The Psychometric Properties of Patient Preferences in Osteoporosis

ANN CRANNEY, DOUGLAS COYLE, BA PHAM, JACQUELINE TETROE, GEORGE WELLS, ELAINE JOLLY, and PETER TUGWELL

ABSTRACT. ***Objective.*** Osteoporosis is a chronic disease manifested by wrist, vertebral, and hip fractures that results in significant morbidity and burden to society. About 30% of postmenopausal women have osteoporosis according to the WHO criteria. Women with one vertebral fracture have a 4-fold increased risk of a subsequent fracture. The goal of treatment is to prevent fractures and improve quality of life. Preferences or utilities are now recommended for incorporating quality of life into evaluations of the cost effectiveness of new therapeutic interventions. We evaluated the psychometric properties of preference based measures in osteoporosis.

Methods. Preference scenarios were constructed with a health state classification system. The reliability and validity of the feeling thermometer and the standard gamble was assessed by interviewing 42 women from 4 different patient groups. The sensitivity to change of the feeling thermometer and standard gamble was compared with the Health Utilities Index Mark 2 (HUI2) and SF-36. All subgroups were reassessed about 2 months after their first interview.

Results. Preference measurement was feasible in women of different age groups. The reliability coefficients for health states ranged from 0.65 to 0.87. The preference scores for the marker states demonstrated content validity. Convergent validity of the feeling thermometer was supported by a significant correlation with the HUI2 ($r = 0.38$, $p < 0.05$) and the physical health summary of the SF-36 ($r = 0.56$, $p < 0.005$). The standard gamble did not correlate with the HUI2 ($r = 0.15$) or the feeling thermometer ($r = 0.09$), but correlated with 2 domains of the SF-36. The preference measures were sensitive to change, with the efficiency scores ranging from 0.78 to 1.0.

Conclusion. Preference measurements in the evaluation of osteoporosis are feasible. The feeling thermometer and standard gamble appear to be related to different aspects of health related quality of life. Both instruments were sensitive to change over a 2 month period. (J Rheumatol 2001;28:132–7)

Key Indexing Terms:
PREFERENCES OSTEOPOROSIS QUALITY OF LIFE FRACTURES

Osteoporosis is a debilitating disease that results in a significant burden to society[1,2]. Fifteen out of 100 women will fracture their hip during their lifetime[3]. It is estimated that 10% of these women die within the first 6 months of their hip fracture and only one-third will return to their prefracture level of functioning. Postmenopausal women are particularly susceptible to the complications of osteoporosis, which is characterized by the presence of fractures of the wrist, spine, and hip.

There are currently many treatments available to prevent and treat osteoporosis, including hormonal replacement therapy (HRT), bisphosphonates, raloxifene, calcitonin, and vitamin D derivatives. Unfortunately, women may be required to take medications for many years to prevent an osteoporotic related fracture. Current resources for health care treatments are limited and not all osteoporosis treatments that demonstrate efficacy can be supported. It is important that we identify cost effective therapies that can prevent osteoporosis. Analysis of cost effectiveness should not be restricted only to clinical outcomes but also should incorporate health related quality of life (QOL) in the form of preference measures (utilities) rather than disease-specific instruments.

Preference measures assign values to particular health states to reflect the level of satisfaction that a person associates with a particular health state[4]. Preference measures such as the feeling thermometer can provide a summary score (ranging from 0 to 1) of health related quality of life (HRQOL). This score can then be used as a quality adjustment factor in determining quality adjusted life years (QALY) in cost-utility analyses[5]. However, preference mea-

From the Department of Medicine and Division of Gynecology and Reproductive Endocrinology, Ottawa Hospital; Clinical Epidemiology Unit, Loeb Health Research Institute, University of Ottawa; and Thomas C. Chalmers Centre for Systematic Reviews, Children's Hospital of Eastern Ontario Research Institute, Ottawa, Ontario, Canada.

A. Cranney was a recipient of an Arthritis Society Fellowship when this work was conducted.

A. Cranney, MD, MSc, Department of Medicine, Ottawa Hospital; D. Coyle, MSc, Clinical Epidemiology Unit, Loeb Health Research Institute, University of Ottawa; B. Pham, MSc, M Math, Thomas C. Chalmers Centre for Systematic Reviews, Children's Hospital of Eastern Ontario Research Institute; J. Tetroe, BA, MA, Clinical Epidemiology Unit; G. Wells, PhD, Clinical Epidemiology Unit, University of Ottawa; E. Jolly, MD, Division of Gynecology and Reproductive Endocrinology; P. Tugwell, MD, MSc, Department of Medicine, Ottawa Hospital.

Address reprint requests to Dr. A. Cranney, 737 Parkdale, Room 460, Ottawa Hospital, Civic Campus, Ottawa, Ontario K1Y 1J8, Canada.

Submitted February 28, 2000 revision accepted July 17, 2000.

sures may be less responsive to change than disease-specific instruments.

Researchers have elicited preferences in women taking HRT, both with and without osteoporosis[6,7], but no published studies could be found on preferences in patients using other therapeutic regimens. Fordham elicited QOL from 132 hip fracture patients at one and 2 years using the EuroQol, with a mean score of 0.61 and 0.60, respectively, suggesting that QOL did not decrease one year after fracture[8]. Gabriel, *et al* elicited preferences in patients one to 5 years after vertebral or hip fracture using an automated time trade-off (U-titer)[9]. The mean current health score for vertebral fracture subjects was 0.81 (n = 94) and 0.70 for hip fracture subjects (n = 37).

The feasibility, reliability, and validity of preference measurement techniques in osteoporosis are still uncertain. We investigated if it was feasible and valid to elicit preferences in different groups of women with osteoporosis. We set out to answer a number of questions: (1) Would preferences be reliable and sensitive to change in women who sustain a fracture? (2) Do the different measures used to elicit preferences correlate with one another? (3) Would HRQOL differ according to the type of fracture? (4) Finally, does inclusion of side effects in the preference scenario have a negative effect on the value given to that health state?

MATERIALS AND METHODS

Study sample. A cohort of 42 women was followed over a 2 month period. There were 4 subgroups: Group 1: women who had not experienced a fracture but who were osteopenic and undergoing HRT for osteoporosis prevention; and Groups 2–4: women with a recent wrist, spine, or hip fracture. For the purpose of this validation study, 10–15 patients per subgroup were interviewed. An exact sample size calculation was difficult due to insufficient data on preferences in patients with osteoporosis.

Women in Group 1 had recently commenced HRT and were identified through the menopause and rheumatology clinics.

Women over age 50 with osteoporotic related fractures of the hip, vertebrae, or wrist were identified after their fracture and interviewed at baseline, and on a second occasion 2 months later. Subjects were identified through the fracture clinics and the orthopedic and geriatric wards. All fractures were confirmed by radiograph.

Patients were excluded if they were unable/unwilling to give consent, not fluent in English, severely depressed, deaf or visually impaired, or had a pathological fracture.

Demographic information was recorded for each patient at baseline, including name, level of education, type of fracture, use of HRT. Comorbid conditions such as heart disease, respiratory conditions, and malignancy were elicited from the patient although a formal questionnaire was not used.

Health status and preference measures. Study interviews were performed at baseline and there were 2 month followup interviews with each patient. Re-interviews were conducted in a sample of patients from each of the 4 subgroups at 2 weeks to evaluate test–retest reliability.

Preferences were elicited using the feeling thermometer (current health and 5 different marker states), the standard gamble (current health), and the Health Utilities Index Mark 2 (HUI2). The feeling thermometer (also known as rating scale) is derived from psychological scaling[10] and consists of a horizontal or vertical scale anchored by defined endpoints: death (zero = worst imaginable state) at one end and perfect health (100 = best imaginable state) at the other. Patients were requested to rate the desirability of different health (marker) states and their own health state along an interval scale between the anchors. The anchors used were perfect health at the top and death at the bottom. After the feeling portion of the interview was completed, the standard gamble was administered using a probability wheel as a visual aid to facilitate understanding. Each patient was offered a choice between remaining in their current health and a gamble with chance p to obtain perfect health and a chance 1 – p of death. Chance p was systematically varied using the ping-pong approach.

The standard gamble requires patients to choose between living in a less than optimal chronic health state or taking a gamble with a treatment with a certain probability (p) of perfect health, but also risking a corresponding probability (1 – p) of immediate death. The more severe the health state evaluated, the higher the risk of death the individual will be willing to take to avoid the current state.

The HUI2 is an indirect method of preference elicitation. Two steps are involved. First, a person's health status is elicited along several dimensions using a questionnaire. Second, a preference for that particular health state is derived, based on values obtained from previous populations[11].

Health status was measured at baseline and at 2 months using the SF-36 generic measure[12,13]. The Medical Outcomes Survey SF-36 is a 36 item questionnaire divided into 8 subscales detailing HRQOL: Physical function, Role function as limited by physical problems, Bodily pain, General health perceptions, Vitality, Social function, Role function as limited by emotional problems, and Mental health. Roughly 80–85% of the reliable variance in the 8 scales is explained by 2 domains, which results in the construction of 2 summary measures — the Physical Component summary and the Mental Health Component summary. The SF-36 is reliable and internally consistent; normative data is available[12].

A script for conducting the preference assessments was developed to standardize the interview and make the process clearer. Marker health states were constructed using a health state classification system format that requires the investigator to identify the attributes that define the disease problem, resulting in a comprehensive description of the levels of functioning. The description of each marker state covered 6 dimensions: activities of daily living, self-care functions, anxiety and depression, leisure activities, pain, and side effects from treatment. To assess the relative importance of side effects on the overall preference score, identical scenarios were developed for patients with and without side effects from their medication. The side effects described with hormonal replacement were bleeding, breast tenderness, mood swings, weight gain, headache, and nausea. After the scenarios were developed, 2 experts in the area of health economics/preference assessment and 4 experts (physicians, nurses) involved in the treatment of osteoporosis/menopause were requested to examine the script for content validity. At each assessment, the feeling thermometer was used to derive preference values for the 5 marker states. Patients were then requested to describe their own current health state according to the same health state classification system.

Analyses: Test-retest reliability. Reliability coefficients were calculated to evaluate short term test-retest reliability for the marker states, and current health values elicited by the feeling thermometer and standard gamble. Test-retest reliability was assessed by the patient's valuation of the different marker states — their evaluation of their current health and the standard gamble at 2 weeks compared to the baseline assessment, using an intraclass correlation (ICC)[14,15]. The reliability coefficient was calculated based on the difference between 2 assessments using one way analysis of variance (ANOVA), with the subject as an independent factor and the difference between scores as the dependent factor. The accepted standard for the reliability coefficient is a value greater than 0.70[16].

Sensitivity to change. Sensitivity to change relates to whether the estimation technique is able to detect the smallest clinically important improvement. An efficiency score (over the 2 months) was calculated for each measurement technique using $E = d/SD_d$, where d is the mean of the change scores for the group and SD_d is the standard deviation of the change measure[17]. Each individual was asked if they had improved subjectively, which was used as a criterion of clinical change, and the efficiency measure was calculated for individuals who felt that they had improved.

Construct validity. Each of the preference measurement techniques was

assessed in terms of construct validity, defined as the ability to produce results that are consistent with theoretically derived hypotheses concerning the constructs that are being measured. Construct validity can be assessed by 2 measures. Convergent validity refers to the extent to which a selected measurement technique agrees with other accepted measurement techniques, and divergent validity refers to the fact that the instrument should not correlate with unrelated variables or constructs.

In this study, construct validity was assessed by verification of the following hypotheses:

(1) Preferences derived from the hip and vertebral fracture subjects (using the feeling thermometer and standard gamble) would be lower than scores for the wrist fracture and HRT subjects.

(2) Scores from the feeling thermometer and standard gamble should correlate with certain domains of the SF-36 (such as the physical function domain). These correlations were assessed using Spearman's correlation coefficient.

RESULTS

Patient characteristics. All interviews were completed except in 2 individuals who were lost to followup. The time range to complete the full interview was 45–60 minutes.

The degree of comorbidity was higher in the vertebral (45%) and hip fracture (63%) groups, compared to both the wrist fracture and HRT groups (9%) (Table 1).

Baseline SF-36 scores for the 4 different groups are shown in Table 1. Overall, there was a trend for the scores to be lower in the hip fracture group, except on the domains of general health, emotional role, and mental health, where the hip fracture group rated themselves higher than the HRT group.

Construct validity. The results of preferences for current health agree with the a priori axioms that hip and vertebral fracture preferences should be lower than those of wrist or non-fracture patients, at least for the feeling thermometer method.

Tables 2 and 3 present the baseline and followup preferences for the patient's current health using the feeling thermometer. The results reveal a gradient of severity according to the group assessed, with highest preferences for the HRT group and lowest for the hip fracture subjects. On the standard gamble, women with a hip fracture and those undergoing HRT rated the highest health (0.91) compared to 0.88 and 0.86 for

Table 2. Baseline scores for current health according to group: mean (standard deviation) and range.

Group	Feeling Thermometer	Standard Gamble	Health Utilities Index 2
HRT, n = 11	0.92 (0.08)	0.90 (0.11)	0.80 (0.10)
	0.78, 1.0	0.70, 1.0	0.62, 0.92
Wrist, n = 11	0.84 (0.11)	0.87 (0.19)	0.86 (0.06)
	0.63, 0.95	0.50, 1.0	0.75, 0.92
Vertebral fracture,	0.76 (0.13)	0.84 (0.20)	0.79 (0.22)
n = 10	0.50, 0.95	0.50, 1.0	0.25, 0.92
Hip, n = 10	0.71 (0.11)	0.91 (0.12)	0.67 (0.12)
	0.50, 0.85	0.75, 1.0	0.53, 0.89

HRT: hormone replacement therapy.

Table 3. Followup preference scores for current health: mean (standard deviation) and range.

Group	Feeling Thermometer	Standard Gamble	Health Utilities Index Mark 2
HRT	0.88 (0.12)	0.93 (0.07)	0.82 (0.07)
	0.6–1.0	0.8–1.0	0.69–0.92
Wrist fracture	0.88 (0.07)	0.91 (0.15)	0.87 (0.07)
	0.75–0.95	0.5–1.0	0.70–0.95
Vertebral fracture	0.83 (0.08)	0.91 (0.10)	0.76 (0.14)
	0.73–0.97	0.70–1.0	0.43–0.92
Hip fracture	0.76 (0.18)	0.84 (0.18)	0.71 (0.09)
	0.40–0.95	0.50–1.0	0.58–0.82

Table 1. Baseline patient characteristics.

	Patient Group			
	Hormone Replacement	Wrist Fracture	Vertebral Fracture	Hip Fracture
Number approached	15	20	20	40
Number eligible	13	19	15	15
Sample size (number consenting)	11	11	10	10
Age: median (range), yrs	56.0 (45–69)	68.0 (51–80)	75.5 (65–88)	79.5 (63–91)
Education > Grade 13 level, %	45	45	40	50
Comorbidity, %	9	9	45	63
Mean SF-36 score				
Physical functioning	85	76	51	42
Role - physical	70	30	25	22
Body pain	63	58	44	50
General health	72	79	72	81
Vitality	60	60	58	55
Social functioning	82	73	67	69
Role emotional	79	51	86	83
Mental health	76	82	79	85
Physical health summary	47	43	23	32
Mental health summary	50	51	43	58

wrist and vertebral fracture groups, respectively. The HUI2 preferences for current health were similar to preferences derived from the feeling thermometer, except the HRT group had lower preferences when elicited by the HUI2. The mean standard gamble preferences were higher than both the HUI2 and feeling thermometer preferences throughout the wrist, vertebral, and hip fracture groups. Lower HUI2 scores for women with hip and vertebral fractures were a result of lower physical functioning and more problems with self-care (HUI2) and mobility. In contrast, the standard gamble preferences were higher in the hip fracture group (mean 0.91) and were similar to the preference measurements for the HRT group.

In Table 4, preferences for marker states are presented for baseline and followup. The preferences for the marker states decreased in value according to the severity of the marker condition. The mean values for the different marker states were as follows: mild (wrist fracture) ranged from 0.86 to 0.91, moderate (vertebral fracture) from 0.50 to 0.61, and 0.30 to 0.37 for the severe marker state (hip fracture), which confirms consistency between the 4 subgroups. There was also less than 10% difference between the marker states at baseline and followup, confirming reliability.

Both the feeling thermometer and the HUI2 correlated with the following domains of the SF-36: Physical functioning, Social functioning, Role physical, Body pain, and the Physical health summary (Table 5). The standard gamble, however, correlated more strongly with the general health dimension of the SF-36 ($r = 0.59$, $p = 0.0001$) as well as with Role emotional, Role social, and Mental health summary. The feeling thermometer correlated significantly with the HUI2 values.

Table 4. Baseline and followup preference scores for marker states elicited with the feeling thermometer*.

Group	Marker States, Mean Preference Score (range)				
	Wrist Fracture without Side Effects	Wrist Fracture with Side Effects	Vertebral Fracture	Hip Fracture without Side Effects	Hip Fracture with Side Effects
HRT					
Baseline	0.88 (0.77–1.00)	0.76 (0.60–0.95)	0.61 (0.35–0.8)	0.47 (0.25–0.80)	0.37 (0.20–0.65)
Followup	0.88 (0.75–1.00)	0.73 (0.60–0.95)	0.57 (0.27–0.85)	0.45 (0.20–0.70)	0.37 (0.10–0.70)
Wrist					
Baseline	0.86 (0.70–0.95)	0.75 (0.55–0.90)	0.50 (0.25–0.65)	0.39 (0.15–0.55)	0.30 (0.10–0.50)
Followup	0.90 (0.80–0.95)	0.78 (0.60–0.85)	0.54 (0.27–0.70)	0.42 (0.20–0.60)	0.36 (0.20–0.55)
Vertebrae					
Baseline	0.91 (0.75–0.99)	0.80 (0.70–0.90)	0.56 (0.40–0.85)	0.42 (0.25–0.65)	0.37 (0.20–0.55)
Followup	0.92 (0.75–0.98)	0.84 (0.70–0.90)	0.61 (0.50–0.75)	0.46 (0.35–0.60)	0.40 (0.25–0.55)
Hip					
Baseline	0.86 (0.65–0.95)	0.78 (0.55–0.90)	0.50 (0.25–0.70)	0.38 (0.20–0.55)	0.33 (0.27–0.50)
Followup	0.86 (0.65–0.98)	0.78 (0.40–0.87)	0.54 (0.40–0.70)	0.41 (0.30–0.55)	0.32 (0.20–0.50)

*Scale anchors: 1 = perfect health, 0 = death.

Table 5. Correlation coefficients between baseline preferences and health status outcomes (n = 42).

	Feeling Thermometer, Current Health	Standard Gamble, Current Health	HUI2	HAQ, Pain
Feeling thermometer	—	0.09	0.38*	–0.36*
Standard gamble	0.09	—	0.15	–0.20
HUI	0.38*	0.15	—	–0.33*
SF-36 dimensions				
Physical functioning	0.56***	0.15	0.55***	–0.39
Role–physical	0.44***	0.22	0.39**	–0.14
Body pain	0.32**	0.25	0.43***	–0.34*
General health	0.14	0.59***	0.05	–0.40**
Vitality	0.14	0.29	0.36*	–0.18
Social functioning	0.52***	0.31*	0.35*	–0.29
Role–emotional	0.02	0.29	0.08	–0.11
Mental health	0.03	0.26	0.03	–0.01
Physical health summary	0.60***	0.22	0.62***	–0.36*
Mental health summary	–0.07	0.36*	–0.09	0.004

*p < 0.05, **p < 0.01, ***p < 0.005. HUI2: Health Utilities Index Mark 2, HAQ: Health Assessment Questionnaire.

Table 6. Reliability coefficients. Test-retest reliability: between test and retest scores 2 weeks apart (n = 10).

Marker State, Feeling Thermometer	Reliability Coefficient
Wrist fracture/side effects	0.79
Wrist fracture/no side effects	0.72
Vertebral fracture	0.65
Hip fracture	0.85
Hip fracture/no side effects	0.87
Current health	
Feeling thermometer	0.83
Standard gamble	0.83

Test-retest reliability. For current health, the reliability coefficient was 0.83 for both the feeling thermometer and the standard gamble. The reliability coefficients for the different marker states exceeded the a priori criteria for acceptance of 0.70, except for the vertebral fracture state (Table 6).

Sensitivity to change. The larger the value, the greater the sensitivity to change of an instrument. The efficiency score for the feeling thermometer, standard gamble, and HUI2 were 1.0, 1.0 and 0.8, respectively.

DISCUSSION

This study was designed to assess psychometric properties of patient preferences for current health in subjects initiating hormone replacement and patients who recently experienced an osteoporotic related fracture.

The methods we used to develop the marker states for the feeling thermometer paid special attention to content validity of the descriptions of the marker states. A comprehensive system of levels of functioning was used in the description of the marker states, along with a description of side effects. Some of the studies on preferences have neglected to involve physicians, nurses, and patients in the development of health state descriptions. After development of the osteoporotic health states, we evaluated reliability, construct validity, and sensitivity to change of the preferences elicited via the standard gamble and the feeling thermometer.

The patients' education levels were similar in all 4 groups, which is relevant, since earlier studies have shown that education level may affect an individual's preference values[10].

The test-retest reliability of the values using the feeling thermometer and standard gamble in patients with osteoporotic related fractures was acceptable[14,18] and comparable to other studies, with 5 of 7 values exceeding 0.75[19].

The values elicited using all 3 instruments revealed a gradient according to group, with the highest values in the HRT group, which confirms content validity. The standard gamble values were higher than the feeling thermometer values in the fracture subgroups[9].

The 4 subgroups of women provided similar scores in their rating of the marker states (Table 4). One of the limitations of this study was the small numbers in each subgroup — it was not possibe to test for differences between subgroups.

The feeling thermometer and standard gamble proved sensitive to change, although to assess the sensitivity to change adequately, a longer followup period after fracture would be necessary, since 2 months may be an inadequate period to detect change in some patients. One of the limitations of the standard gamble is the ceiling effect that is noted on the baseline values (Table 2).

The values we obtained for the vertebral and hip fracture subgroups using the feeling thermometer and the HUI2 were comparable to those obtained by Gabriel, *et al* using the time trade-off, HUI2, and rating scale[9]. Gabriel, *et al* obtained a mean value of 0.76 for a vertebral fracture group using a rating scale, which is identical to our value elicited using the feeling thermometer. For the hip fracture group, Gabriel, *et al* obtained a mean value of 0.72 compared to our value of 0.71 elicited with the feeling thermometer. Similarly, the values we obtained using the HUI2 were almost identical with Gabriel's results.

Construct validity was confirmed, as results supported the a priori hypotheses, indicating a higher preference score for the HRT and wrist fracture subjects than for the vertebral and hip fracture subjects. The only exception to this was the higher scores on the standard gamble for the fracture subjects. The scenarios for the marker states that contained side effects resulted in lower values than the scenarios without side effects. Convergent validity of the preferences obtained by the feeling thermometer was supported by correlations with 4 domains and the Physical health summary of the SF-36. The feeling thermometer also correlated with the HUI2, which was expected given that the feeling thermometer script and the 6 attributes used are derived from the HUI2. The standard gamble values did not correlate with the HUI2, and correlated with only 2 completely different domains — general health and social functioning — of the SF-36. The responses for the different marker states were distributed over the feeling thermometer, a phenomenon that could be attributed to "response-spreading"[20] rather than reflecting values for the true severity of the different conditions. However, we attempted to avoid response-spreading by following recommendations that the response continuum is made clear and the endpoints of the scale are well defined[21].

On a purely theoretical basis, the standard gamble has some attributes that are not present in the other instruments. Researchers argue that the standard gamble is the only correct method since it is derived from von Neuman-Morgenstern utility theory and has a strong axiomatic base[22]. The standard gamble has the "element of risk" attitude that is not part of the feeling thermometer. This may explain why the results of the standard gamble did not correlate with the feeling thermometer values. The standard gamble values reflect different aspects of health status than the feeling thermometer, perhaps because the 2 instruments are derived from different scientific paradigms.

Lalonde, *et al* evaluated the psychometric properties of preferences in subjects with coronary heart disease and also found that the feeling thermometer and standard gamble seemed to assess different aspects of quality of life[23].

The results of this study have important implications for the use of different methods to elicit preferences in subjects with osteoporosis. Both methods are reliable based on test-retest results. Some of the older women had more difficulty comprehending the standard gamble and this may limit the feasibility of these measurements. The standard gamble, however, would be useful in decision analyses with treatments that involved a risk of death, where it would be important to incorporate an individual's attitudes toward risk into the preference measurement.

In conclusion, the validity of the current health scores of the feeling thermometer and the standard gamble were evaluated in comparison to scores derived from the SF-36 and HUI2. We found that it was feasible to use the feeling thermometer and the standard gamble to determine preference values and document QOL in patients with osteoporotic related fractures.

The differences between the preferences for current health obtained with the feeling thermometer and the standard gamble and their validity testing with the SF-36 and HUI2 are important. These results suggest that the feeling thermometer and standard gamble may be measuring different aspects of health related quality of life, which may have implications for their use as research tools. Further validity testing in the area of sensitivity to change is necessary before preferences can be used as a measure of QOL in economic evaluations for osteoporosis therapies.

Our results confirm that hip and spine fractures are associated with reduced HRQOL, particularly in the Physical function domain. Longterm therapy with medications is required for the prevention of osteoporotic fractures, so it is essential that we evaluate the effects of these medications on HRQOL, including the effect of medication related side effects. Future work should focus on whether HRQOL changes over time after a hip fracture, and how various interventions affect health related quality of life.

REFERENCES

1. Goeree R, O'Brien B, Pettit D, Cuddy L, Ferraz M, Adachi J. An assessment of the burden of illness due to osteoporosis in Canada. J Soc Obstet Gynecol 1996;18:S15-24.
2. Eddy D, Johnston CC, Cummings SR, et al. Osteoporosis: review of the evidence for prevention, diagnosis, and treatment and cost-effectiveness analysis. Osteoporos Int 1998;81-8.
3. Cummings S, Black DM, O'Fallon WM, Wahner HW, Riggs BL. Bone density at various sites for prediction of hip fractures. Lancet 1993;341:72-5.
4. Bennett KJ, Torrance GW. Measuring health state preferences and utilities: rating scale, time trade-off, and standard gamble techniques. In: Spilker B, editor. Quality of life and pharmacoeconomics in clinical trials. 2nd ed. Philadelphia: Lippincott-Raven; 1996:253-65.
5. Gold M, Patrick DL, Torrance GW, et al. Identifying and valuing outcomes. In: Gold M, Siegel J, Russell L, Weinstein M, editors. Cost-effectiveness in health and medicine. New York: Oxford University Press; 1997:82-134.
6. Daly E, Gray A, Barlow D, McPherson K, Roche M, Vessey M. Measuring the impact of menopausal symptoms on quality of life. BMJ 1993;307:836-40.
7. Tosteson ANA, Houpt L, Kneeland TS, et al. Longitudinal analysis of utility data from a randomized controlled health economics trial [abstract]. Med Decis Making 1997;17:540.
8. Fordham RJ, Northey K, Rosenberg M. Validating the EuroQol using the Sickness Impact Profile: Longterm outcome after hip fractures. Health Economists' Study Group Meeting, Bristol, UK, 1995.
9. Gabriel SE, Kneeland T, 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.
10. Torrance GW. Social preferences for the health states. An empirical evaluation of three measurement techniques. Socioecon Plan Sci 1976;10:129-36.
11. Torrance GW, Furlong W, Feeny D, Boyle M. Multi-attribute preference functions — Health Utilities Index. Pharmacoeconomics 1995;7:503-20.
12. McHorney CA, Ware JE, Raczek A. The MOS 36-item Short-form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. Med Care 1993;31:247-63.
13. Ware J. The SF-36 Health Survey. In: Spilker B, editor. Quality of life and pharmacoeconomics in clinical trials. 2nd ed. Philadelphia: Lippincott-Raven; 1996:337-45.
14. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Control Clin Trials 1991;12:142S-58S.
15. Muller R, Buttner P. A critical discussion of intraclass correlation coefficients. Stat Med 1994;13:2465-76.
16. Donner AP, Eliasziw M. Sample size requirements for reliability studies. Stat Med 1987;6:441-8.
17. Anderson JJ, Chernoff MC. Sensitivity to change of rheumatoid arthritis clinical trial outcome measures. J Rheumatol 1993; 20:535-7.
18. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1967;86:420-8.
19. Froberg DG, Kane RL. Methodology for measuring health-state preferences. II: scaling methods. J Clin Epidemiol 1989;42:459-71.
20. Bass EB, Wills S, Scott IU, et al. Preference values for visual states in patients planning to undergo cataract surgery. Med Decis Making 1997;17:324-30.
21. Kaplan RM, Ernst JA. Do category rating scales produce biased preference weights for a health index? Med Care 1983;21:193-207.
22. von Neuman J, Morgenstern O. Theory of games and economic behavior. Princeton, NJ: Princeton University Press; 1994.
23. Lalonde L. Psychometric properties of valuational quality of life measures in coronary heart disease prevention and treatment [abstract]. Med Decis Making 1997;17:517.