A Model to Estimate Health Utilities Index Mark 3 Utility Scores from WOMAC Index Scores in Patients with Osteoarthritis of the Knee

PAUL GROOTENDORST, DEBORAH MARSHALL, DAN PERICAK, NICHOLAS BELLAMY, DAVID FEENY, and GEORGE W. TORRANCE

ABSTRACT. ***Objective.*** To develop a formula to translate Western Ontario and McMaster University Osteoarthritis Index (WOMAC) scores collected in clinical trials of patients with osteoarthritis (OA) into Health Utilities Index Mark 3 (HUI3) utility scores for application in economic evaluation.

Methods. Data from a previously published open-label randomized controlled trial of appropriate care with hylan G-F 20 versus appropriate care without hylan G-F 20 in 255 outpatients with knee OA. We estimated linear regression models of HUI3 scores using various functions of WOMAC, demographics, and clinical variables. Out-of-sample predictive performance of the models was assessed using the mean absolute error and several other criteria.

Results. The preferred formula included WOMAC pain, stiffness, function subscales, demographic variables; it accounted for almost 40% of the variation in the HUI3 utility scores. At the group level, absolute differences between predicted and actual overall HUI3 utility scores were < 0.001 and not statistically significantly different from zero.

Conclusion. A formula was derived from the WOMAC index to estimate overall utility scores based on the HUI3 for studies of patients with OA for whom utility has not been recorded. Researchers can estimate overall utility scores, compute quality-adjusted life-years, and perform cost-utility analyses within a defined range of certainty. (J Rheumatol 2007;34:534–42)

Key Indexing Terms:
OSTEOARTHRITIS KNEE QUALITY-ADJUSTED LIFE-YEARS
WESTERN ONTARIO AND McMASTER UNIVERSITY OSTEOARTHRITIS INDEX
HEALTH UTILITIES INDEX REGRESSION ANALYSIS

Healthcare payers routinely use evidence on value for money when considering coverage of new healthcare interventions. Interventions used to manage osteoarthritis (OA) and other chronic diseases that affect primary morbidity, not mortality, are often valued in terms of their effect on health-related quality of life (HRQOL). Most studies assessing health interventions in patients with OA measure HRQOL using the Western Ontario and McMaster University Osteoarthritis Index (WOMAC); the WOMAC measures the degree of pain, mobility, and stiffness in patients with OA. Payers, however, typically require a generic measure of HRQOL, that is, a measure that can be used to prioritize interventions used to treat a variety of health problems. Health-state utilities are a commonly-used generic HRQOL measure. To generate health-state utilities, one starts with a health status classification system that is capable of describing a wide variety of health states. The Health Utilities Index Mark 3 (HUI3) system, for instance, describes 972,000 different health states

From the Faculty of Pharmacy, University of Toronto, Toronto; Department of Economics, McMaster University, Hamilton; Department of Health Economics and Outcomes Research, i3 Innovus, Burlington; Health Utilities Inc., Dundas, Ontario; Departments of Economics and Public Health Sciences, University of Alberta, and Institute for Health Economics, Edmonton, Alberta, Canada; The University of Queensland, Centre of National Research on Disability and Rehabilitation Medicine (CONROD), Queensland, Australia; and Kaiser Permanente Northwest Center for Health Research, Portland, Oregon, USA.

Supported by the Alberta Improvements for Musculoskeletal Disorders Study (AIMS). AIMS was initiated by the Alberta Ministry of Health to improve the care and quality of life for patients with acute and chronic musculoskeletal disorders and is funded by unrestricted grants from Merck Frosst Canada Inc. and Pfizer Canada Ltd. An independent advisory committee was assembled to oversee the study design, analysis, and interpretation of the findings. i3 Innovus is an independent health economics research organization responsible for the design and execution of the study. The authors thank Genzyme Corporation for permitting the data from the hylan G-F study[1] to be used for this study. D. Feeny and G. Torrance have a proprietary interest in Health Utilities Incorporated, Dundas, ON, Canada. HUInc. owns the copyright to and distributes HUI materials. N. Bellamy has a proprietary interest in WOMAC™ and owns the copyright and trademark to and distributes WOMAC materials.

P. Grootendorst, PhD, Associate Professor, Leslie Dan Faculty of Pharmacy, University of Toronto; D. Marshall, PhD, VP,Global Health Economics and Outcomes Research, i3 Innovus; D. Pericak, MMath, PStat, Senior Manager, Health Economics and Outcomes Research, i3 Innovus; N. Bellamy, MD, MSc, MBA, DSc, FRCPC, FRCP (Glas, Edin), FACP, FRACP, Professor and Director, CONROD, The University of Queensland; D. Feeny, PhD, Center for Health Research, Northwest/Hawai'i/Southeast Kaiser Permanente Northwest Region; G.W. Torrance, PhD, Professor Emeritus, McMaster University and Principal Consultant, i3 Innovus.

Address reprint requests to D. Pericak, Health Economics and Outcomes Research, i3 Innovus, 1016-A Sutton Drive, Burlington, Ontario L7L 6B8, Canada. E-mail: dan.pericak@i3innovus.com

Accepted for publication November 9, 2006.

Personal non-commercial use only. The Journal of Rheumatology Copyright © 2007. All rights reserved.

Downloaded on September 30, 2020 from www.jrheum.org

defined by the combinations of levels of function in each of 8 health attributes, including attributes potentially affected by OA: ambulation, dexterity, and pain and discomfort. Next, one attaches a utility score to each possible health state; this score describes the desirability (based on a preference survey of the general population) of the state relative to the 2 anchor states: normal health (usually assigned a utility score of one) and death (usually assigned a utility score of zero). Using the health-state classification and utility function, analysts can compare the utility value of morbidity reductions produced by several candidate interventions targeted at different populations and/or different diseases. Further, these utility values can be combined with evidence on the mortality effects of alternative interventions to rank the interventions' impact on quality-adjusted life-years (QALY). Numerous studies have examined the utility[1,2] associated with OA[3,4] and computed cost-utility analyses, a common approach for evaluating the economic influence of medical care technologies where effects are expressed as quality-adjusted survival, allowing for comparisons across different diseases and interventions.

Most published effectiveness studies in OA measure HRQOL using just the WOMAC. How then can the value of OA treatments be compared to the value of treatments for other conditions? One option — repeating these OA studies using a generic HRQOL measure — would be very costly. A more practical option would be to map WOMAC scores into a generic measure of HRQOL.

We attempted to map the WOMAC into HUI3 utility scores that can ultimately be used in a cost-utility analysis.

MATERIALS AND METHODS

Data source. Data came from a previously reported multicenter, open-label randomized clinical trial (RCT) of patients with OA of the knee randomized (1:1) to either "appropriate care with hylan G-F 20" (AC+H) or to "appropriate care without hylan G-F 20" (AC) and followed for one year. Appropriate care was defined by the Guidelines for the Medical Management of Osteoarthritis of the Knee proposed by the American College of Rheumatology[5]. Patients in this study had symptomatic knee OA of mild to moderate severity that was previously treated with analgesics. Appropriate care could include medications such as analgesics, nonsteroidal antiinflammatory drugs, corticosteroid injections, supportive measures such as education and counseling, weight loss, joint rest, application of heat or ice, the use of devices, physical therapy, arthroscopy, and total joint replacement. Patients randomized to the AC+H group could receive more than one course of hylan G-F 20 treatment in the study knee (i.e., the knee most symptomatic or with the most predominant musculoskeletal problem) if medically warranted, and could receive bilateral treatment if their contralateral knee was affected.

Patients were assessed by the clinical investigator at baseline and 12 months. Both WOMAC and HUI3 were measured at multiple intervals (0, 1, 2, 4, 6, 8, 10, and 12 mo) for 255 patients and were completed by telephone interview from Months 1 to 12. The results showed that hylan G-F 20, when used in conjunction with appropriate care, provides improvement in outcomes that is both clinically important and statistically significant as measured by the change in the WOMAC pain score (38% reduction with hylan G-F 20 and appropriate care vs 13% for appropriate care only; $p = 0.0001$) and HUI3 utility score (0.071 increase; $p < 0.05$)[4,6].

WOMAC Index. The WOMAC has been well validated[7,8] and is widely used in clinical studies of patients with hip and knee OA of varying degrees of severity[8,9]. It is a valid, reliable, and responsive measure of outcome, and has been used in diverse clinical and interventional environments[9,10].

The WOMAC Likert 3.0 scale used in the RCT is a self-administered disease-specific HRQOL instrument that evaluates pain, mobility, and stiffness in a series of 24 questions[9,10]. Each respondent rated the extent of his/her arthritis symptom severity or degree of restriction of usual activities on a 5 point Likert scale from 1 (none) to 5 (extreme). The WOMAC Likert 3.0 provides scores for the 3 subscales: pain (minimum 0, maximum 20; based on 5 questions), stiffness (minimum 0, maximum 8; based on 2 questions), physical function (minimum 0, maximum 68; based on 17 questions); and an aggregate total score (minimum 0, maximum 30) calculated by using a weighted average of the subscale scores. In all cases, the higher the score, the worse the disability. A full description for calculating subscale and aggregate scores is provided in the scoring manual[9].

HUI3. The HUI3 was used in the RCT to generate health-state utility scores[11-13]. HUI3 utility scores have interval scale properties; hence one can compute cross-sectional and time-based differences in scores. The HUI3 has been used in many clinical trials and 5 major Canadian population health surveys, providing extensive data on population norms[14]. The performance of the HUI3 system has been tested extensively[15-18].

Subjects filled out the self-complete HUI3 questionnaire. This form consists of 12 questions to determine functional status in each of 8 attributes (dimensions) of health; each attribute has 5 or 6 levels of functional capacity ranging from normal function to severe disability. The 8 attributes are: vision (5 levels), hearing (6), speech (5), ambulation (6), dexterity (6), emotion (5), cognition (6), and pain and discomfort (5).

The HUI3 system includes a mapping of each 8 element health state into an overall utility score and 8 single-attribute utility scores[19]. The utility scoring formulas are calculated from a multiplicative multiattribute utility function based on preference measurements obtained from a random sample of the general population. Single-attribute utility scores range from 0.0 (most severe disability) to 1.0 (no disability, normal). Overall HUI3 utility scores, which are the focus of this study, range from –0.36 (the score attached to the worst HUI3 health state) to 0.0 (dead) to 1.0 (perfect health). A negative score implies a health state worse than death. Complete instructions on calculating attribute levels and utility scores are provided in the scoring manual[20].

Representation of WOMAC scores in the prediction model. We considered several alternative representations of the WOMAC variables for purposes of predicting HUI3 utility scores. These were: (1) responses to each of the 24 WOMAC questions represented using 4 binary indicators (producing $24 \times 4 = 96$ indicators in total) to represent the response categories: "mild," "moderate," "severe," "extreme." The response "none" was the baseline category; (2) as above, except the response categories "mild" and "moderate" were collapsed into one level, as were the response categories "severe" and "extreme." This reduced the number of indicators from 96 to 48; (3) WOMAC subscale scores: pain, stiffness, and function; (4) WOMAC subscale scores: pain, stiffness, and function and pairwise interaction terms (i.e., pain × stiffness, pain × function, function × stiffness, pain^2, stiffness^2, function^2) to account for possible nonlinearities; and (5) the WOMAC aggregate total score (total) and its square (total^2).

Representation of demographic and clinical variables in the prediction model. Several demographic (age at baseline and sex) and clinical variables (years since onset of OA in the study knee, and Kellgren radiographic grade) were considered important in predicting utility and hence were included in the prediction model. Although these additional clinical variables should improve the predictive performance of the model, some users might not have data on these covariates. We therefore considered several specifications, which exclude some or all of these covariates: (1) no covariates; (2) demographic covariates; (3) demographic covariates + years since onset of OA in the study knee; (4) demographic covariates + years since onset of OA in the study knee + Kellgren radiographic grade; where demographic covariates = age, age^2, sex; years since onset of OA in the study knee = years since onset of OA, years since onset of OA^2; Kellgren radiographic grade = indicators of radiographic grade I, grade II, grade III, grade IV; the radiographic grade 0 was the baseline category.

Personal non-commercial use only. The Journal of Rheumatology Copyright © 2007. All rights reserved.

Downloaded on September 30, 2020 from www.jrheum.org

The 5 representations of the WOMAC and 4 combinations of demographic/clinical covariates yielded 20 different candidate sets of predictor variables.

Estimators. We elected to model HUI3 utility scores as a conventional linear function of the covariates. It was unclear, however, which estimation technique was appropriate, given that the models were estimated using longitudinal data. Given that our goal is prediction, and not the estimation of the individual model measures per se, ordinary least squares (OLS) should yield unbiased predictions even with longitudinal data. On the other hand, predictions might be more precise if the longitudinal nature of the data were explicitly modeled. A conventional fixed effects specification is inappropriate given that estimates are conditional on the subjects in the data, making it impossible to generalize to other subjects. A random effects (RE) specification that decomposes the model error term into a conventional subject- and measurement-specific error and a subject-specific, time-invariant error could be appropriate. We therefore considered both the OLS and RE estimators. Further details on model estimation are provided in the Appendix.

Prediction model selection. We required a method to select the "best" model from the 40 candidate prediction models (2 estimators and 20 sets of covariates). We elected to assess the models on their ability to predict the utility scores of "out-of-sample" subjects — that is, those subjects whose data were not used to estimate the models[21]. Models that fare well on this criterion are also likely to do well in routine use. The measures of each candidate model were thus estimated using a randomly chosen subset of two-thirds of the subjects; the fitted model was used to predict the HUI3 scores of the remaining one-third of the subjects. The difference between the actual and predicted HUI3 scores in the prediction subsample — the "prediction error" — was used to generate the mean absolute error (MAE), which is the average absolute prediction error[21]. The MAE was identified *a priori* as the primary criterion for model selection as it provided a directly interpretable measure of the prediction performance of individual HUI3 scores.

The MAE likely varies from sample to sample, for 2 reasons. First, it depends on the (random) division of the sample into estimation and prediction subsamples, as well as the idiosyncratic components of utility of individuals in the prediction subsample. Second, it also likely depends on the sampling error in the estimates of the model measures. Since the MAE is a random variable, we inferred the precision of our estimator of MAE from the empirical 95% confidence interval (CI) of the bootstrapped distribution of MAE. Specifically, for each candidate model, the model selection procedure (i.e., the selection of the prediction subsample, model estimation, and estimation of MAE) was repeated 500 times, thereby generating 500 MAE estimates. The model with the lowest mean MAE was selected as the preferred model.

For comparison purposes we also evaluated the predictive performance of the models using several other prediction performance statistics: (1) root mean square error (RMSE): the RMSE is the positive square root of the average squared prediction error. In contrast to the MAE, the RMSE attaches greater weight to larger errors and favors prediction models that do not produce particularly large errors at the expense of models that are off by a modest amount; (2) intraclass correlation coefficient (ICC): the ICC, a measure of agreement between subject's predicted and observed scores, was estimated for each model from a 2-way mixed model ANOVA[22,23]. ICC is a ratio of between-subject variance over the sum of the between-subject and within-subject variance. The ICC measure is similar to MAE in that when ICC = 1, MAE = 0. To implement the ICC, a model of the following form was estimated:

$$HUI3[ijt] = \alpha + x[t]ß + v[i] + \varepsilon[ijt]$$

where: HUI3[ijt] = HUI3 score for patient *i* at time *t* measured according to method; *j* = (actual, predicted); α, ß = measures to be estimated; x[t] = time indicators (to pick up systematic variation in HUI3 scores of patient *i* measured according to method *j* over time); v[i] = patient-specific random effect; ε[ijt] = overall random effect. The total residual variance of the model was calculated as: $Var(v[i] + \varepsilon[ijt]) = \tau^2 + \sigma^2$, i.e., the total variance is the sum of a between-subject component (τ^2) and a within-subject component (σ^2). The ICC is the proportion of the total residual variance (denominator) that is due to residual variability between subjects (numerator): $\tau^2/(\tau^2 + \sigma^2)$; (3) mean error (ME), the average prediction error, reflects the accuracy of predictions of group average HUI3 utility scores.

Assessing the prediction precision of the preferred model. In addition to generating HUI3 predictions, users of the prediction model need some sense of how accurate their predictions typically are. Precision will likely vary by the value of the predicted HUI3 score, with extreme values (less than 0 and close to 1) predicted with less precision than more frequently occurring intermediate values. Precision also likely depends on whether one wants to predict mean HUI3 scores for a group of patients or HUI3 scores for individual patients. Intuitively, group-level predictions are more precise because particularly large prediction errors of individuals within the group tend to cancel each other out. Given that most users of the formula would likely be interested in group-level predictions, group-level prediction error CI for all subjects and for subjects in various ranges of the predicted HUI3 score were estimated. This was accomplished using the following procedure. Using the preferred prediction model specification, we predicted utility scores using each subject's baseline WOMAC and other characteristics. We then computed each subject's prediction error; this yielded a data set containing the predicted baseline HUI3 score and associated prediction error for each of the 255 subjects. Forecast precision was assessed by applying a bootstrap procedure to these data. Specifically: (1) we randomly sampled persons, with replacement, from the prediction errors of the 255 subjects, groups of size x = (10, 25, 50, 100, 200, 400); (2) we computed the group mean HUI3 predictions and prediction errors; (3) we repeated steps (1) to (2) 5000 times, thereby generating a distribution of mean prediction errors for each of the 6 group sizes and for various ranges of the predicted HUI3 score. The 2.5th and 97.5th percentiles of the empirical prediction error frequency distribution were used to estimate the prediction error CI.

Finally, we assessed the fraction of absolute subject-specific prediction errors from the primary model that exceeded 0.03. For overall HUI3 scores, differences of 0.03 or more are considered to be clinically important[24-26]. To do so, we randomly sampled, with replacement, 10,000 prediction errors from the baseline prediction errors of the 255 subjects, and tabulated the resulting absolute prediction errors.

RESULTS

Demographics (Table 1). In the trial, 255 patients were randomized to receive appropriate care or appropriate care with hylan G-F 20. Twenty-four subjects dropped out of the study. The average age at baseline was 63.1 years, with a mean of 9.5 years since onset of OA in the study knee. Seventy percent of subjects were female. Fifty percent of subjects had Kellgren radiographic grade III or higher.

Health status at baseline (Table 2). The majority of baseline responses to each WOMAC question for each subscale fell in the moderate to severe range with representation in almost all of the "none," "mild," and "extreme" categories; an appendix containing a summary of the response to each WOMAC and HUI3 question at baseline is available from the author. The mean total WOMAC score was 18.03 [standard deviation (SD) = 3.95]. The WOMAC function subscale score was highest, followed by the pain and stiffness subscales scores (the higher the score, the worse the problem).

The mean overall HUI3 utility score at baseline was 0.48 (SD 0.23); the range of the HUI3 score was –0.21 to 0.92 (Table 2). Single-attribute utility scores revealed that the mean

Personal non-commercial use only. The Journal of Rheumatology Copyright © 2007. All rights reserved.

Downloaded on September 30, 2020 from www.jrheum.org

Table 1. Baseline demographics.

Characteristic	Mean	SD
Continuous Variables (n = 255)		
Age (yrs)	63.06	9.98
Weight (kg)	86.61	19.32
Years since onset of OA in study knee	9.45	9.59
BMI	32.51	7.63
Discrete Variables (n = 255), %		
Other knee affected by OA	84.7	
Other knee requires treatment	54.1	
Joints affected by OA [R or L hip, or spine, or interphalangeal joints (hand), or thumb carpal metacarpal joint, or 1st MTP]	99.6	
Sex		
Female	70.4	
Race		
Caucasian	92	
Black	2	
Asian	3	
Other	3	
Prior surgery of the study knee	31	
Household income, $		
0–20,000	23	
20,000–39,000	34	
40,000–59,000	21	
60,000–79,000	10	
80,000–99,000	4	
100,000+	6	
Radiographic grade		
0	3	
I	11	
II	26	
III	34	
IV	26	

BMI: body mass index = weight in kg/(height in m^2). OA: osteoarthritis.

Table 2. Baseline WOMAC total and subscale scores and HUI3 utility scores.

WOMAC Scores (n = 255)	Mean (SD)
WOMAC total score	18.03 (3.95)
WOMAC subscale scores	
Pain (0 to 20)	11.64 (2.81)
Stiffness (0 to 8)	5.08 (1.46)
Function (0 to 68)	39.87 (9.25)
HUI3 Utility scores (n = 255)	
Overall	0.48 (0.23)
Single-attribute utility scores	
Vision	0.93 (0.88)
Hearing	0.93 (0.19)
Speech	0.98 (0.08)
Ambulation	0.92 (0.14)
Dexterity	0.95 (0.13)
Emotion	0.92 (0.16)
Cognition	0.92 (0.16)
Pain	0.53 (0.25)

HUI3: Health Utility Index — Mark 3.

pain utility score was the lowest (0.53) among all the 8 attributes.

Prediction model. There were 2186 observations from 255 patients with a maximum of 8 assessments each. A total of 353 observations were dropped due to missing covariates, leaving 1833 complete records.

We initially compared the OLS to RE estimators. In each specification considered, OLS forecast precision performed better than the RE forecasts using the MAE criterion. Moreover, the models in which WOMAC was represented as a function of its subscale scores (pain, stiffness, function, pain × stiffness, pain × function, stiffness × function, $pain^2$, $stiffness^2$, $function^2$) dominated the other representations. In what follows we report the results of the OLS estimation of the 4 remaining candidate models — i.e., those distinguished by the choice of demographic and clinical covariates. The prediction performance of these models as measured by all 4 selection statistics (MAE, RMSE, ICC, and ME) is shown in Table 3. Models that included the WOMAC subscale scores and their pairwise interaction terms, as well as demographics (age, age^2, sex), and the years of OA in the study knee (yrs of OA, yrs of OA^2) performed marginally better than did models without these covariates. The ICC and ME selection statistics favored models that also included indicators of Kellgren radiographic grade, although again differences in prediction performance were slight. This is clear from Figure 1, which displays the bootstrapped distribution for the 4 best prediction models for each selection statistic.

Table 4 reports the estimates and standard errors of the prediction model with the best MAE score. All variables were retained in the model regardless of statistical significance. This equation is:

$$\begin{aligned}\text{Predicted HUI3 utility score} = {} & 0.5274776 + 0.0079676 \times \text{Pain} + 0.0065111 \times \text{Stiffness} - 0.0059571 \times \text{Function} + \\ & 0.0019928 \times \text{Pain} \times \text{Stiffness} + 0.0010734 \times \text{Pain} \times \text{Function} + 0.0001018 \times \text{Stiffness} \times \text{Function} - 0.0030813 \\ & \times \text{Pain}^2 - 0.0016583 \times \text{Stiffness}^2 - 0.000243 \times \text{Function}^2 + 0.0113565 \times \text{Age in years} - 0.0000961 \times \text{Age in years}^2 \\ & - 0.0172294 \times \text{Female} - 0.0057865 \times \text{Years since onset of OA in the study knee} + 0.0001609 \times \text{Years since onset of OA in the study knee}^2\end{aligned}$$

To illustrate, this model predicts that a 56-year-old woman with pain, stiffness, and function scores of 10, 5, and 24, respectively, who had OA for 2.5 years would have a HUI3 score of 0.68. This model accounted for 39% of the variation in HUI3 utility scores. At the group level, the absolute difference between the predicted and the actual utility scores was < 0.001 and not statistically significantly different from zero. The function subscale score was statistically significant and negative, both alone and as a higher order term in the model.

If Kellgren radiographic grade is available in addition to information on WOMAC scores, demographics, and years

Personal non-commercial use only. The Journal of Rheumatology Copyright © 2007. All rights reserved.

Downloaded on September 30, 2020 from www.jrheum.org

Table 3. Comparison of model performance criteria.

Performance Criterion	Model	Mean	95% CI
MAE (Primary Model)	Model 3	0.1628	0.1457 to 0.1779
	Model 4	0.1638	0.1473 to 0.1797
	Model 1	0.1645	0.1486 to 0.1798
	Model 2	0.1652	0.1488 to 0.1813
RMSE	Model 3	0.2065	0.1846 to 0.2273
	Model 4	0.2083	0.1864 to 0.2294
	Model 1	0.2083	0.1872 to 0.2290
	Model 2	0.2095	0.1868 to 0.2310
ICC	Model 4	0.5572	0.4723 to 0.6262
	Model 3	0.5557	0.4696 to 0.62934
	Model 2	0.5379	0.4536 to 0.61341
	Model 1	0.5359	0.45134 to 0.6110
ME	Model 4	–0.0003	–0.0445 to 0.0441
	Model 2	–0.0005	–0.0431 to 0.0427
	Model 3	–0.0006	–0.0422 to 0.0397
	Model 1	–0.0007	–0.04421 to 0.0412

RMSE: Root mean square error; MAE: Mean absolute error; ME: mean error; ICC: Intraclass correlation coefficient. WOMAC = f\{Pain, Stiffness, Function, (Pain × Stiffness), (Pain × Function), (Stiffness × Function, Pain^2, Stiffness^2, Function^2\}. DEMOG = f \{Age, Age^2, Sex\}. YRSOA = f \{Years since onset of OA, Years since onset of OA^2\}. X-Ray = f \{Kellgren radiographic grade I, II, III, IV\}. Model 1 = WOMAC; Model 2 = WOMAC + DEMOG; Model 3 = WOMAC + DEMOG + YRSOA; Model 4 = WOMAC + DEMOG + YRSOA + X-Ray.

since onset of OA, an alternative prediction model can be applied. This alternative model (Appendix, Model 4) accounts for a similar proportion of the variation in HUI3 utility scores (40%) as the primary model and performs slightly better than the primary model on the ME criterion.

Turning next to the forecast precision of the primary model, Table 5 reports the proportion of observations for which the absolute difference between the predicted and actual baseline HUI3 utility scores exceeded 0.03, which is commonly considered to be a clinical or policy important difference[26]. Only 10% of absolute differences were less than 0.03; the majority of differences were in excess of 0.10.

The primary model does not appear to produce sufficiently precise individual level predictions. How did it fare at predicting group-level baseline HUI3 scores? Figure 2 and Table 6 display the group-level prediction precision of the HUI3 utility scores by group size and by the size of the predicted HUI3 score. For each of the group sizes considered, most group-level predictions were between 0.4 and 0.5, although the proportion of predictions that fell between 0.4 and 0.5 increased with group size (Figure 2). Prediction errors tended to shrink with group size as well: with groups of size 10, errors range from –0.2 to +0.2. When group size is 25, almost all errors are between –0.1 and +0.1. Table 6 provides 95% prediction error CI specific to one's group size and estimated HUI3 score. For example, if the mean predicted HUI3 utility score in a group of size 25 was 0.45, then the 95% CI around the group-level prediction would be (0.45 – 0.06) to (0.45 + 0.08) = 0.39 to 0.53; prediction error CI for group sizes between 10 and 400 can be interpolated. Note that even with 10,000 bootstrap replications, there were insufficient realizations to construct CI for all combinations of group mean predicted HUI3 values and group size.

DISCUSSION

Generic measures of HRQOL are commonly used in clinical and policy research, and are particularly useful for economic appraisal of interventions to manage diseases that reduce functional capacity. Most studies assessing health interventions in patients with OA, however, measure HRQOL using the WOMAC, which is an OA-specific HRQOL measure. Our study offers a method for estimating generic HRQOL scores from individual-level WOMAC scores collected in trials that did not collect data on generic HRQOL.

We measured generic HRQOL using HUI3 utility scores. The HUI3 prediction models explained between 39% and 40% of the variation in the HUI3 scores of subjects in our data, all of whom had mild to moderate OA of the knee. This predictive performance is moderately lower than the performance of models that predict utility scores based on the Medical Outcomes Study Short Form-36 Health Survey (SF-36) and SF-12 with ranges of performance of 47% to 55%[27-30]. Moreover, our models were unable to make reliable predictions on individual subject utility scores. For instance, about 90% of predictions exceeded actual HUI3 utility scores by 0.03; 0.03 is commonly held to be a clinically important difference in HUI3 scores. The average absolute prediction error was about 0.16.

Some of the prediction error can be attributed to the measurement error in the 2 instruments. The reliability of the WOMAC and HUI3 instruments is in the range of 0.72 to 0.77[31]; Cronbach's alpha for the Likert version of the

Personal non-commercial use only. The Journal of Rheumatology Copyright © 2007. All rights reserved.

Downloaded on September 30, 2020 from www.jrheum.org

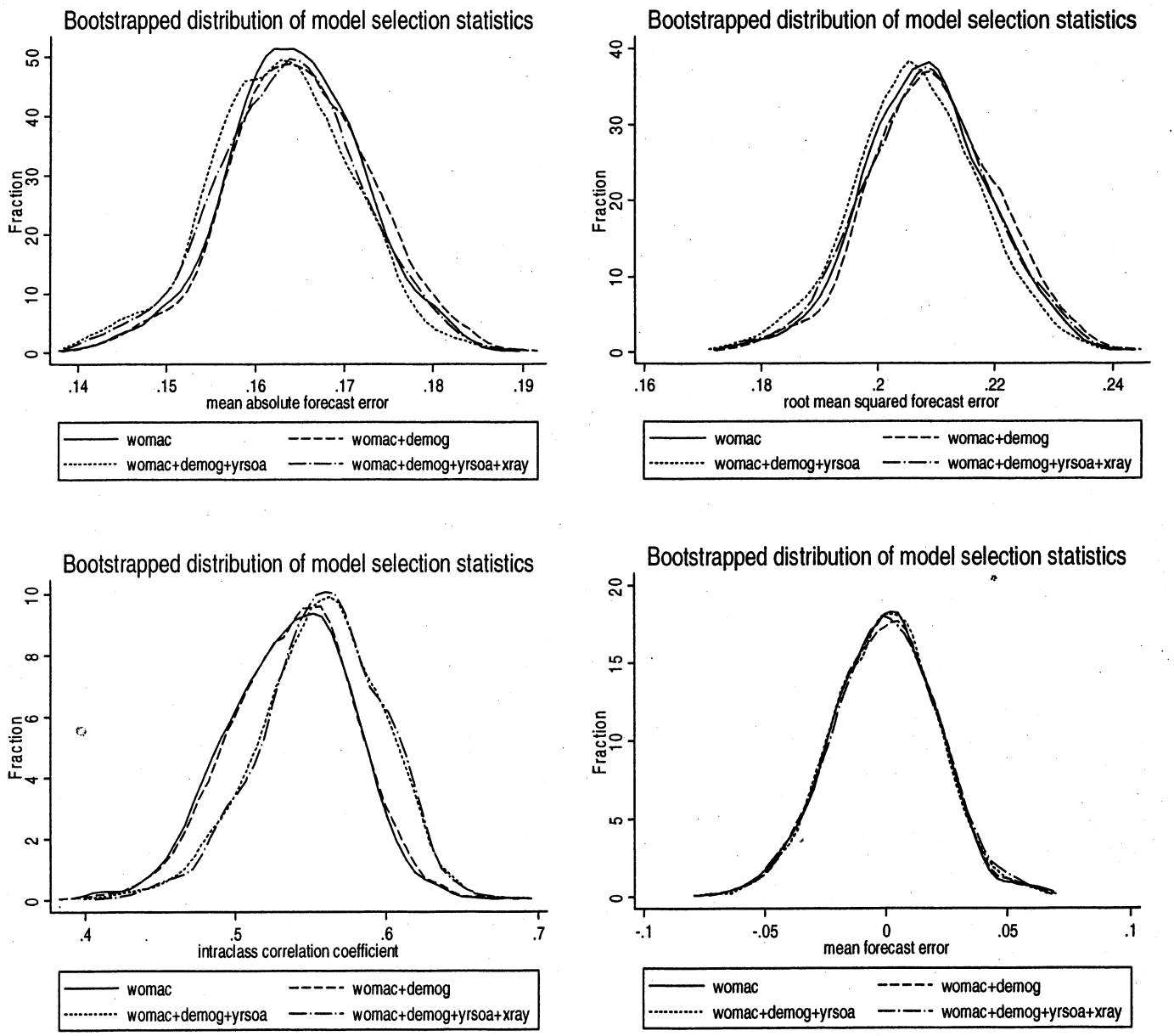

Figure 1. Bootstrapped distribution using model selection statistics MAE, RMSE, ICC, and ME. WOMAC = *f*[Pain, Stiffness, Function, (Pain × Stiffness), (Pain × Function), (Stiffness × Function), Pain^2, Stiffness^2, Function^2]; DEMOG = *f*[Age, Age^2, Sex]; YRSOA = *f*[Yrs since onset of OA, Yrs since onset of OA^2]; X-RAY = *f*[Kellgren radiographic grade I, II, III, IV].

WOMAC from the 2 original validation studies is 0.75 to 0.97[9,10]. The majority of the prediction error, however, is likely due to the variation in HUI3 scores that is attributable to variation in vision, emotion, hearing, and other domains recorded in the HUI3 that are unaffected by OA[27-30].

While the models were unable to precisely predict HUI3 scores of individual subjects, the models were able to generate reasonably precise group-average HUI3 scores. The reason is that subject-specific forecast errors tend to cancel each other out when taking the group average. We therefore recommend that the formula be used to estimate HUI3 utility scores on the basis of group average characteristics (WOMAC scores, demographics and severity of OA). This is not overly restrictive — economic appraisal studies typically compare treatment-group-specific average HRQOL and costs.

To allow users to assess the precision of their forecasts, we have provided estimates of the forecast error and CI for various values of group average HUI3 utility score predictions and group sizes. Prediction performance for groups of patients with typical levels of dysfunction (HUI3 predictions between 0.4 and 0.6) are reasonable. For instance, the 95% CI for a group of 50 OA patients predicted to have an HUI3 score of between 0.4 and 0.6 would be about 0.10 utility units in width.

Our preferred model predicts HUI3 scores using data on

Personal non-commercial use only. The Journal of Rheumatology Copyright © 2007. All rights reserved.

Downloaded on September 30, 2020 from www.jrheum.org

Table 4. Prediction model for estimating HUI3 scores from WOMAC (primary model).

Variable	Beta Coefficient	SE	t	p > |t|	95% CI
Pain	0.0080	0.0067	1.20	0.232	–0.0051 to 0.0210
Stiffness	0.0065	0.0120	0.54	0.588	–0.0171 to 0.0301
Function	–0.0060	0.0022	–2.70	0.007	–0.0103 to –0.0016
Pain × Stiffness	0.0020	0.0020	0.98	0.328	–0.0020 to 0.0060
Pain × Function	0.0010734	0.0004	2.86	0.004	0.0003 to 0.0018
Stiffness × Function	0.0001	0.0007	0.15	0.882	–0.0012 to 0.0014
Pain^2	–0.0031	0.0007	–4.42	0.000	–0.0045 to –0.0017
Stiffness^2	–0.0017	0.0025	–0.67	0.500	–0.0065 to 0.0032
Function^2	–0.0002	0.0001	–3.14	0.002	–0.0004 to –0.0001
Age	0.0114	0.0052	2.18	0.029	0.0011 to 0.0216
Age^2	–0.0001	0.00004	–2.32	0.020	–0.0002 to –0.00001
Female	–0.0172	0.0105	–1.65	0.099	–0.0377 to 0.0033
Years since onset of OA	–0.0058	0.0013	–4.36	0.000	–0.0084 to –0.0032
Years since onset of OA^2	0.0002	0.00003	4.98	0.000	0.0001 to 0.0002
Intercept	0.5275	0.1610	3.28	0.001	0.2117 to 0.8433

WOMAC Pain, Stiffness, Function, Age at baseline, and Years since onset of OA were continuous variables. Adjusted R-squared: 0.3895, SE: standard error.

Table 5. Important differences between observed and predicted HUI3 utility scores at the individual level. Predictions are from the primary prediction model reported in Table 4.

Important Difference	n = 10,000	%
|Deltal		
|Deltal < 0.01	494	4.9
0.01 < |Deltal ≤ 0.03	499	5.0
0.03 < |Deltal ≤ 0.05	856	8.6
0.05 < |Deltal ≤ 0.10	1,697	17.0
|Deltal > 0.10	6,454	64.5

WOMAC subscale scores, as well as their pairwise interactions, demographics, and years since onset of OA. We also report, however, several other prediction models that performed almost as well as the preferred model on the MAE criterion (or slightly better than the preferred model on the other criteria). All of these models predict HUI3 using WOMAC subscale scores and their pairwise interactions, but include as additional covariates various combinations of the demographic, years since onset of OA, and radiographic grade as additional predictors. We present these more parsimonious models

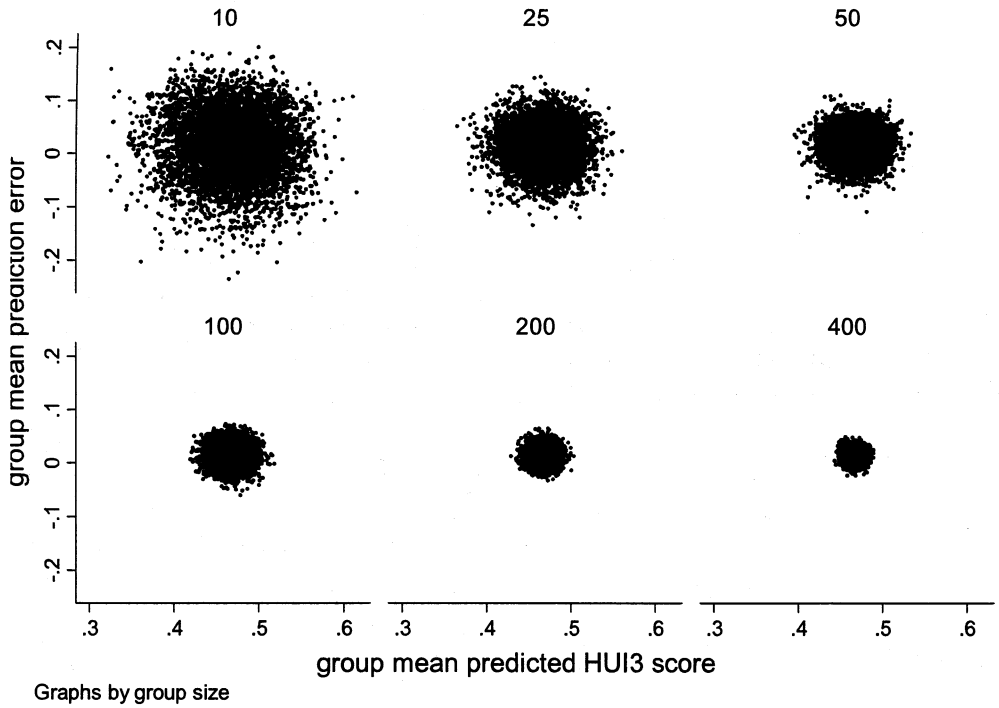

Figure 2. Group mean prediction errors, by group size (10, 25, 50, 100, 200, 400) and group mean predicted HUI3 score.

Personal non-commercial use only. The Journal of Rheumatology Copyright © 2007. All rights reserved.

Downloaded on September 30, 2020 from www.jrheum.org

Table 6. Group level prediction precision and estimated prediction error by range of predicted HUI3 score and group size. There were no realizations for some of the combinations of predicted HUI3 score and group size.

Range of Predicted HUI3 Score	Group Size	Lower 95% CI of Forecast Error Distribution	Upper 95% CI of Forecast Error Distribution	CI Width
0.3 ≤ mean pred HUI3 < 0.4	10	–0.106	0.120	0.226
0.3 ≤ mean pred HUI3 < 0.4	25	–0.039	0.074	0.113
0.3 ≤ mean pred HUI3 < 0.4	50	0.031	0.036	0.005
0.4 ≤ mean pred HUI3 < 0.5	10	–0.103	0.125	0.228
0.4 ≤ mean pred HUI3 < 0.5	25	–0.059	0.084	0.143
0.4 ≤ mean pred HUI3 < 0.5	50	–0.037	0.065	0.102
0.4 ≤ mean pred HUI3 < 0.5	100	–0.021	0.050	0.071
0.4 ≤ mean pred HUI3 < 0.5	200	–0.011	0.040	0.051
0.4 ≤ mean pred HUI3 < 0.5	400	–0.004	0.033	0.037
0.5 ≤ mean pred HUI3 < 0.6	10	–0.105	0.114	0.218
0.5 ≤ mean pred HUI3 < 0.6	25	–0.065	0.079	0.144
0.5 ≤ mean pred HUI3 < 0.6	50	–0.042	0.065	0.107
0.5 ≤ mean pred HUI3 < 0.6	100	–0.034	0.040	0.074
0.5 ≤ mean pred HUI3 < 0.6	200	–0.014	0.017	0.031
0.6 ≤ mean pred HUI3 < 0.7	10	–0.074	0.107	0.181

for the benefit of users who lack data on these additional covariates. The group-level forecast accuracy of these more parsimonious models is very similar to the accuracy of the preferred model.

There are several limitations of our study. First, the prediction model ideally applies to patients with OA of the knee similar to patients included in the RCT from which the data were obtained. The ranges of values (minimum to maximum) for the patients in the RCT were: WOMAC pain (4 to 20), WOMAC stiffness (1 to 8), WOMAC function (13 to 66), age (40 to 87 yrs), years since onset of OA in the study knee (0.25 to 52 yrs), and radiographic grade (0 to IV). It seems likely that our results would generalize well to patients with OA of the hip and OA in general, but this remains unknown. Second, as stated earlier, the models can predict with reasonable accuracy HUI3 utility scores given group-level average WOMAC, demographic, and clinical characteristics. Predictions for individual subjects are highly inaccurate. The prediction models presented here should therefore not replace the use of a utility instrument in future clinical studies. Third, there are some concerns that the HUI3 system is not sufficiently sensitive to changes in function of the lower limbs and hands and fingers. Future research should investigate whether other generic HRQOL instruments, such as the EuroQoL[18,20,26,32], capture the sequelae of OA better.

With these limitations, our results offer a method for estimating health utility scores from existing WOMAC scores collected in trials where there is no other means to do so. To our knowledge, this is the first study that has attempted to establish a formula to translate WOMAC scores to HUI3 scores. Investigators who need to estimate utilities should be able to use this model for estimating QALY for decision and cost-utility analyses.

APPENDIX

Prediction models of the following form were estimated using linear regression:

$$(utility[it]^{\lambda} - 1)/\lambda = \alpha + \beta^{T}X[it] + \delta^{T}W[it] + \gamma^{T}Z[it] + \varepsilon[it]$$

where utility[it] refers to HUI3 utility score for patient i at measurement t; X[it] is a vector of WOMAC scores; W[it] is a vector of demographic variables; and Z[it] is a vector of clinical variables for patient i at measurement t; ß, δ, and γ are conformable vectors of unknown measures, and ε[it] is a random disturbance. Only complete data were included in the regression modeling, but otherwise all observations were used for model estimation.

Equation 3 allows for utility to be transformed using the "Box-Cox" transformation: $(utility[it]^{\lambda} - 1)/\lambda$. The transformation takes on several special cases, depending on the value of λ. When λ = 1, utility remains linear; when λ = –1, the inverse value of utility results; and when λ = 0, the log of utility results. The choice of transformation was assessed by estimating λ along with the unknown model measures using a linear model incorporating a variety of different sets of covariates. The Box-Cox transformation is undefined for HUI3 scores that were not strictly positive[33]; these observations were therefore removed when estimating λ. In each case, the estimated value of λ was most consistent with the linear specification (λ = 1). The HUI3 utility score therefore remained untransformed.

Prediction model for estimating HUI3 scores from WOMAC. For the convenience of users who lack information on clinical and/or demographic characteristics of subjects for whom HUI3 utility score predictions are required, we give estimates of prediction models in which these variables were removed.

Model 1: WOMAC subscales with their interactions and their square terms only.

Predicted HUI3 utility score = 0.8228595 + 0.0102259 × Pain + 0.0100088 × Stiffness – 0.0074078 × Function + 0.0029179 × Pain × Stiffness + 0.0011218 × Pain × Function + 0.0000113 × Stiffness × Function – 0.0034621 × $Pain^2$ – 0.0026578 × $Stiffness^2$ – 0.0002302 × $Function^2$

Model 2: WOMAC subscales with their interactions and their square terms as well as clinical variables of age and its second-order term and sex.

Predicted HUI3 utility score = 0.5897559 + 0.0096178 × Pain + 0.0074462 × Stiffness – 0.0068115 × Function + 0.0026835 × Pain × Stiffness + 0.0010606 × Pain × Function + 0.0001674 × Stiffness × Function – 0.0032652 × $Pain^2$ – 0.0027313 × $Stiffness^2$ – 0.0002396 × $Function^2$ + 0.0087576 × Age – 0.0000744 × Age^2 – 0.0251635 × Female

Personal non-commercial use only. The Journal of Rheumatology Copyright © 2007. All rights reserved.

Downloaded on September 30, 2020 from www.jrheum.org

Model 3 (primary model): WOMAC subscales, their interactions and their square terms, age, OA duration (in the study knee) and their square terms as well as sex.

Predicted HUI3 utility score = 0.5274776 + 0.0079676 × Pain + 0.0065111 × Stiffness – 0.0059571 × Function + 0.0019928 × Pain × Stiffness + 0.0010734 × Pain × Function + 0.0001018 × Stiffness × Function – 0.0030813 × $Pain^2$ – 0.0016583 × $Stiffness^2$ – 0.0002430 × $Function^2$ + 0.0113565 × Age – 0.0000961 × Age^2 – 0.0172294 × Female – 0.0057865 × OA duration + 0.0001609 × $OA\ duration^2$

Model 4: WOMAC subscales, their interactions and their square terms, age, duration of OA (in the study knee) and their square terms as well as sex and Kellgren radiographic grade.

Predicted HUI3 utility score = 0.5044234 + 0.0084581 × Pain + 0.0035271 × Stiffness – 0.0054986 × Function + 0.0021076 × Pain × Stiffness + 0.0010924 × Pain × Function + 0.0000657 × Stiffness × Function – 0.0031476 × $Pain^2$ –0.0012983 × $Stiffness^2$ – 0.0002503 × $Function^2$ + 0.0141542 × Age – 0.0001192 × Age^2 – 0.0212673 × Female – 0.0055305 × OA duration + 0.0001571 × $OA\ duration^2$ – 0.0837873 × radiographic grade I – 0.0458229 × grade II – 0.0572539 × grade III – 0.0872116 × grade IV

REFERENCES

1. Chapman RH, Stone PW, Sandberg EA, Bell C, Neumann PJ. A comprehensive league table of cost utility ratios and a sub-table of 'panel worthy' studies. Med Decis Making 2000;20:451-67.
2. Torrance GW, Tugwell P, Amorosi S, Chartash E, Sengupta N. Improvement in health utility among patients with rheumatoid arthritis treated with adalimumab (a human anti-TNF monoclonal antibody) plus methotrexate. Rheumatology Oxford 2004;43:712-8. Epub 2004 Mar 23.
3. Marshall DA, Strauss ME, Pericak D, Buitendyk M, Codding C, Torrance GW. Economic evaluation of controlled-release oxycodone vs oxycodone-acetaminophen for osteoarthritis pain of the hip or knee. Am J Manag Care 2006;12:205-14.
4. Torrance GW, Raynauld JP, Walker V, et al. Canadian Knee OA Study Group. A prospective, randomized, pragmatic, health outcomes trial evaluating the incorporation of hylan G-F 20 into the treatment paradigm for patients with knee osteoarthritis (Part 2 of 2): economic results. Osteoarthritis Cartilage 2002;10:518-27.
5. Hochberg MC, Altman RD, Brandt KD, et al. Guidelines for the medical management of osteoarthritis. Part II. Osteoarthritis of the knee. Arthritis Rheum 1995;38:1541-6.
6. Raynauld JP, Torrance GW, Band PA, et al, in collaboration with the Canadian Knee OA Study Group. A prospective, randomized, pragmatic, health outcomes trial evaluating the incorporation of hylan G-F 20 into the treatment paradigm for patients with knee osteoarthritis (Part 1 of 2): clinical results. Osteoarthritis Cartilage 2002;10:506-17.
7. Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. Validation study of WOMAC: a health status instrument for measuring clinically-important patient-relevant outcomes following total hip or knee arthroplasty in osteoarthritis. J Orthop Rheumatol 1998;1:95-108.
8. Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. Validation study of WOMAC: A health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. J Rheumatol 1988;15:1833-40.
9. Bellamy N. WOMAC Osteoarthritis Index user guide. Version VII. Brisbane, Australia; 2002.
10. http://www.womac.com (Accessed November 22, 2006).
11. Feeny D, Furlong W, Boyle M, Torrance GW. Multi-attribute health status classification systems, health utilities index. Pharmacoeconomics 1995;7:490-502.
12. Torrance GW, Furlong W, Feeny D, Boyle M. Multi-attribute preference functions, health utilities index. Pharmacoeconomics 1995;7:503-20.
13. Torrance GW, Feeny DH, Furlong W. Health Utilities Index. Technical document, July 1992. Hamilton, Ontario: McMaster University; 1992.
14. http://www.healthutilities.com/update.htm (Accessed November 22, 2006).
15. Torrance GW, Furlong W, Feeny D. Health utility estimation. Exp Rev Pharmacoeconomics Outcomes Res 2002;2:99-108.
16. Eiser C, Morse R. The measurement of quality of life in children: past and future perspectives. J Dev Behav Pediatr 2001;22:248-56.
17. Feeny DH, Torrance GW, Furlong WJ. Health Utilities Index. In: Spilker B, editor. Quality of life and pharmacoeconomics in clinical trials. 2nd ed. Philadelphia: Lippincott-Raven Press; 1996:239-52.
18. Furlong WJ, Feeny D, 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.
19. 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.
20. Torrance GW, Feeny D, Furlong W, Goldsmith C, De Pauw S, Zhu Z. Health Utilities Index Mark 3 (HUI3): second preliminary multiplicative multi-attribute and single-attribute utility scoring functions. Hamilton: McMaster University; 1998.
21. Kennedy P. A guide to econometrics. 5th ed. Cambridge, MA: MIT Press; 2003.
22. Schuck P. Assessing reproducibility for interval data in health-related quality of life questionnaires: which coefficient should be used? Qual Life Res 2004;13:571-86.
23. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Control Clin Trials 1991;12 Suppl:142S-58S.
24. Drummond M. Introducing economic and quality of life measurements into clinical studies. Ann Med 2001;33:344-9.
25. Grootendorst P, Feeny D, Furlong W. Health Utilities Index Mark 3: evidence of construct validity for stroke and arthritis in a population health survey. Med Care 2000;38:290-9.
26. Horsman J, Furlong W, Feeny D, Torrance GW. The Health Utilities Index (HUI®): concepts, measurement properties and applications. Health Qual Life Outcomes 2003;1:54.
27. Nichol MB, Sengupta N, Globe DR. Evaluating quality-adjusted life years: estimation of the Health Utility Index (HUI) from the SF-36. Med Decis Making 2001;21:105-12.
28. Franks P, Lubetkin EI, Gold MR, Tancredi DJ. Mapping the SF-12 to preference-based instruments: convergent validity in a low-income, minority population. Med Care 2003;41:1277-83.
29. Franks P, Lubetkin EI, Gold MR, Tancredi DJ, Jia H. Mapping the SF-12 to the EuroQol EQ-5D Index in a national US sample. Med Decis Making 2004;24:247-54.
30. 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 Making 1997;17:1-9.
31. Jones CA, Feeny D, Eng K. Test-retest reliability of Health Utilities Index scores: evidence from hip fracture. Int J Technol Assess Health Care 2005;21:393-8.
32. Rabin R, de Charro F. EQ-5D: A measure of health status from the EuroQol Group. Ann Med 2001;33:337-43.
33. Box GEP, Cox DR. An analysis of transformations. J Roy Stat Soc 1964;26:211-43.

Personal non-commercial use only. The Journal of Rheumatology Copyright © 2007. All rights reserved.

Downloaded on September 30, 2020 from www.jrheum.org