International Journal of Technology Assessment in Health Care, **19:2** (2003), 362–372.

COMPARING COMMUNITY-PREFERENCE–BASED AND DIRECT STANDARD GAMBLE UTILITY SCORES: EVIDENCE FROM ELECTIVE TOTAL HIP ARTHROPLASTY

David Feeny
University of Alberta

Christopher Blanchard
American Cancer Society

Jeffrey L. Mahon
Robert Bourne
Cecil Rorabeck
Larry Stitt
Susan Webster-Bogaert
University of Western Ontario

Abstract

Objectives: Do utility scores based on patient preferences and scores based on community preferences agree? The purpose is to assess agreement between directly measured standard gamble (SG) utility scores and utility scores from the Health Utilities Index Mark 2 (HUI2) and Mark 3 (HUI3) systems.
Methods: Patients were assessed repeatedly throughout the process of waiting to see a surgeon, waiting for surgery, and recovery after total hip arthroplasty (THA). Group mean scores are compared using paired t-tests. Agreement is assessed using the intraclass correlation coefficient (ICC).
Results: The mean SG, HUI2, and HUI3 (SD) scores at assessment 1 are 0.62 (0.31), 0.62 (0.19), and 0.52 (0.21); n = 103. At assessment 2, the means are 0.67 (0.30), 0.68 (0.30), and 0.58 (0.22); n = 84. There are no statistically significant differences between group mean SG and HUI2 scores. Mean SG and HUI3 scores are significantly different. ICCs are low.

Financial Support: The study, "The Effect of Waiting for Elective Hip Arthroplasty on Health-Related Quality of Life," was supported by a grant from Physician Services Incorporated (PSI) of Ontario to Dr. Jeffrey Mahon (Grant 94-30). The analyses reported in this study were supported by grants from the Alberta Heritage Foundation for Medical Research (AHFMR 199909) and the Institute of Health Economics (IHE) to David Feeny. PSI, AHFMR, and IHE played no role in the design, interpretation, or analysis of the project and have not reviewed or approved of this manuscript.

The authors gratefully acknowledge the help of the patients and surgeons who participated in the study (London, Ontario, orthopedic surgeons, Drs. Harvey Bailey, David Chess, Wayne Grainger, Paul Kim, and Robert McCalden). The authors also acknowledge the assistance of Dr. Michael Farnworth. Finally, the authors acknowledge the helpful comments provided by the reviewer.

Conclusions: At the mean level for the group, SG and HUI2 scores match closely. At the individual level, agreement is poor. HUI2 scores were greater than HUI3 scores. HUI2 and HUI3 are appropriate for group level analyses relying on community preferences but are not a good substitute for directly measured utility scores at the individual level.

Keywords: Utilities, Standard gamble, Health utilities index, Multi-attribute utility functions, Hip arthroplasty, Preferences, Agreement

Results from economic evaluations and assessments of health-related quality of life (HRQL) play an increasingly important role in clinical and resource allocation decisions in health and health care policy. Prominent Canadian (2) and U.S. (12) guidelines recommend the use of preference-based measures that embody community preferences for health outcomes. There is, however, very little evidence comparing the preference scores directly provided by patients to those based on community preferences. This study compares direct utility scores from patients undergoing total hip arthroplasty (THA) to utility scores based on community preferences.

Preference-based measures are an important family of HRQL measures that provide information on the value of health states (15). The conventional scale assigns a score of 0.00 to dead and 1.00 to perfect health.

There are two major categories of preference-based measures of HRQL: direct and multi-attribute (also referred to as indirect) (23). Direct measures involve asking respondents to value health states. Preference scores are elicited using techniques such as the visual analog scale or using choice-based techniques such as the time tradeoff (TTO) or standard gamble (SG). Respondents may be asked to value hypothetical health states or the respondent's subjectively defined current health state (SDCS). In the latter case, the respondent assesses the current state of their health and provides a valuation for that health state.

In the multi-attribute approach the respondent completes a questionnaire based on a multi-attribute health-status classification system. Prominent multi-attribute systems include the Health Utilities Index (HUI) (6;10), the Quality of Well Being Scale (QWB) (20), and the EuroQol (EQ-5D) (4;21). Health states derived from multi-attribute systems are valued using multi-attribute utility functions estimated using preference measurements from random samples of the general population (community preferences). Thus, in the multi-attribute approach, the respondent provides an assessment of their current health status but does not provide a valuation of it.

Direct preference assessment involves several advantages. Preference scores are obtained from the recipients of the care and reflect both their assessment of current health and their valuation of it. The direct approach, however, is also demanding. In practice, it typically requires trained professional interviewers supported by carefully constructed and tested study-specific interview scripts as well as the use of props (8). The cognitive demands of preference-elicitation interviews exceed the demands of providing self-assessment information on health status; thus, the direct approach is not suitable for all types of respondents.

Seen in this context, obtaining preference scores indirectly by using a multi-attribute system can be quite attractive. How closely would such scores agree with directly measured ones at the group level? How closely would such scores agree at the individual level? In this study, directly measured SG scores will be compared with scores derived from the Health Utilities Index Mark 2 (HUI2) and Mark 3 (HUI3) systems. The scoring functions for HUI2 and HUI3 are both based on preference scores obtained using the standard gamble. Thus, the comparison are not confounded by any difference in method of preference elicitation.

METHODS

Respondents

Respondents were patients enrolled in a study to describe HRQL associated with waiting for and after THA. The study has been described previously and will be only briefly described here (17). Patients were contacted after their primary care physician referred them to an orthopedic surgeon for evaluation of THA. Patients referred to any one of seven surgeons in London, Ontario, were contacted to ascertain their interest in and eligibility for the study. Eligible and consenting patients were invited to attend the outpatients' clinic for a baseline assessment.

Assessments

At baseline (assessment 1, A1), information was collected on sociodemographic variables, employment status, medications, duration and nature of hip problems, and comorbid conditions (17). A battery of health status and HRQL were administered, including the 6-Minute Walk test, Medical Outcomes Study Short Form 36, the Western Ontario McMaster Osteoarthritis Index, the Harris Hip Scale, the State-Trait Anxiety Inventory, and a standardized questionnaire based on HUI2 and HUI3.

HUI2 and HUI3

HUI2 describes health status using seven dimensions of health: sensation (vision, hearing, and speech), mobility, emotion, cognition, self-care, pain, and fertility (5;6;24); the health status classification systems for HUI2 and HUI3 are available at http://www.fhs.mcmaster.ca/hug/index.htm). (Fertility was not assessed in this study.) Each of the six dimensions or attributes of health status assessed for HUI2 in the study has four or five levels ranging from highly impaired to normal. HUI2 health states are described as a seven-element vector, one level for each attribute (with fertility set at level 1, normal). HUI2 states are scored using a multiplicative multi-attribute utility function- based standard gamble scores obtained from members of the general population (community preferences) using the conventional 0.00 = dead and 1.00 = perfect health scale (24). In the HUI2 project the lottery in the SG consisted of perfect health and dead.

HUI3 includes eight attributes: vision, hearing, speech, ambulation, dexterity, emotion, cognition, and pain. In HUI3, there are five or six levels per attribute. HUI3 health states are eight-element vectors, scored using a multiplicative multi-attribute utility functions based on standard gambles scores obtained from a random sample of the general population (community preferences) using the conventional scale (7;9). In the HUI3 project, the lottery in the SG consisted of perfect health and the all-worst HUI3 state. (The HUI3 all-worst health state consists of the lowest level (most disabled) for each of the eight attributes, for instance blind for vision, etc.)

Direct Assessment of Utility Scores

The baseline assessment also involved the direct assessment of preferences for six health states, including three marker states. The marker states were hypothetical health states constructed to correspond to mild, moderate, and severe osteoarthritis (OA) potentially suitable for THA. Each health state description included structured information on pain and stiffness, mobility, use of analgesics, pain and sleep, ability to do housework and chores, and ability to engage in social activities (16). The health-state descriptions for the marker states focused on the domains of health status identified by the orthopedic surgeons as important in the context of OA and THA. The health states were not described using either HUI system.

Each health state was evaluated using a vertical visual analog scale, in this case the Feeling Thermometer (FT) (8). The top of the scale was labeled as"Most Desirable"; the bottom of the scale was labeled as "Least Desirable." In addition to the three marker states, patients evaluated perfect health (lack of OA), dead, and their own SDCS on the FT.

After providing scores on the FT, patients evaluated each of the same health states using the SG. To assist patients in understanding the SG, a Chance Board (CB) was used. Patients were offered a choice between a lottery and a sure thing. The sure thing consisted of an intermediately ranked health state. Each of the marker states and the patient's SDCS were evaluated separately on the CB. The lottery consisted of perfect health with probability p and dead with probability 1-p. The probability p is systematically varied until the patient is indifferent between the lottery and the sure thing. The better the sure-thing health state, the higher the probability of perfect health required to be indifferent between the lottery and the sure thing. The SG scores for each patient's current health are the focus of the analyses reported in the study.

Schedule of Assessments

There were two types of assessments: complete and partial. The complete assessments involved the administration of the health status and HRQL instruments as well as the direct assessment of preferences (marker states and current health state). Partial assessments involved only the administration of the health status and HRQL instruments but did not include the direct assessment of utility scores. Only scores from complete assessments that permit the comparison of direct and multi-attribute scores obtained at the same point in time are reported in the study. Starting at six months after baseline and every six months thereafter, patients were asked to return to the outpatients' clinic for complete assessments. Patients who waited for surgery for more than three months did a complete assessment just before surgery. (Patients not recommended for THA were dropped from the study.) Final and complete assessments were done at least three months after THA. Interviews were conducted by a trained interviewer. Because two criteria (time since baseline and time since surgery) were used to determine the schedule of assessments for each patient, the timing of assessments after baseline do not follow a single pattern. In addition, the number of complete assessments per patient varied, depending on how long the patient waited for surgery.

Statistical Analyses

Paired t-tests are used to compare mean SG and HUI scores at each assessment. SG and HUI scores are interval-scale data. Therefore, agreement between SG and HUI2 scores for the same person at the same point in time is assessed using an intraclass correlation coefficient (ICC) (3;22). The ICC is based on a two-way mixed-effects analysis of variance (ANOVA) model in which the respondent is the random factor and method (direct SG or HUI) is the fixed factor. Similarly, agreement between SG and HUI3 scores is assessed using an ICC. Levels of agreement are interpreted using standards proposed by Guyatt et al. (14): strongly correlated >0.50, moderately correlated 0.35 to 0.50, weakly correlated 0.20 to 0.34, negligible or not correlated 0.00 to 0.19. The extent of agreement is assessed both at the individual level. Analyses were conducted using SPSS Version 11.0.

Ethics Approval

The study was approved by the local Human Ethics Committee of the University of Western Ontario. All patients provided signed informed consent.

RESULTS

Of 215 patients who completed the baseline assessment, 123 were placed on the waiting list for THA; 92 were not (17). Two patients did not have surgery (risk of complications due to coronary artery disease); seven had their THA elsewhere. Of the remaining 114 patients, 15 failed to return for a post-THA assessment (14 refused, 1 died).

The mean time (SD) between assessment 1 and assessment 2 (A2) was 5.2 months (2.0). The mean time between A2 and A3 was 5.5 months (1.8). The mean time between A3 and A4 was 5.0 months (2.3). (Only four patients completed a fifth assessment; data from the fifth assessment are not presented here.) Because the sample sizes for A3 (n = 35) and A4 (n = 12) are small, results from those assessments are reported for descriptive purposes.

As noted in the Methods section, two criteria were used to schedule assessments: time since baseline and time since surgery. For instance, twenty-five patients did A2 while waiting for surgery, thirteen patients did A2 just before surgery, and seventy patients did A2 after surgery. Thus, A2 (or any other assessment other than A1) does not necessarily correspond to a fixed point in the natural history of waiting for THA or recovery after it.

SG scores for the SDCS, HUI2, and HUI3 scores were available (no missing data) for 103 patients at the baseline assessment, 84 patients at A2, 35 patients at A3, and 12 patients at A4. Descriptive information on utility scores is presented in Table 1. Tables 2 and 3 provide the complete list of HUI2 and HUI3 vectors for patients at the baseline assessment.

Results from paired sample t-tests comparing the mean SG and HUI2 (and mean SG and HUI3) scores at each assessment are reported in Table 4. There are no statistically significant differences between mean SG and mean HUI2 scores at any of the assessments. The magnitude of the differences between mean SG and mean HUI2 scores is, in general, small and in some cases trivial. At A2 and A3, the mean SG scores are lower than the mean HUI2 score; at A4, the mean SG score is higher than the mean HUI2 score.

Table 1. Standard Gamble (SG), HUI2, and HUI3 Scores at Assessment 1, Assessment 2, Assessment 3, and Assessment 4

Assessment period	No.	Min	Max	Median	Mean	SD
A1						
SG	103	0.05	0.95	0.75	0.62	0.31
HUI2	103	0.16	0.94	0.65	0.62	0.19
HUI3	103	−0.08	0.97	0.54	0.52	0.21
A2						
SG	84	0.05	0.95	0.75	0.67	0.30
HUI2	84	0.24	0.95	0.73	0.68	0.18
HUI3	84	0.05	0.97	0.59	0.58	0.22
A3						
SG	35	0.05	1.00	0.75	0.65	0.28
HUI2	35	0.11	0.95	0.75	0.70	0.19
HUI3	35	−0.19	0.97	0.64	0.57	0.24
A4						
SG	12	0.15	1.00	0.95	0.75	0.31
HUI2	12	0.29	0.95	0.72	0.70	0.19
HUI3	12	0.32	0.91	0.62	0.61	0.20

Note: Sg means standard gamble utility score for subjectively-defined current health state; HUI = Health Utilities Index Mark; No. = number of patients; SD = standard deviation. The utility score for dead = 0.00; the utility score for perfect health = 1.00.

Table 2. Enumeration of Health Utilities Index Mark 2 Vectors at Baseline

Sensation	Mobility	Emotion	Cognition	Self-care	Pain	Fertility	Frequency	%	Score
2	2	1	1	1	3	1	8	7.8	0.80
2	2	1	1	1	2	1	4	3.9	0.92
2	3	1	1	2	3	1	4	3.9	0.64
2	3	2	1	2	4	1	4	3.9	0.43
2	1	1	1	1	2	1	3	2.9	0.92
2	2	2	2	2	3	1	3	2.9	0.65
2	3	2	1	2	3	1	3	2.9	0.59
2	3	1	1	2	4	1	3	2.9	0.47
2	1	1	1	2	2	1	2	1.9	0.89
2	2	1	1	2	3	1	2	1.9	0.75
2	2	1	1	2	4	1	2	1.9	0.55
2	2	1	2	1	4	1	2	1.9	0.53
2	2	2	1	1	3	1	2	1.9	0.71
2	2	2	1	2	3	1	2	1.9	0.69
2	2	2	2	1	3	1	2	1.9	0.67
2	3	1	1	1	3	1	2	1.9	0.66
2	3	1	1	2	5	1	2	1.9	0.25
2	3	1	2	1	2	1	2	1.9	0.72
1	2	1	1	1	2	1	1	0.1	0.94
1	2	1	1	1	3	1	1	0.1	0.81
1	2	3	2	1	4	1	1	0.1	0.45
1	3	1	1	2	3	1	1	0.1	0.67
1	3	1	1	2	4	1	1	0.1	0.49
1	3	2	2	1	3	1	1	0.1	0.61
2	1	1	1	1	3	1	1	0.1	0.80
2	1	1	2	1	2	1	1	0.1	0.87
2	1	1	2	2	2	1	1	0.1	0.84
2	1	2	1	2	4	1	1	0.1	0.52
2	1	2	2	1	3	1	1	0.1	0.70
2	2	1	1	1	4	1	1	0.1	0.57
2	2	1	1	2	2	1	1	0.1	0.86
2	2	1	2	1	2	1	1	0.1	0.84
2	2	1	2	1	3	1	1	0.1	0.73
2	2	1	2	2	3	1	1	0.1	0.71
2	2	1	2	2	5	1	1	0.1	0.28
2	2	2	1	1	5	1	1	0.1	0.29
2	2	2	1	2	2	1	1	0.1	0.79
2	2	2	1	2	5	1	1	0.1	0.27
2	2	2	1	3	4	1	1	0.1	0.47
2	2	2	2	1	2	1	1	0.1	0.78
2	2	2	2	1	4	1	1	0.1	0.49
2	2	2	2	2	4	1	1	0.1	0.48
2	3	1	1	1	1	1	1	0.1	0.79
2	3	1	1	1	4	1	1	0.1	0.48
2	3	1	1	4	5	1	1	0.1	0.20
2	3	1	2	1	3	1	1	0.1	0.62
2	3	1	2	1	5	1	1	0.1	0.25
2	3	1	2	2	4	1	1	0.1	0.44
2	3	2	1	1	3	1	1	0.1	0.61
2	3	2	1	3	3	1	1	0.1	0.55
2	3	2	2	1	2	1	1	0.1	0.66
2	3	2	2	1	3	1	1	0.1	0.58
2	3	2	2	2	3	1	1	0.1	0.56
2	3	2	2	4	4	1	1	0.1	0.32
2	4	1	1	1	3	1	1	0.1	0.56
3	1	1	1	1	2	1	1	0.1	0.82
3	1	1	1	1	3	1	1	0.1	0.71

https://doi.org/10.1017/S0266462303000321 Published online by Cambridge University Press

Table 2. (*Continued*)

Sensation	Mobility	Emotion	Cognition	Self-care	Pain	Fertility	Frequency	%	Score
3	1	1	2	1	3	1	1	0.1	0.68
3	2	1	1	1	3	1	1	0.1	0.69
3	2	1	1	2	3	1	1	0.1	0.67
3	2	1	2	1	2	1	1	0.1	0.75
3	2	1	2	2	2	1	1	0.1	0.73
3	2	1	2	2	4	1	1	0.1	0.46
3	2	2	2	2	5	1	1	0.1	0.23
3	3	1	1	1	3	1	1	0.1	0.59
3	3	1	2	2	3	1	1	0.1	0.54
3	3	2	2	2	5	1	1	0.1	0.19
3	4	2	2	2	5	1	1	0.1	0.16
4	4	1	1	2	2	1	1	0.1	0.38

Note: N = 103; 69 unique vectors; 50 unique health states.

Table 3. Enumeration of Health Utilities Index Mark 3 Vectors at Baseline

Vision	Hearing	Speech	Ambulation	Dexterity	Emotion	Cognition	Pain	Frequency	%	Score
2	1	1	2	1	1	1	4	9	8.7	0.59
2	1	1	2	1	1	1	3	5	4.9	0.75
2	1	1	2	1	1	3	3	5	4.9	0.70
2	1	1	3	1	1	1	4	4	3.9	0.52
2	1	1	3	1	2	1	4	4	3.9	0.47
2	1	1	3	1	1	1	5	3	2.9	0.26
2	1	1	3	1	1	3	5	3	2.9	0.62
2	1	1	3	2	1	1	4	3	2.9	0.47
2	1	1	2	1	1	1	5	2	1.9	0.32
2	1	1	2	1	1	3	4	2	1.9	0.54
2	1	1	2	1	1	4	4	2	1.9	0.43
2	1	1	2	1	2	1	4	2	1.9	0.54
2	1	1	3	1	1	3	4	2	1.9	0.47
2	1	1	1	1	1	1	3	2	1.9	0.84
1	1	1	2	1	1	1	2	1	0.1	0.85
1	1	1	2	1	1	1	3	1	0.1	0.78
1	1	1	2	1	2	5	4	1	0.1	0.19
1	1	1	3	1	1	1	3	1	0.1	0.69
1	1	1	3	1	1	1	4	1	0.1	0.54
1	1	1	3	1	1	3	3	1	0.1	0.64
1	3	1	2	1	1	1	3	1	0.1	0.65
1	3	1	3	4	1	1	4	1	0.1	0.24
2	1	1	1	1	1	1	1	1	0.1	0.97
2	1	1	1	1	1	1	4	1	0.1	0.66
2	1	1	1	1	1	3	3	1	0.1	0.78
2	1	1	1	1	2	1	2	1	0.1	0.85
2	1	1	1	1	2	1	3	1	0.1	0.70
2	1	1	1	1	2	3	3	1	0.1	0.72
2	1	1	1	2	1	1	3	1	0.1	0.78
2	1	1	2	1	1	1	2	1	0.1	0.83
2	1	1	2	1	2	3	3	1	0.1	0.64
2	1	1	2	2	1	1	3	1	0.1	0.70
2	1	1	2	2	1	1	4	1	0.1	0.54
2	1	1	2	2	1	2	4	1	0.1	0.47
2	1	1	2	2	2	1	4	1	0.1	0.50
2	1	1	2	3	1	1	3	1	0.1	0.62
2	1	1	2	3	1	4	4	1	0.1	0.33
2	1	1	2	3	2	1	4	1	0.1	0.43

https://doi.org/10.1017/S0266462303000321 Published online by Cambridge University Press

Table 3. (*Continued*)

Vision	Hearing	Speech	Ambulation	Dexterity	Emotion	Cognition	Pain	Frequency	%	Score
2	1	1	3	1	1	1	1	1	0.1	0.78
2	1	1	3	1	2	2	3	1	0.1	0.54
2	1	1	3	1	2	4	5	1	0.1	0.13
2	1	1	3	2	3	4	4	1	0.1	0.23
2	1	1	3	3	1	1	4	1	0.1	0.41
2	1	1	4	1	1	1	4	1	0.1	0.38
2	1	1	4	1	1	3	4	1	0.1	0.35
2	1	1	4	1	2	1	4	1	0.1	0.35
2	1	1	4	2	1	1	5	1	0.1	0.14
2	1	1	5	1	1	1	3	1	0.1	0.41
2	1	2	3	1	2	2	4	1	0.1	0.36
2	1	2	5	1	2	3	5	1	0.1	0.04
2	1	3	1	1	1	1	2	1	0.1	0.78
2	1	3	3	1	2	5	5	1	0.1	−0.05
2	2	1	1	1	1	4	2	1	0.1	0.65
2	2	2	2	1	2	3	3	1	0.1	0.59
2	2	1	3	1	1	1	4	1	0.1	0.47
2	3	1	1	1	1	1	3	1	0.1	0.71
2	3	1	2	1	1	3	4	1	0.1	0.44
2	3	1	3	1	1	1	4	1	0.1	0.42
2	3	1	4	1	1	1	5	1	0.1	0.11
2	4	1	2	1	1	3	3	1	0.1	0.48
2	4	1	2	2	1	1	4	1	0.1	0.36
2	4	1	3	1	2	1	4	1	0.1	0.31
2	4	2	2	1	3	3	5	1	0.1	0.05
2	6	1	5	1	1	1	5	1	0.1	−0.08
3	1	1	1	3	1	1	3	1	0.1	0.60
3	1	1	2	1	1	1	4	1	0.1	0.50
3	1	1	2	1	1	5	3	1	0.1	0.24
3	4	1	1	1	1	3	3	1	0.1	0.46
4	4	1	2	1	1	3	2	1	0.1	0.41

Note: N = 103; 69 unique vectors; 42 unique health states.

Table 4. Paired t-Tests and Intraclass Correlations (ICC) Comparing Standard Gamble and HUI2 and HUI3 Scores

Comparisons at assessment	No.	t-Test (g)	ICC (i)	95% CI (i)
A1				
SG-HUI2	103	0.02	0.09	−0.10 to 0.28
SG-HUI3	103	2.94*	0.09	−0.10 to 0.27
A2				
SG-HUI2	84	−0.40	0.06	−0.15 to 0.27
SG-HUI3	84	2.48*	0.09	−0.13 to 0.29
A3				
SG-HUI2	35	−0.85	0.22	−0.11 to 0.52
SG-HUI3	35	1.34	0.17	−0.17 to 0.47
A4				
SG-HUI2	12	1.05	0.77	0.38 to 0.93
SG-HUI3	12	1.94	0.55	−0.01 to 0.85

Note: SG = standard gamble utility score; HUI, Health Utilities Index Mark; g = group; i = individual; CI = confidence interval; No. = number; A1 = assessment 1; A2 = assessment 2; A3 = assessment 3; A4 = assessment 4.
* $p < .05$.

https://doi.org/10.1017/S0266462303000321 Published online by Cambridge University Press

Table 5. Mean Differences (Postsurgery minus Presurgery) in Utility Scores

Measure	Mean difference	Standard deviation
SG	0.17	0.32
HUI2	0.18	0.17
HUI3	0.25	0.20

Note: SG = standard gamble; HUI = Health Utilities Index Mark.

There are statistically significant differences between mean SG and mean HUI3 scores at A1 and at A2. The magnitude of the differences in mean SG and mean HUI3 scores is, in general, large. Furthermore, at all four assessments, mean SG scores are higher than mean HUI3 scores.

It is also worth noting that the differences between group mean HUI2 and HUI3 scores are substantial and systematic. Mean HUI2 scores exceed mean HUI3 scores.

Differences between mean scores pre- and postsurgery are also of interest. The mean differences (postsurgery minus presurgery) in utility scores are displayed in Table 5. The SG and HUI2 provide virtually identical information about the effects of THA at the group level. HUI3 provides similar information on the magnitude of the effects of THA.

Individual-level ICCs (and 95% confidence intervals) comparing SG and HUI2 (and SG and HUI3) scores at each assessment are presented in Table 4. Agreement at A1and A2 is low, in most cases negligible. Agreement is higher at A3 but still at best weak. Agreement is strong at A4 but based on results for only 12 patients.

DISCUSSION

At the group mean level, HUI2 scores appear to match direct SG scores accurately. In general, the differences between mean SG and mean HUI2 scores are not large and the direction of the difference is not systematic. The estimated estimates of the effects of THA on HRQL are virtually identical.

In comparing group mean SG and HUI3 scores, the differences are larger and HUI3 are systematically lower than SG scores. A possible reason for the difference is that the multiplicative multi-attribute utility function for HUI3 ranges from −0.36 (most disabled state in HUI3, all attributes at their lowest level) to 0.00 (dead) to 1.00 (perfect health). In the HUI3 system, there are many states with scores worse than dead. Indeed at A1 and at A3, the minimum HUI3 scores observed were negative. However, there were only two individuals with negative scores at A1 (−0.05, −0.08); if both scores were set to zero, the mean HUI3 score at A1 would still be 0.52. Similarly, only one patient had a negative HUI3 score at A3 (−0.19); the mean at A3 would be 0.58 instead of 0.57 if that score were set to zero.

A second potential source of the systematic difference between SG and HUI3 scores is that the HUI3 scoring function was based on SG questions involving a lottery between perfect health and the all-worst HUI3 health state, but the lottery in the SG done by patients involved perfect health and dead. Nonetheless, the SG and HUI3 provide similar estimates of the effects of THA.

That HUI2 scores may be systematically lower than HUI3 scores, especially when the burden of morbidity is substantial, has been noted previously (18). One of the sources of this difference is that there are very few states in the HUI2 system with scores worse than dead. The HUI2 scale runs from −0.03 (all-worst HUI2 state) to 0.00 (dead) to 1.00 (perfect health). The extent to which states were considered worse than dead was not quantified in

detail in the HUI2 scoring function project (24) but was in the HUI3 project (7;9). Another source of difference is the greater descriptive power of the HUI3 system. In this study, HUI2 covered six attributes with four or five levels per attribute. In contrast, HUI3 covered eight attributes with five or six levels per attribute. Although HUI3 scores are systematically lower than SG and HUI2 scores, the differences in mean scores between baseline and A2 are similar for all three sets of scores. Similarly, the differences in mean scores pre- and postsurgery are similar.

Even though the group mean HUI2 scores match the group mean SG scores reasonably well, agreement (ICCs) is, in general, low. Much the same pattern is evident in assessing agreement between SG and HUI3 scores. This finding suggests that, at the individual level, HUI2 and HUI3 scores are poor substitutes for the SG score. Although SG and HUI scores are all utility scores, there are important conceptual differences between them. With the SG for the SDCS, each patient reflects on their current health and the value that they attach to that health state. With the HUI, each patient provides information on how they view their current health using the definitions of health status of each of the HUI systems. Furthermore, the resulting HUI2 and HUI3 health states are valued using community preferences embodied in the multi-attribute utility scores functions. The community preferences captured in the HUI scoring functions represent mean utilities that suppress the considerable interperson heterogeneity in preferences. In contrast, the SG results retain that heterogeneity. (Gorbatenko-Roth et al. (13) discuss the importance of individual differences in preferences.) That there would be a close match at the group mean level but little agreement at the individual level is, therefore, not remarkable.

Agreement between scores from multi-attribute systems and directly measured preference scores at the group mean level but a lack of agreement at the individual level has been noted in several studies. Results similar to those reported here were reported by Gabriel et al. (11), Nichol et al. (19), and Albertsen, Nease, and Potosky (1). Given the widely observed heterogeneity in preference scores for health states, it is not surprising that community preferences and the preference of each individual patient match poorly.

CONCLUSIONS AND POLICY IMPLICATIONS

In selecting a utility measure for use in a study with the intention of conducting group-level analyses, HUI2 is a good substitute for the directly measured SG scores. Furthermore, the burden on the investigator and the patients associated with the use of HUI2 and HUI3 is considerably lower than the burden imposed by the direct assessment of utility scores. For studies relying on group-level analyses, the use of systems such as HUI also conforms with recommendations in the national guidelines for economic and HRQL assessments in Canada (2) and in the United States (12). In assessing the effects of THA in this study, the SG and HUI2 provide virtually identical estimates, whereas the HUI3 results are similar. In this case, similar answers are obtained at the group level using patient or community preferences.

For studies for which it is important to have individual-specific preference information, however, multi-attribute systems are not a good substitute for directly measured preference scores. More research on agreement between directly measured and multi-attribute scores is needed.

REFERENCES

1. Albertsen PC, Nease RF, Potosky AL. Assessment of patient preferences among men with prostate cancer. *J Urol*. 1998;159:158-63.

2. Canadian Coordinating Office for Health Technology Assessment. *Guidelines for economic evaluation of pharmaceuticals: Canada*. 2nd ed. Ottawa: Canadian Coordinating Office for Health Technology Assessment; November, 1997.
3. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures: Statistics and strategies for evaluation. *Control Clin Trials*. 1991;12:142S-158S.
4. Essink-Bot ML, Stouthard MEA, Bonsel GJ. Generalizability of valuations on health states collected with the EuroQol questionnaire. *Health Econ*. 1993;2:237-46.
5. Feeny D, Furlong W, Barr RD, et al. A comprehensive multiattribute system for classifying the health status of survivors of childhood cancer. *J Clin Oncol*. 1992;10:923-8.
6. Feeny DH, Torrance GW, Furlong WJ. Health Utilities Index. In: Bert Spilker, ed. *Quality of life and pharmacoeconomics in clinical trials*. 2nd ed. Philadelphia: Lippincott-Raven Press; 1996:239-52.
7. Feeny D, Furlong W, Torrance GW, et al. Multi-attribute and single-attribute utility functions for the health utilities index Mark 3 system. *Med Care*. 2002;40:113-28.
8. Furlong W, Feeny D, Torrance GW, Barr R, Horsman J. Guide to design and development of health-state utility instrumentation. Ontario, Canada: McMaster University Centre for Health Economics and Policy Analysis Working Paper No 90-9; June 1990.
9. Furlong W, Feeny D, Torrance GW, et al. Multiplicative multi-attribute utility function for the health utilities index Mark 3 (HUI3) system: A technical report. Ontario, Canada: McMaster University Centre for Health Economics and Policy Analysis Working Paper No. 98-11.
10. Furlong WJ, Feeny DH, Torrance GW, Barr RD. The Health Utilities Index (HUI) system for assessing health-related quality of life in clinical studies. *Ann Med*. 2001;33:375-84.
11. Gabriel SE, Kneeland TS, Melton LJ, et al. Health-related quality of life in economic evaluations for osteoporosis: Whose values should we use? *Med Decis Making*. 1999;19:141-8.
12. Gold MR, Siegel JE, Russell LB, Weinstein MC, eds. *Cost-effectiveness in health and medicine*. New York: Oxford University Press; 1996.
13. Gorbatenko-Roth KG, Levin IP, Altmaier EM, Doebbeling BN. Accuracy of health-related quality of life assessment: What is the benefit of incorporating patients' preferences for domain functioning? *Health Psychol*. 2001;20:136-40.
14. Guyatt GH, Berman LB, Townsend M, Pugsley SO, Chambers LW. A measure of quality of life in clinical trials in chronic lung disease. *Thorax*. 1987;42:773-8.
15. Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. *Ann Intern Med*. 1993;118:622-9.
16. Laupacis A, Bourne R, Rorabeck C, et al., Effect of elective total hip replacement upon health-related quality of life. *J Bone Joint Surg Am*. 1993;75:1619-26.
17. Mahon J, Bourne R, Rorabeck C, et al. The effect of waiting for elective total hip arthroplasty on health-related quality of life. *Can Med Assoc J*. 2002;167:1115-21.
18. Neumann PJ, Sandberg EA, Araki SS, et al. A comparison of HUI2 and HUI3 utility scores in Alzheimer's disease. *Med Decis Making*. 2000;20:413-22.
19. Nichol G, Llewellyn-Thomas HA, Thiel EC, Naylor CD. The relationship between cardiac functional capacity and patients' symptom-specific utilities for angina. *Med Decis Making*. 1996;16:78-85.
20. Patrick DL, Bush J, Chen M. Methods for measuring levels of well-being for a health status index. *Health Serv Res*. 1973;8:228-45.
21. Rabin R, de Charro F. EQ-5D: A measure of health status from the EuroQol group. *Ann Med*. 2001;33:337-43.
22. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. *Psychol Bull*. 1979;86:420-8.
23. Torrance GW. Measurement of health state utilities for economic appraisal: A review. *J Health Econ*. 1986;5:1-30.
24. Torrance GW, Feeny DH, Furlong WJ, Barr RD, Zhang Y, Wang Q. Multi-attribute preference functions for a comprehensive health status classification system: Health utilities index Mark 2. *Med Care*. 1996;34:702-22.