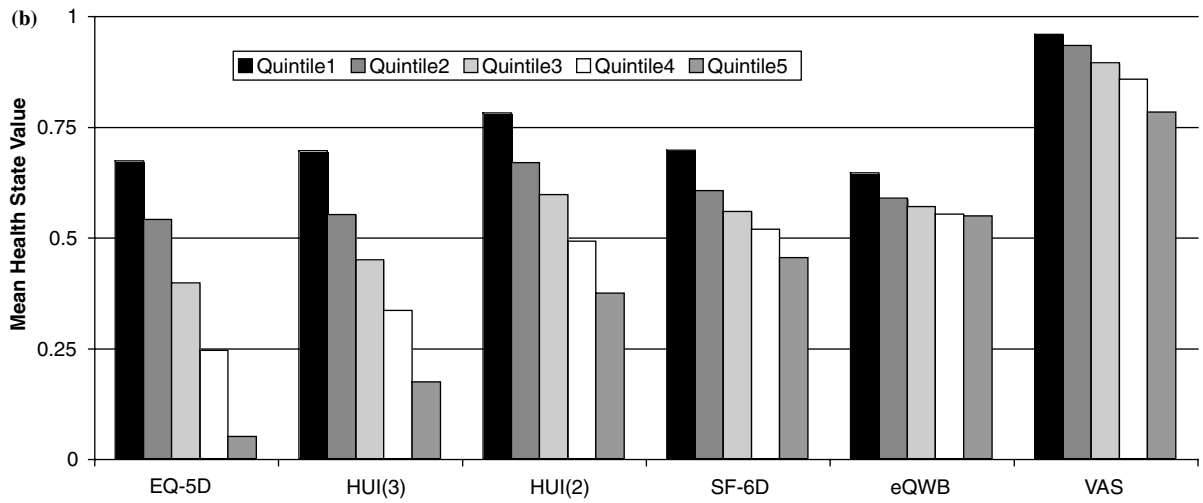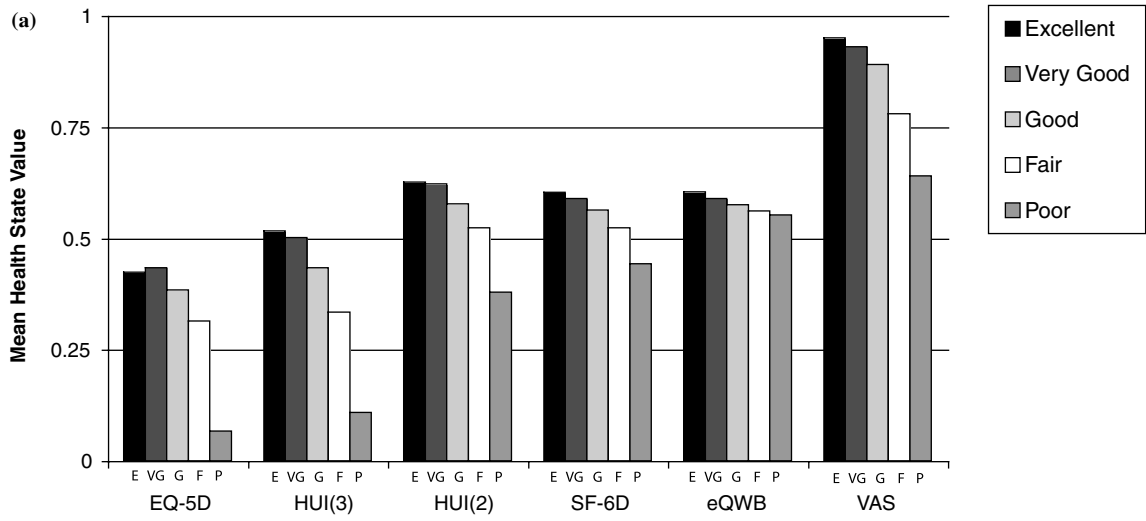
(a)

(b)

(c)

Table 3. Spearman rank correlations between instruments at baseline

	HUI(3)	HUI(2)	SF-6D	eQWB	VAS
EQ-5D	0.67	0.68	0.69	0.52	0.41
HUI(3)		0.78	0.72	0.53	0.45
HUI(2)			0.67	0.51	0.41
SF-6D				0.53	0.46
eQWB					0.30

All p-values < 0.001.

mean health state values between systems were noted. Although all systems showed statistically significant differences in health state values when baseline comparisons were made between those who were very dissatisfied with their symptoms and those who were not, the magnitude of this difference ranged widely across systems (e.g., from 0.30 for EQ-5D to 0.04 for eQWB). While longitudinal performance is critical in the determination of health state values, it is worthwhile to consider the role of these baseline differences in final estimates of cost per QALY gained. Were these differences to carry over into a formal incremental cost-effectiveness analysis, the potential for qualitatively different results is readily apparent.

Other studies have compared health state values derived from these systems; however, ours included all of these instruments in a large head-to-head comparison of persons with back and leg symptoms with specific, confirmed diagnoses. Hollingworth et al. [20] studied SF-36-derived health valuations of participants with low back pain at baseline and concluded that instrument choice may have a significant effect on health state estimates and effect size calculations. Suarez-Almazor et al. [28], in their longitudinal study of low back pain, reported a difference in means of 0.11 for EQ-5D and HUI(2), but the sample size was too small for the difference to reach statistical significance. Gabriel

◄

Figure 2. Trends in mean health state values with global health, physical disability, and symptom satisfaction at baseline. (a) Trends in mean health state values with global health rating by instrument. VAS trends shown are for transformed group means. (b) Trends in mean health state values with ODI quintiles (quintile 1 is lowest disability). VAS trends shown are for transformed group means. (c) Trends in mean health state values with level of symptom satisfaction by instrument. VAS trends shown are for transformed group means.

et al. [24] reported a statistically significant difference between eQWB and HUI(2) values in their cross-sectional study of 382 participants with osteoporotic fracture. Belanger et al. [21] reported a statistically significant difference between mean EQ-5D and HUI(3) values in a subgroup of respondents to their general health survey that reported chronic back pain. Conner-Spady et al. found a range of 0.13 in mean health state values at baseline in their longitudinal studies of EQ-5D, HUI, and SF-6D in persons with musculoskeletal disease [23, 40]. In contrast, Brazier et al. found little difference in mean health state values between EQ-5D and SF-6D [22]. Figure 1 illustrates the differences in the distribution of values at baseline. The range of values covered by EQ-5D and HUI(3) is large in contrast with SF-6D and eQWB, which may indicate a potential limitation for the SF-36-derived instruments.

Our study highlights some of the challenges in measuring the validity of health measurement tools without a gold standard. To address this problem, we tracked instrument performance with self-perceived overall level of health, ODI quintiles, and level of satisfaction with symptoms. The instruments functioned as expected in their trends with each of these measures. There were some differences between instruments in their ability to discriminate between levels of health. For example, eQWB could not distinguish between fair and poor health. Hollingworth et al. reported similar findings for eQWB and SF-6D [20]. Overall, there was little evidence for a superior instrument, though the limitations noted for the eQWB suggest that it is somewhat inferior. These observations do not apply to the actual QWB Index, which was not used in this study.

Though correlation between instruments is often used as a measure of convergent validity, there is no consensus on the interpretation of various levels of correlations for this application. Suarez-Almazor reported low correlations between EQ-5D, HUI, and SF-36 and suggested that they measure somewhat different concepts of health. Having been developed for different purposes, it is not surprising that various tools would reach only modest correlations. In their review of health measurement instruments, McDowell and Newell [41] describe why a correlation of 0.60 can be viewed as extremely high, given that correlation is

Table 4. Percentage of responses by dimension and level for measures at baseline

Instrument	Percent of responses for each level of severity[a]					
EQ-5D	Level 1		Level 2			Level 3
Mobility	14.6		83.4			2.0
Self care	61.4		37.9			0.7
Usual Activities	10.5		69.7			19.8
Pain/discomfort	1.9		60.1			38.0
Anxiety/depression	55.0		41.0			4.1
HUI(3)	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
Vision	25.6	71.1	0.9	1.4	0.8	0.2
Hearing	90.9	1.6	2.0	3.3	1.0	1.2
Speech	95.8	2.2	2.0	0.1	0.0	NA[b]
Cognition	67.0	5.2	12.5	12.3	3.1	0.0
Emotion	37.6	30.0	22.9	8.8	0.8	NA
Pain	0.3	6.4	24.6	44.4	24.2	NA
Ambulation	25.1	54.5	8.6	8.3	2.3	1.2
Dexterity	90.8	7.6	0.6	0.8	0.2	0.1
HUI(2)	Level 1	Level 2	Level 3	Level 4	Level 5	
Sensation	24.7	64.6	9.3	1.4	NA	
Mobility	25.1	54.5	16.8	3.5	0.0	
Emotion	38.8	47.1	11.2	2.8	0.1	
Cognition	67.0	30.6	2.5	0.0	NA	
Self-care	62.3	36.0	0.4	1.3	NA	
Pain	0.7	17.6	36.4	29.3	15.9	
SF-6D	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
Physical functioning	1.1	5.3	20.2	49.7	12.6	11.0
Role limitations	10.1	39.9	2.5	47.5	NA	NA
Social functioning	17.6	20.0	29.9	23.3	9.2	NA
Pain	0.3	2.5	10.3	23.7	40.8	22.4
Mental health	17.6	30.6	34.0	14.7	3.2	NA
Vitality	1.7	17.0	32.4	27.8	21.2	NA

[a] Numbers of levels are 3 for EQ-5D dimensions, 5 or 6 for HUI(3) dimensions, 4 or 5 for HUI(2) and 4 to 6 for SF-6D dimensions; the highest level for each dimension is the most severe.
[b] NA indicates not applicable.

limited to the square root of the product of the instruments' reliability. In this context, we found very high correlations between EQ-5D, HUI(3), HUI(2) and SF-6D. Conner-Spady et al. reported similar correlation coefficients at baseline for EQ-5D, HUI(3) and SF-6D [23, 40]. Intermediate level correlations were found in our study between eQWB and HUI(3), HUI(2), EQ-5D, and SF-6D. The lowest correlations were noted between VAS and all other instruments (Table 3). These findings indicate that EQ-5D, HUI(3), HUI(2), and SF-6D measure essentially the same construct.

Our validity tests support the idea that all of the instruments measure quality of life to some degree, but they do not provide sufficient information to guide the selection of one instrument over another. Where no clearly superior method exists, both practical and design aspects of measurement tools may indicate which tool would be better suited for measurement of the quality of life of persons with back and leg symptoms. For example, the reference time frame used in the descriptive system varies among tools. The EQ-5D asks the respondent to rate health status today, while others refer to health status in a larger time frame: over the past few months or 'usual.' However, some instruments, such as HUI, have multiple versions available. It would be important to consider how these

characteristics fit the nature of the study and the timing of testing. In the case of back and leg symptoms, it may be necessary to differentiate between usual and present status due to exacerbation.

Brazier et al. [29] have reported that one theoretical advantage of the SF-6D over EQ-5D is its larger SF-36-based descriptive system, but suggested that future research should assess whether item selection for SF-6D caused limitations to its descriptive ability and indicates that SF-6D may not benefit from the descriptive richness of the original SF-36. It has also been suggested that one advantage of HUI is the large number of distinct health states it describes [12]. We observed that SF-6D, HUI(3), and HUI(2) identified much larger numbers of unique health states than EQ-5D, which is not surprising since EQ-5D has fewer available health states. It is noteworthy that the large range of unique health states available from SF-6D is not reflected in a wide range of health state values.

In theory, EQ-5D extracts information from five distinct health dimensions, HUI(3) from eight, and SF-6D from six. From correlation analyses, Belanger et al. [21] concluded that EQ-5D represents fewer than five independent dimensions, while HUI(3) essentially addresses all eight. Keeping in mind that this decomposed analysis fits HUI(3)'s design better than that of EQ-5D, it still raises the question of what EQ-5D and HUI(3) are measuring relative to different dimensions. For example, in the study by Belanger et al. for the three dimensions HUI(3) and EQ-5D have in common, people were less likely to report perfect health on the EQ-5D than on the HUI(3) and in the pain dimension, which is an important component for the back pain population, 80.4 % reported perfect health on the HUI compared to 58.9% on the EQ-5D [21]. Our study identified differences in tools at the highest and lowest levels of physical function and at the most severe level for pain, but not at the level of no pain as in Belanger et al.'s study.

For those with back and leg symptoms, average baseline health is well below that of the general population [2]; therefore, the ability to characterize health status at the lowest levels is particularly important. Hollingworth et al. [20] noted the potential for floor effects in all of the SF-36-derived measurement tools. Suarez-Almazor et al. [28] reported a floor effect in SF-36 subscales for participants who had deteriorated. Brazier et al. [22] reported that for osteoarthritis participants 92% reported the lowest level of physical function on the SF-6D while none did on the EQ-5D. Taylor et al. [42] found floor effects in the role physical and role emotional subscales of the SF-36 and recommended using version 2 of the SF-36 to eliminate this problem. We used version 1 of the SF-36. We did not find evidence of a floor effect for index values, but noted that each instrument had large proportions of participants at the floor for either pain or physical function (Table 4). Most notable was that 38% of participants reported the most severe level of pain on the EQ-5D. Concern about ceiling effects for milder health states have been described for EQ-5D [21, 27]. In contrast to the use of these instruments for measuring general population health, our study provides evidence that the previously observed ceiling effect for EQ-5D is not a limitation among persons with back and leg symptoms who have specific diagnoses (IDH, SpS, and/or DS). Hollingworth et al. [20] described an artificial ceiling for eQWB as well, noting a maximum possible score of 0.84. In our study, all other instruments produced health valuations above eQWB's highest value. Though we would note that the limited range of values available from the eQWB and SF-6D is a limitation of the instruments, our study did not find evidence of a ceiling effect *per se*. In the pain dimension none of the tools demonstrated a significant proportion at the ceiling. However, in the mobility dimension 25.1% of participants were at the ceiling for HUI(3) and HUI(2). This compares to 14.6% for EQ-5D and 1.1% for SF-6D. Feeny et al. [36] have suggested that HUI(3) may not detect the impact of musculoskeletal problems that do not involve the lower limbs or hands and fingers. It is unclear whether the differences are due to these issues.

In our study, EQ-5D described far lower quality of life than all of the other instruments. Prior reports have been conflicting. Brazier et al. [22] reported similar mean health state values between EQ-5D and SF-6D among a sub-group of back pain patients. In their analysis, they noted the implications of the 'N3 term' in the EQ-5D York scoring function, which subtracts an additional 0.269 from the score for persons at the lowest level in any dimension. The SF-6D has a similar scoring function, but the coefficient associated with a

response in one of the most severe levels is −0.032, which is much smaller than the EQ-5D N3 term. Our findings are consistent with Conner-Spady et al. [23, 40], who found a lower baseline mean health state value from EQ-5D, as well as the greatest health state difference over 12 months. This issue is particularly important for measurement of a population such as ours, with health problems that dramatically affect pain and function. Brazier et al. found that the pain dimension heavily influences the index values for EQ-5D and SF-6D [22]. This, combined with the large proportion (45%) of participants using the N3 term, implies that the N3 term could have accounted for the bimodal EQ-5D distribution and contributed heavily to the lower mean health state value for EQ-5D as reported in our study. By lowering baseline values, this scoring function could cause an increase in the change scores ultimately used to estimate incremental cost-effectiveness ratios in economic evaluations [23, 40, 43]. Nevertheless, the question remains whether the N3 term appropriately describes or inappropriately exaggerates severe health states.

Our study has several limitations that warrant comment. First, although a definitive answer to the question, 'which instrument measures change best?' requires longitudinal data, our findings suggest that choice of instrument may matter. To date only three relatively small longitudinal studies of preference-based measurement tools have been reported for persons with musculoskeletal problems [23, 28, 40]. One raises important issues regarding differences in responsiveness of EQ-5D, HUI, and SF-36 in persons with chronic low back pain whose condition deteriorated [28]. Two other studies reported that EQ-5D, HUI and SF-6D appeared to measure the same construct, but that health state values and change scores differed between instruments [23, 40]. None of these studies addressed the very specifically defined population of participants included in SPORT. Second, our study did not explore the effects of valuation method on health state values. Conner-Spady et al. have reported that different scaling methods can change potential QALYs gained [44]. Differences exist in the methods used to value health states in each of the instruments. Third, the design of our study did not allow us to estimate an instrument order effect. The extent to which a change in instrument order would influence baseline health state values is unknown. Finally, our study population provides information about persons with confirmed diagnosis of spinal disorders (IDH, SpS, DS) for whom surgery is an appropriate treatment option. Our analysis includes both those who consented to random treatment allocation within the RCT and those who agreed to participate in an observational cohort. This makes our study more generalizable than a typical RCT. However, whether our findings can be generalized to persons with chronic, non-specific back pain for whom surgery is not indicated, or to persons with problems unrelated to the low back, is unknown.

In summary, we found important differences in health state values derived from different preference-based health status measurement instruments at baseline, which could have an impact on incremental cost-effectiveness ratios and public policy. Our study contributes to the evidence that there may not be a superior instrument. For now, researchers should choose a measurement tool that best fits the condition under investigation and their study design. Our analysis suggests that the eQWB may be more limited for describing persons with debilitating health states such as back and leg symptoms associated with lumbar spine disorders than are EQ-5D and HUI. Longitudinal data from SPORT will provide more definitive information about the performance of these instruments in measuring change in populations with confirmed lumbar spine disorders. In particular, the ability of the instruments to measure change at the lowest health levels will be very important in defining effectiveness for use in economic evaluation for this population. Nonetheless, at the present time, it should be noted that health state values derived from different tools are not interchangeable, and therefore caution should be used when comparing incremental cost-effectiveness ratios across studies.

Investigator/Site list

James N. Weinstein, Principal Investigator (SPORT)
Dartmouth Medical School
Hanover, New Hampshire 03755

Dartmouth-Hitchcock Medical Center, Lebanon, NH
    William A Abdu, MD, MS
New York University Medical Center's Hospital for Joint Diseases, NY, NY
    Thomas Errico, MD
Hospital for Special Surgery, NY, NY
    Frank P. Cammisa, MD; Harvinder Sandhu, MD
Kaiser-Permanente, Oakland, CA
    Harley Goldberg, DO
Nebraska Foundation for Spinal Research, Omaha, NE
    Michael Longley, MD
Rothman Institute @ Thomas Jefferson University, Philadelphia, PA
    Todd Albert, MD/Alan Hilibrand, MD
Rush-Presbyterian-St. Luke's Medical Center, Chicago, IL
    Gunnar Andersson, MD, PhD/Howard An, MD
The Emory Clinic, Decatur, GA
    Scott Boden, MD
University of California Medical Center, San Francisco, CA
    Serena Hu, MD
University Hospitals of Cleveland, Cleveland, OH
    Sanford Emery, MD
Washington University Hospital, St. Louis, MO
    Lawrence Lenke, MD
William Beaumont Hospital, Royal Oak, MI
    Harry Herkowitz, MD

References

1. Gibson J, Grant I, Waddell G. Surgery for lumbar disc prolapse. Cochrane Database of Systematic Reviews 2000(4).
2. Fanuele JC, Birkmeyer NJ, Abdu WA, Tosteson TD, Weinstein JN. The impact of spinal problems on the health status of patients. Spine 2000; 25(12): 1509–1514.
3. Cherkin D, Deyo R, Loeser J, Bush T, Waddell G. An international comparison of back surgery rates. Spine 1994; 19(11): 1201–1206.
4. Conditions of the Spine. In: Birkmeyer JD (ed), The Dartmouth Atlas of Musculoskeletal Health Care. Chicago: AHA Press; 2000. 28–58.
5. Ciol M, Deyo R, Howell E, Kreif S. An assessment of surgery for spinal stenosis: Time trends, geographic variations, complications, and reoperations. J Am Geriatr Soc 1996; 44(3): 285–290.
6. Malter AD, Larson EB, Urban N, Deyo R. Cost-effectiveness of lumbar discectomy for the treatment of herniated intervertebral disc. Spine 1996; 21(9): 1048–1054.
7. Kuntz KM, Snider RK, Weinstein JN, Pope MH, Katz JN. Cost-effectiveness of fusion with and without instrumentation for patients with degenerative spondylolisthesis and spinal stenosis. Spine 2000; 25(9): 1132–1139.
8. Birkmeyer NJ, Weinstein JN, Tosteson AN, et al. Design of the Spine Patient Outcomes Research Trial (SPORT). Spine 2002; 27(12): 1361–1372.
9. Gold MR, Siegel JE, Russell LB, Weinstein MC. Cost-Effectiveness in Health and Medicine. New York: Oxford University Press; 1996.
10. Brazier J, Deverill M, Green C, Harper R, Booth A. A review of the use of health status measures in economic evaluation. Health Technol Asses 1999; 3(9): 57–81.
11. Torrance GW, Furlong W, Feeny D, Boyle M. Multiattribute preference functions. Health Utilities Index. Pharmacoeconomics 1995; 7(6): 503–520.
12. Feeny D, Furlong W, Torrance GW, et al. Multiattribute and single-attribute utility functions for the Health Utilities Index Mark 3 System. Med Care 2002; 40(2): 113–128.
13. Boyle MH, Furlong W, Feeny D, Torrance GW, Hatcher J. Reliability of the health utilities index–mark III used in the 1991 cycle 6 Canadian general social survey health questionnaire. Qual Life Res 1995; 4(3): 249–257.
14. Dolan P. Modelling valuations for EuroQol health states. Med Care 1997; 35(11): 1095–1108.
15. Brooks R. EuroQol: The current state of play. Health P 1996; 37(1): 53–72.
16. Dolan PR J. Modelling valuations for EQ-5D health states. An alternative model using differences in valuations. Med Care 2002; 40(5): 442–446.
17. Kaplan RM, Anderson JP. A general health policy model: update and applications. Health Serv Res 1988; 23(2): 203–235.
18. Fryback DG, Dasbach EJ, Klein R, et al. The Beaver Dam Health Outcomes Study: Initial catalogue of health-state quality factors. Med Decis Mak 1993; 13(2): 89–102.
19. Fryback DG, Lawrence WF, Martin PA, Klein R, Klein BE. Predicting Quality of Well-Being scores from the SF-36: Results from the Beaver Dam Health Outcomes Study. Med Decis Mak 1997; 17(1): 1–9.
20. Hollingworth W, Deyo RA, Sullivan SD, et al. The practicality and validity of directly elicited and SF-36 derived health state preferences in patients with low back pain. Health Econ 2002; 11(1): 71–85.
21. Belanger A, Berthelot J-M, Guimond E, Houle C. A Head-to-head comparison of two generic health status measures in the household population: McMaster Health Utilities Index (Mark 3) and the EQ-5D. Ottawa: Statistics Canada,

1332

Health Analysis and Modelling Group; 2000 Final Revision April 2000.

22. 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; 9–31.

23. 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; 235–245.

24. Gabriel SE, Kneeland TS, Melton LJ, IIIrd, et al. Health-related quality of life in economic evaluations for osteoporosis: whose values should we use? Med Decis Mak 1999; 19(2): 141–148.

25. Glick HA, Polsky D, Willke RJ, Schulman KA. A comparison of preference assessment instruments used in a clinical trial: Responses to the visual analog scale from the EuroQol EQ-5D and the health utilities index [erratum appears in Med Decis Mak 1999 Oct–Dec; 19(4): 511]. Med Decis Mak 1999; 19(3): 265–275.

26. Krahn M, Ritvo P, Irvine J, et al. Patient and community preferences for outcomes in prostate cancer: Implications for clinical policy. Medical Care 2003; 41(1): 153–164.

27. Macran S, Weatherly H, Kind P. Measuring population health: A comparison of three generic health status measures. Med Care 2003; 41(2): 218–231.

28. Suarez-Almazor ME, Kendall C, Johnson JA, Skeith K, Vincent D. Use of health status measures in patients with low back pain in clinical settings. Comparison of specific, generic and preference-based instruments. Rheumatology 2000; 39(7): 783–790.

29. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ 2002; 21(2): 271–292.

30. Brazier J, Usherwood T, Harper R, Thomas K. Deriving a preference-based single index from the UK SF-36 Health Survey. J Clin Epidemiol 1998; 51(11): 1115–1128.

31. Ware JE, Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992; 30(6): 473–483.

32. Shmueli A. The SF-36 profile and health-related quality of life: an interpretive analysis. Qual Life Res 1998; 7: 187–195.

33. Fairbank JC, Pynsent PB. The oswestry disability index. Spine 2000; 25(22): 2940–2952; Discussion 2952.

34. Fairbank JC, Couper J, Davies JB, O'Brien JP. The Oswestry low back pain disability questionnaire. Physiotherapy 1980; 66(8): 271–273.

35. Bombardier C, Hayden J, Beaton DE. Minimal clinically important difference. Low back pain: Outcome measures. J Rheumatol 2001; 28(2): 431–438.

36. Feeny D, Furlong W, Boyle M, Torrance GW. Multi-attribute health status classification systems. Health utilities index. Pharmacoeconomics 1995; 7(6): 490–502.

37. Tosteson AN. Preference-based health outcome measures in low back pain. Spine 2000; 25(24): 3161–3166.

38. Tosteson AN, Hammond CS. Quality-of-life assessment in osteoporosis: health-status and preference-based measures. Pharmacoeconomics 2002; 20(5): 289–303.

39. Torrance GW, Zhang Y, Feeny D, Furlong W, Barr R. Multi-attribute preference functions for a comprehensive health status classification system. McMaster University Centre for Health Economics and Policy Analysis 1992; Working Paper Series No. 92–18.

40. Conner-Spady B, Suarez-Almazor ME. Variation in the estimation of quality-adjusted life-years by different preference-based instruments. Med Care 2003; 41(7): 791–801.

41. McDowell I, Newell C. Measuring Health: A Guide to Rating Scales and Questionnaires. New York: Oxford University Press; 1996.

42. Taylor SJ, Taylor AE, Foy MA, Fogg AJB. Responsiveness of common outcome measures for patients with low back pain. Spine 1999; 24(17): 1805–1812.

43. Cronbach LJ, Furby L. How should we measure 'change' - or should we? Psychol Bull 1970; 74(1): 68–80.

44. Conner-Spady B, Voaklander DC, Suarez-Almazor ME. The effect of different EuroQol weights on potential QA-LYs gained in patients with hip and knee replacement. 17th Plenary Meeting of the EuroQol Group; 2000. 127–137.

Address for correspondence: Anna N.A. Tosteson, HB 7505 Clinical Research, Dartmouth-Hitchcock Medical Center, One Medical Center Drive, Lebanon, NH 03756, USA
Phone: +1 603 653 3568; Fax: +1 603 653 3554
E-mail: Anna.N.A.Tosteson@dartmouth.edu