Agreement about Identifying Patients Who Change over Time : Cautionary Results in Cataract and Heart Failure Patients

David Feeny, Karen Spritzer, Ron D. Hays, Honghu Liu, Theodore G. Ganiats, Robert M. Kaplan, Mari Palta and Dennis G. Fryback

Med Decis Making 2012 32: 273 originally published online 18 October 2011
DOI: 10.1177/0272989X11418671

Agreement about Identifying Patients Who Change over Time: Cautionary Results in Cataract and Heart Failure Patients

David Feeny, PhD, Karen Spritzer, BA, Ron D. Hays, PhD, Honghu Liu, PhD, Theodore G. Ganiats, MD, Robert M. Kaplan, PhD, Mari Palta, PhD, Dennis G. Fryback, PhD

Background. *Preference-based measures of health-related quality of life all use the same dead = 0.00 to perfect health = 1.00 scale, but there are substantial differences among measures.* ***Objective.*** *The objective was to examine agreement in classifying patients as better, stable, or worse.* ***Methods.*** *The EQ-5D, Health Utilities Index Mark 2 and Mark 3, Quality of Well-Being–Self-Administered scale, Short-Form 36 (Short-Form 6D), and disease-targeted measures were administered prospectively in 2 clinical cohorts. The study was conducted at academic medical centers: University of California, Los Angeles; University of California, San Diego; University of Wisconsin–Madison; and University of Southern California. Patients undergoing cataract extraction surgery with lens replacement completed the 25-item National Eye Institute Visual Function Questionnaire (NEI-VFQ-25). Patients newly referred to congestive heart failure specialty clinics completed the Minnesota Living with Heart Failure Questionnaire (MLHF). In both cohorts, subjects completed surveys at baseline and at 1 and 6 months. The NEI-VFQ-25 and MLHF were used as gold standards to assign patients to categories of change. Agreement was assessed using* κ. ***Results.*** *There were 376 cataract patients recruited. Complete data for baseline and the 1-month follow-up were available on all measures for 210 cases. Using criteria specified by Altman, agreement was poor for 6 of 9 pairs of comparisons and fair for 3 pairs. There were 160 heart failure patients recruited. Complete data for baseline and the 6-month follow-up were available for 86 cases. Agreement was negligible for 5 pairs and fair for 1. The study was conducted on selected patients at a few academic medical centers.* ***Conclusions.*** *The results underscore the lack of interchangeability among different preference-based measures.* ***Key words:*** *scale development/validation; severity of illness measures; population-based studies; access to care; outcomes research; public health; psychometric methods/scaling; preventive medicine/screening; education; family medicine; performance measurement; preventive services; primary care; randomized trial methodology; risk factor evaluation.* ***(Med Decis Making 2012;32:273–286)***

Preference-based measures of health-related quality of life (HRQL) are needed for monitoring population health and for program evaluation for comparative effectiveness research. Most importantly, these measures are required for estimating quality-adjusted life years (QALYs). A number of widely used, generic, preference-based measures are available such as the EQ-5D,[1] Health Utilities Index Mark 2 (HUI2) and Mark 3 (HUI3),[2] the Quality of Well-Being–Self-Administered scale (QWB-SA),[3] and Short-Form 6D (SF-6D).[4,5] Although these measures share a common core[6,7] and all include items on mobility, mental health, and pain, there are also important differences with respect to which attributes (dimensions or domains of health status) are included. HUI and the QWB-SA include vision, hearing, speech, and dexterity; the EQ-5D and SF-6D do not. The QWB-SA is unique in that it includes 58 symptoms or health problems, only some of which are included in the other measures. These measures also differ in the range of function or symptom severity covered in each attribute. The QWB-SA asks respondents if they have or do not have a problem such as pain and stiffness; in contrast, HUI and SF-6D have gradients such as the categories mild, moderate, and severe pain.

These measures also differ with respect to the methods that were used to elicit preference scores with which to estimate their respective multiattribute

DOI: 10.1177/0272989X11418671

scoring functions, the methods for estimating those functions, and their functional forms.[8] For instance, the QWB-SA scoring function is based on valuations using the visual analog scale (VAS) and a linear additive scoring function. SF-6D is based on the standard gamble (SG) and an ad hoc modified linear additive functional form. EQ-5D is based on the time tradeoff (TTO) and an ad hoc modified linear additive functional form. HUI is based on transformed VAS and SG scores and a multiplicative functional form.

It is therefore not surprising that several investigators who have used 2 or more measures have concluded that the scores from these measures are not interchangeable.[9–14] Further, there is evidence from prospective studies that the estimates of absolute and/or relative change (responsiveness, including effect size [ES] and the standardized response mean [SRM])[15] often do not agree.[12,16–19]

The objective of this article was to examine agreement among the above measures in classifying patients into the same categories of change: We want to know if the measures agree on which patients get better, remain stable, or get worse. Data from 2 prospective cohort studies that employed all 5 of the above measures as well as disease-targeted measures are used to assess agreement among these measures: one study of patients undergoing cataract surgery, and the other of patients referred for treatment for congestive heart failure by a specialty clinic.

This article builds on an earlier work[19] based on the data from the same study. That work provided cohort-level estimates of responsiveness (SRM) for each of the 5 preference-based measures in each of the 2 cohorts. Responsiveness varied among measures and across cohorts. Results from that work underscore the lack of interchangeability of scores among these measures.

This article asks an important follow-up question: Even if overall responsiveness differs among measures, do they agree on who gets better, who gets worse, and who was stable?

METHODS

Patients

Subjects for both components of the study had to be at least 35 years of age, able to give informed consent, able to hear and understand instructions in English, and have sufficient vision and ability in reading and writing English to complete questionnaires.[19]

Cataract surgery. Patients were undergoing cataract extraction surgery with lens replacement. Patients were excluded if undergoing simultaneous glaucoma, corneal, or vitroretinal procedures or if they were unable to read large print versions of questionnaires.

Heart failure. Patients were newly referred to congestive heart failure clinics. Inclusion criteria included evidence of the presence of heart failure for at least 3 months defined as a left ventricular ejection fraction less than 40%. Patients classified as class IV in the New York Heart Association system, those with a recent ($\leq$6 months) myocardial infarction, unstable angina, recent ($\leq$3 months) coronary artery bypass graft surgery, those on the heart transplant list, or those with recent ($\leq$3 months) ventricular tachycardia were excluded.

Participants were recruited from 4 academic medical centers: the University of California, Los

Received 27 August 2010 from The Center for Health Research, Kaiser Permanente Northwest and Health Utilities Incorporated, Portland, OR (DF); Department of Medicine, University of California, Los Angeles (KS, RDH); School of Dentistry, University of California, Los Angeles (HL); Department of Family and Preventive Medicine, University of California, San Diego (TGG); Department of Health Services Research, University of California, Los Angeles (RMK); and Department of Population Health Sciences, University of Wisconsin-Madison (MP, DGF). An earlier version of this work was presented at the 2010 meeting of Health Technology Assessment International, Dublin, 6-9 June 2010, and at the 17th Annual Meeting of the International Society for Quality of Life Research, London, 27-30 October 2010. This work was supported by grant P01AG020679 from the National Institute on Aging (NIA). Drs. Kaplan and Hays were also provided support by National Institutes of Health (NIH) grant 1 P01 AG020679-01A2, the UCLA Claude D. Pepper Older Americans Independence Center, NIH/NIA grant 5P30AG028748, and Centers for Disease Control (CDC) grant U48 DP000056-04. Dr. Hays also received support from the UCLA Resource Center for Minority Aging Research/Center for Health Improvement in Minority Elderly (P30AG021684) and the UCLA/DREW Project EXPORT (P20MD000148 and P20MD000182). The funding agreement ensured the independence of the authors in the design, conduct, interpretation, data, writing, and publishing of the article. The granting agencies have neither read nor approved of the contents of the article. Dr. Feeny has a proprietary interest in Health Utilities Inc. (Dundas, Ontario, Canada). Health Utilities Inc. distributes copyrighted Health Utilities Index (HUI) materials and provides methodological advice on the use of HUI. None of the other authors declare a conflict of interest. Revision accepted for publication 21 June 2011.

Address correspondence to David Feeny, PhD, The Center for Health Research, Kaiser Permanente Northwest, 3800 North Interstate Avenue, Portland, OR 97227-1110; telephone: (503) 528-3937; fax: (503) 335-2428; e-mail:david.feeny@kpchr.org.

Angeles (UCLA); the University of California, San Diego (UCSD); the University of Wisconsin; and the University of Southern California (cataract patients). The study was approved by the Institutional Review Boards at each of these institutions (UCLA IRB #G05-06-096-11, UCSD Project #070435, Wisconsin M-2005-1171, and USC #HS-06-00493).

Procedures

At enrollment, patients were given a packet of self-administered questionnaires to complete and mail back to the UCSD Health Services Research Center (HSRC) within 7 days. The HSCR mailed out the same packet for the 1- and 6-month follow-up surveys.

Measures

The study included 5 of the most commonly used preference-based measures.[8] There is substantial evidence on the reliability, cross-sectional construct validity, and responsiveness (longitudinal construct validity) of each of these measures in a wide variety of applications. The study also used a widely used disease-targeted measure for vision (25-item National Eye Institute Visual Function Questionnaire [NEI-VFQ-25])[20–22] and a prominent disease-targeted measure for heart failure (Minnesota Living with Heart Failure Questionnaire [MLHF]).[23–26]

EQ-5D. The health status classification system of EQ-5D includes 5 attributes (mobility, self-care, usual activity, pain/discomfort, and anxiety/depression) with 3 levels (no problem, some problem, and extreme problem) per attribute.[1] The EQ-5D also includes a VAS on which respondents provide a rating of their current overall health; the analyses reported here do not include the VAS scores. Health status at a point in time for a subject is described as a 5-element vector, one level for each attribute. Preference-based scores for EQ-5D health states were derived using a scoring function based on TTO preferences elicited from a random sample of community-dwelling residents of the United States and estimated with an ad hoc modified linear additive utility function.[27] Scores are defined on the conventional scale in which dead = 0.00 and perfect health = 1.00; EQ-5D scores range from –0.11 (states worse than dead) to 1.00.

HUI2. [28,29] HUI2 includes 7 attributes: sensation (vision, hearing, speech), mobility, emotion, cognition, self-care, pain, and fertility. (The item on fertility was not administered in this study; fertility was assumed to be normal at level 1.) There are 4 or 5 levels per attribute in HUI2. The multiplicative HUI2 scoring function is based on preference elicitation using the VAS and SG from a random sample of community-dwelling subjects in Canada.[28] Single-attribute utility scores are on a scale in which 0.00 is the score of the most disabled level in that attribute and 1.00 is the score for level 1 (no problem or disability in that attribute). Overall HUI2 scores vary from –0.03 to 1.00. In addition to the overall HUI2 score, the single-attribute HUI2 sensation score was included in the analyses of data from the cataract cohort because of its relevance as a specific measure of visual function.

HUI3. HUI3 includes 8 attributes (vision, hearing, speech, ambulation, dexterity, emotion, cognition, and pain and discomfort) with 5 or 6 levels per attribute. The multiplicative HUI3 scoring function is based on preference elicitation using the VAS and SG from a random sample of community-dwelling subjects in Canada.[29] Overall HUI3 scores vary from –0.36 to 1.00. In addition to the overall HUI3 score, the single-attribute HUI3 vision score was included in the analyses of data from the cataract cohort because of its relevance.

QWB-SA. The QWB-SA assesses self-reported functioning using a series of questions designed to record limitations in the previous 3 days, within 3 separate domains (mobility, physical activity, and social activity). In addition, QWB-SA includes a series of questions that ask about the presence or absence of different symptom/problem complexes. The 4 domain scores are combined into a total score that provides a numerical point-in-time expression of well-being that ranges from 0.00 for dead to 1.00 for asymptomatic optimum functioning. The original QWB obtained preference ratings of 856 people from the general population.[30] The QWB-SA used convenience samples to model preference for case descriptions, and the models were shown to be highly correlated with the population ratings in the original QWB general population preferences elicitation survey. Scores range from 0.00 to 1.00; 0.09 is the minimum for a living health state. The self-administered QWB-SA has been shown to be highly correlated with the interviewer-administered QWB and to retain the psychometric properties. Extensive evaluations of reliability and validity have been published.[3,30–32]

Self-rated health (SRH). The SRH item,[33] "In general, would you say that your health is excellent,

very good, good, fair, or poor,'' is a widely used measure of overall health and was therefore included in the analyses.

SF-6D. SF-6D is a preference-based measure based on a subset of items from the SF-36 (or SF-12).[4,5,34] SF-6D includes 6 attributes (physical functioning, role limitations, social functioning, pain, mental health, and vitality) with 4 to 6 levels per attribute. The scoring function is based on SG preferences elicited from a random sample of community-dwelling subjects in the United Kingdom and estimated using an ad hoc linear additive functional form.[4]

NEI-VFQ-25. The NEI-VFQ-25 was designed to capture the influence of vision on a number of dimensions of HRQL including emotional well-being and social functioning.[20–22] The NEI-VFQ-25 includes 25 items covering general health, general vision, near vision, distance vision, driving, peripheral vision, color vision, ocular pain, role limitations, dependency, social function, mental health, and expectations. The total score ranges from 0 to 100, with higher scores signifying better (less impaired) vision.

Visual Function Questionnaire–Utility (VFQ-UI). Recently, a preference-based index scoring system has been developed for the NEI-VFQ-25[35] (Kowalski and others [submitted]; Rentz and others [submitted]), the VFQ-UI. The VFQ-UI includes a single item representing each of 6 domains of the NEI-VFQ: near vision (see well up close), distance vision (going out for films, sports events), role function (limited work time due to vision), mental health (worry about doing things that may embarrass because of vision), vision dependency (stay at home because of vision), and social function (see people's reaction to things I say). The items were selected to cover a range of vision-related functioning using Rasch analyses on samples of patients with central vision loss or peripheral vision loss. The VFQ-UI defines 8 vision-related health states ranging from no difficulty to stopped doing work scored on a 0.00 (dead) to 1.00 perfect health range using TTO-derived preference scores.

MLHF. The MLHF includes 21 items covering symptoms, mental health, social life, fatigue, appetite, mobility, sleep, sexual activity, work and recreational activities, and side effects of treatment.[23–26] Overall scores range from 0 to 105, with higher scores signifying greater impairment (lower HRQL).

Criteria for Clinically Important Change

It is important to assess a measured change with respect both to its statistical significance and its clinical importance or magnitude. Guyatt and others[36(p377)] provide a definition of a clinically important difference: ''The MID [minimum important difference] is the smallest difference in score in the domain of interest that patients perceive as important, either beneficial or harmful, and which would lead the clinician to consider a change in the patient's management.'' There are 2 major methods for determining the clinical importance of a given magnitude of change: anchor-based and distribution-based approaches.[36–43] In the anchor-based approach, the change in HRQL score is related to a known anchor. The anchor itself must be an independent measure and be readily interpretable such as the categories of the New York Heart Association functional classification system or ability to climb a flight of stairs. Further, there must be an appreciable association between the anchor and the target measure.[36] In the distribution-based approach, the magnitude of change is compared to some measure of the variability of scores. Cohen's guidance on classifying effect sizes is an example: 0.20, small; 0.50, medium; and 0.80, large.[44] The anchor-based approach provides an estimate of clinically important change, while the distribution-based approach provides a basis for translating raw score change into standardized units that can be used for comparisons with estimates from prior studies or existing rules of thumb.[40]

For this study, a change of 0.03 or more in the overall preference score for each of the preference-based measures is interpreted as a clinically important change.[2,8,11,37,45–56] Empirical estimates of clinically important change (differences) for the 5 preference-based measures vary from 0.01 to 0.08, with 0.03 being well represented in estimates for each of these measures.

For the single-attribute utility scores for HUI2 sensation (which includes vision) and HUI3 vision, the guideline for a clinically important difference is 0.05.[2] For the NEI-VFQ-25, a change of 5.0 or more in the composite score on a 0-to-100 scale is regarded as clinically important.[57] For the MLHF instrument, a change of 5.0 or more in the total score (range = 0–105) is regarded as clinically important.[23–25] For SRH (excellent, very good, good, fair, poor), a movement of one or more categories is considered clinically important.

Statistical Analyses

Previous work[19] indicated that patients undergoing cataract surgery changed substantially between baseline and the 1-month follow-up survey (after surgery) and were typically then stable in the period between the 1- and 6-month follow-ups. Analyses for the cataract cohort therefore focus on change between the baseline to 1-month follow-up. Improvement was more gradual in the heart failure cohort.[19] Analyses focus on the change between baseline and the 6-month follow-up.

Measures of agreement. Relative agreement in direction and size among change scores for the 10 measures used in the cataract cohort and 7 measures used in the heart failure cohort was assessed using an intraclass correlation coefficient (ICC) based on a 2-way mixed analysis of variance model (measures fixed, patients as random). Agreement between the disease-targeted measure (NEI-VFQ-25 for cataracts, MLHF for congestive heart failure) and each of the 5 (EQ-5D, HUI2, HUI3, QWB-SA, and SF-6D) preference-based measures and SRH as to whether patients had improved, were stable, or got worse was assessed using a number of measures including the percentage agreement, κ (unweighted and weighted), and the delta statistic, a measure of agreement that is less sensitive than κ to the marginal distributions.[58] The degree of agreement (κ) was interpreted according to the criteria suggested by Altman[59]: <0.20, poor; 0.21–0.40, fair; 0.41–0.60, moderate; 0.61–0.80, good; and 0.81–1.00, very good. In addition, regarding the 2 disease-specific measures as gold standards, the sensitivity of each of the 6 generic measures to change on the disease-targeted instruments was estimated using receiver operating characteristic (ROC) curve analyses. The ROC analyses determine if the results are sensitive to the choice of the threshold for clinically important change (0.03) on the preference-based measures.

Primary analyses were conducted on a subset of subjects for whom there are complete data at baseline and the 1-month follow-up (cataract cohort) and baseline and the 6-month follow-up (heart failure cohort) for all of the measured included in the analyses. Thus, any differences in agreement across measures will not be the result of differences in the subjects excluded due to missing data. Secondary data analyses were conducted for the larger sample size for which data at baseline and the designated follow-up (all available pairs with complete data) and the sample size vary by pair of measures.

RESULTS

A total of 376 cataract patients and 160 heart failure patients were recruited to the study. The majority of patients were white; cataract patients tended to be female; heart failure patients tended to be male; most cataract patients were >65 years; and the heart failure patients tended to be younger, with the majority in the 45- to 64-year age group (Table 1).

For the cataract cohort, data for baseline and 1-month follow-up assessments were available for 315 of the 376 cases. Complete data for all pairs for all measures were available for 210 cases. The distribution of demographic variables for those with and without complete data was similar, and there were no statistically significant differences between the 2 groups.

For the heart failure cohort, data for baseline and the 6-month follow-up assessments were available for 110 of the 160 cases. Complete data for all pairs for all measures were available for 86 cases. Those with missing data were older than those without missing data, and the difference between the 2 groups was statistically significant.

The overwhelming majority of respondents, 93%, reported that no one helped them to complete the questionnaires; 7% reported receiving help. Among those who received any help, 90% reported that someone read the questions to them; 55% reported that someone wrote the answers on the questionnaire for them; 6% reported that someone answered the questions for them; 4% reported that someone translated the questions into their language for them; and 9% reported some other kind of help. Therefore, the overwhelming majority of responses were based on self-completion and self-assessment.

Scores for each of the measures at baseline and 1 month and the change scores for the cataract cohort are displayed in Table 2. Note that the mean change in the total score for the NEI-VFQ-25 (VFQ_t) of 9.96 exceeds the guideline for a clinically important difference of 5.00. Similarly, the mean change in overall HUI3 scores exceeds the 0.03 clinically important difference guideline. The mean changes in HUI3 vision score and HUI2 sensation score exceed the 0.05 clinically important difference guideline. The mean changes in scores for EQ-5D, QWB-SA, SF-6D, and SRH are less than the guidelines for a clinically important difference. The distribution of change scores for the VFQ_t is displayed in Figure 1. Using the change of 5 or more in VFQ_t score as

Table 1 Demographic Characteristics of the Samples

	Cataract Patients				Heart Failure Patients			
Demographic	Enrolled (n = 376)	Complete Data (in Sample) (n = 210)	Incomplete Data (Not in Sample) (n = 166)	In Sample versus Not	Enrolled (n = 160)	Complete Data (in Sample) (n = 86)	Incomplete Data (Not in Sample) (n = 74)	In Sample versus Not
Age, y								
35–44	5 (1)	3 (1)	2 (1)	chi(2) =	24 (15)	11 (13)	13 (18)	chi(2) =
45–64	115 (31)	71 (34)	44 (27)	2.42	101 (63)	63 (73)	38 (51)	8.96
65–91	256 (68)	136 (65)	120 (72)	Pr = 0.2977	35 (22)	12 (14)	23 (31)	Pr = 0.0113
Race								
White	328 (87)	184 (88)	144 (87)	chi(3) =	126 (79)	74 (86)	52 (70)	chi(3) =
Black	12 (3)	7 (3)	5 (3)	1.20	19 (12)	8 (9)	11 (15)	6.01
Asian	19 (5)	13 (6)	6 (4)	Pr = 0.7535	5 (3)	1 (1)	4 (5)	Pr = 0.1083
Other	4 (1)	2 (1)	2 (1)		2 (1)	2 (2)	0 (0)	
Missing[a]	13 (3)	4 (2)	9 (5)		8 (5)	1 (1)	7 (9)	
Education								
<High school	21 (6)	6 (3)	15 (9)	chi(6) =	20 (13)	7 (8)	13 (18)	chi(6) =
High school graduate	60 (16)	32 (15)	28 (17)	10.81	45 (28)	25 (29)	20 (27)	10.55
Some college	78 (21)	43 (20)	35 (21)	Pr = 0.0943	47 (29)	33 (38)	14 (19)	Pr = 0.1034
2-year associate degree	27 (7)	14 (7)	13 (8)		12 (8)	7 (8)	5 (7)	
4-year college graduate	90 (24)	56 (27)	34 (20)		16 (10)	7 (8)	9 (12)	
Master's degree	57 (15)	38 (18)	19 (11)		9 (6)	3 (3)	6 (8)	
Doctorate/professional	34 (9)	19 (9)	15 (44)		6 (4)	4 (5)	2 (3)	
Missing[a]	9 (2)	2 (1)	7 (4)		5 (3)	0 (0)	5 (7)	
Female	222 (59)	124 (59)	98 (59)	chi(1) = 0.00 Pr = 0.9982	52 (33)	26 (30)	26 (35)	chi(1) = 0.44 Pr = 0.5092

a. Missing not used in tests.

the criterion, 43% of patients improved, 52% were stable, and 4% got worse.

Scores for each of the measures at baseline and 6 months and the mean change scores for the heart failure cohort are displayed in Table 3. Note the mean change of 8.72 in the score for the MLHF exceeds the 5.00 guideline for a clinically important difference. The mean changes in QWB-SA, SF-6D, and HUI3 scores exceed the guideline, while the mean changes in scores for the EQ-5D, HUI2, and SRH do not. The distribution of change scores for the MLHF is displayed in Figure 2. Using the change of 5 or more in MLHF score as the criterion, 47% of patients improved, 35% were stable, and 19% got worse.

Agreement among change scores. The ICC among the 10 measures of change in the cataract cohort was 0.16 (95% confidence interval [CI] = 0.02–0.29). The ICC among the 7 measures of change in the heart failure cohort was 0.07 (95% CI = –0.14 to 0.28).

Agreement in cataract cohort. The percentage agreement varies between 33% and 57% and is displayed in Table 4 along with simple and weighted κ statistics for agreement between pairs of measures in classifying patients as improved, stable, or worse. The simple κ statistics for 6 of the 9 pairs are poor, and the κ statistics for 3 of the pairs are fair. Fair agreement was obtained between the NEI-VFQ-25 total scores and the vision-targeted measures: HUI2 sensation, HUI3 vision, and the VFQ-UI. The results for weighted κ are very similar to the results for the simple κ. Results for the delta statistics are also very similar, ranging from –0.05 (VFQ_t and SRH) to 0.31 (VFQ_t and HUI3 vision) to 0.40 (VFQ_t and VFQ-UI). Area under the curve results for the ROC analyses range from 0.44 (SRH) to 0.67 (HUI3 vision) to 0.72 (VFQ-UI). In many cases in the ROC analyses, the area under the curve is less than 0.60, indicating agreement is little better than one would expect by chance. These results indicate that the lack of agreement is not sensitive to the choice of cut points for clinically important differences. Finally, results from secondary analyses for n = 315 (subjects for whom observations on any measure were available at baseline and at the 1-

Table 2 Baseline, 1-Month, and Change Scores of Cataract Cohort (n = 210)

	Measure	Mean	Median	Standard Deviation	Minimum	Maximum
Baseline	VFQ_t	76.51	80.63	15.42	17.23	98.67
	EQ-5D	0.83	0.83	0.17	0.08	1.00
	HUI2	0.79	0.82	0.17	0.08	1.00
	HUI2 sensation	0.76	0.76	0.14	0.00	1.00
	HUI3	0.66	0.69	0.27	−0.28	1.00
	HUI3 vision	0.80	0.95	0.22	0.00	1.00
	QWB-SA	0.59	0.61	0.14	0.15	1.00
	SF-6D	0.74	0.75	0.12	0.33	1.00
	SRH	2.50	2.00	0.93	1.00	5.00
	VFQ_{ui}	0.86	0.92	0.12	0.41	0.97
1 month	VFQ_t	86.47	90.13	12.57	26.83	100.00
	EQ-5D	0.84	0.83	0.16	0.17	1.00
	HUI2	0.81	0.87	0.19	0.12	1.00
	HUI2 sensation	0.84	0.87	0.17	0.00	1.00
	HUI3	0.72	0.80	0.28	−0.32	1.00
	HUI3 vision	0.91	0.95	0.15	0.00	1.00
	QWB-SA	0.60	0.61	0.14	0.15	0.97
	SF-6D	0.73	0.74	0.12	0.39	1.00
	SRH	2.52	2.00	0.92	1.00	5.00
	VFQ_{ui}	0.90	0.94	0.09	0.46	0.97
Change	VFQ_t	9.96	7.86	13.44	−25.49	71.63
	EQ-5D	0.02	0.00	0.13	−0.51	0.54
	HUI2	0.02	0.03	0.14	−0.61	0.50
	HUI2 sensation	0.08	0.00	0.19	−0.65	1.00
	HUI3	0.05	0.03	0.21	−0.82	0.77
	HUI3 vision	0.12	0.00	0.22	−0.38	1.00
	QWB-SA	0.01	0.00	0.13	−0.63	0.39
	SF-6D	−0.01	0.00	0.09	−0.29	0.22
	SRH	0.02	0.00	0.69	−3.00	2.00
	VFQ_{ui}	0.04	0.01	0.11	−0.24	0.53

Note: EQ-5D = 5-dimension Euro QOL measure; HUI2 = Health Utilities Index Mark 2; HUI3 = Health Utilities Index Mark 3; QWB-SA = Quality of Well-Being–Self-Administered scale; SF-6D = Short-Form 6D; SRH = self-rated health; VFQ_t = total score of National Eye Institute Visual Function Questionnaire (NEI-VFQ-25); VFQ_{ui} = preference-based score based on NEI-VFQ-25.

month follow-up) were very similar to the results reported in Table 4 (data not shown).

Agreement in heart failure cohort. The percentage agreement varies between 19% and 49% and is displayed along with simple and weighted κ statistics for agreement between pairs of measures in classifying patients as improved, stable, or worse for the heart failure cohort in Table 5. The simple κ statistics are negative for 5 pairs, indicating agreement is less than that which would occur by chance. Agreement between the MLFH and SRH is fair. Results for the weighted κ are very similar. The results for the delta statistics also indicate little agreement, ranging from –0.33 (QWB-SA) to 0.26 (SRH). Area under the curve results from the ROC analyses range from 0.31 (QWB-SA) to 0.73 (SRH) and indicate that the results are not sensitive to the choice of cut points. Finally, results from secondary analyses for n = 110 (subjects for whom observations on any measure were available at baseline and at the 6-month follow-up) were very similar to the results reported in Table 5 (data not shown).

Agreement among measures on classification of patients as worse, stable, or improved. Results on the extent of agreement among measures in classifying patients as improved, stable, or deteriorated for the cataract cohort are found in Table 6. Analogous results for the heart failure cohort are found in Table 7. The lack of agreement among measures evident in the ICC results reported above is evident in Tables 6 and 7. Nonetheless, many observations are aligned ''on the diagonal,'' indicating that there is some agreement between the disease-specific measures, VFQ and MLHF, and each of the 5 preference-based

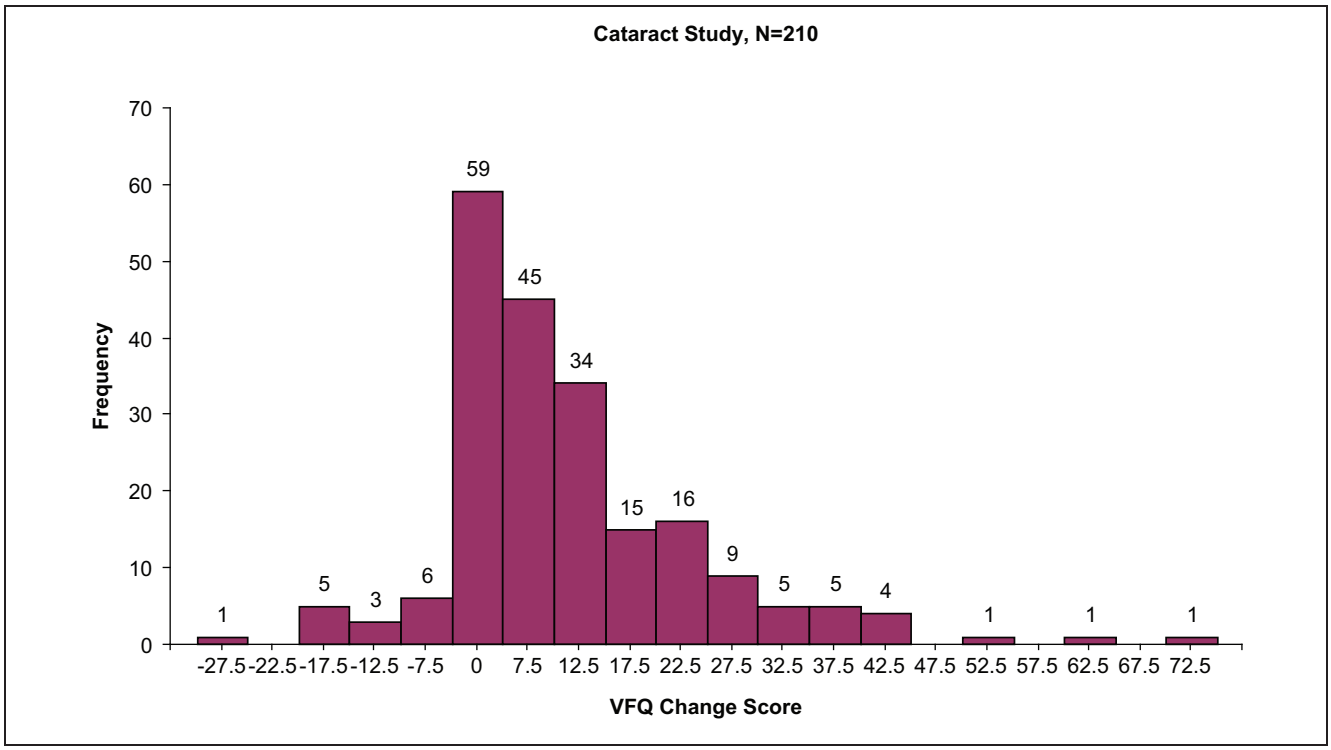

Figure 1 Distribution of change in the 25-item National Eye Institute Visual Function Questionnaire total scores.

Table 3 Baseline, 6-Month, and Change Scores of Heart Failure Cohort (n = 86)

	Measure	Mean	Median	Standard Deviation	Minimum	Maximum
Baseline	MLHF	48.26	48.50	25.40	0.00	101.00
	EQ-5D	0.77	0.79	0.18	0.18	1.00
	HUI2	0.76	0.85	0.22	0.14	1.00
	HUI3	0.62	0.73	0.32	−0.25	1.00
	QWB-SA	0.54	0.55	0.14	0.22	0.87
	SF-6D	0.63	0.63	0.11	0.39	0.93
	SRH	3.79	4.00	0.83	2.00	5.00
6 months	MLHF	39.53	36.50	24.97	0.00	89.00
	EQ-5D	0.76	0.78	0.18	0.21	1.00
	HUI2	0.76	0.84	0.24	0.04	1.00
	HUI3	0.65	0.74	0.33	−0.34	1.00
	QWB-SA	0.58	0.59	0.15	0.21	1.00
	SF-6D	0.66	0.64	0.13	0.41	1.00
	SRH	3.49	4.00	0.98	1.00	5.00
Change	MLHF	−8.72	−7.50	22.12	−69.00	60.00
	EQ-5D	−0.01	0.00	0.18	−0.63	0.52
	HUI2	0.00	0.00	0.17	−0.68	0.43
	HUI3	0.03	0.00	0.23	−0.69	0.87
	QWB-SA	0.04	0.02	0.16	−0.45	0.46
	SF-6D	0.03	0.02	0.11	−0.17	0.42
	SRH	−0.30	0.00	0.90	−3.00	1.00

Note: EQ-5D = 5-dimension Euro QOL measure; HUI2 = Health Utilities Index Mark 2; HUI3 = Health Utilities Index Mark 3; MLHF = Minnesota Living with Heart Failure; QWB-SA = Quality of Well-Being–Self-Administered scale; SF-6D = Short-Form 6D; SRH = self-rated health.

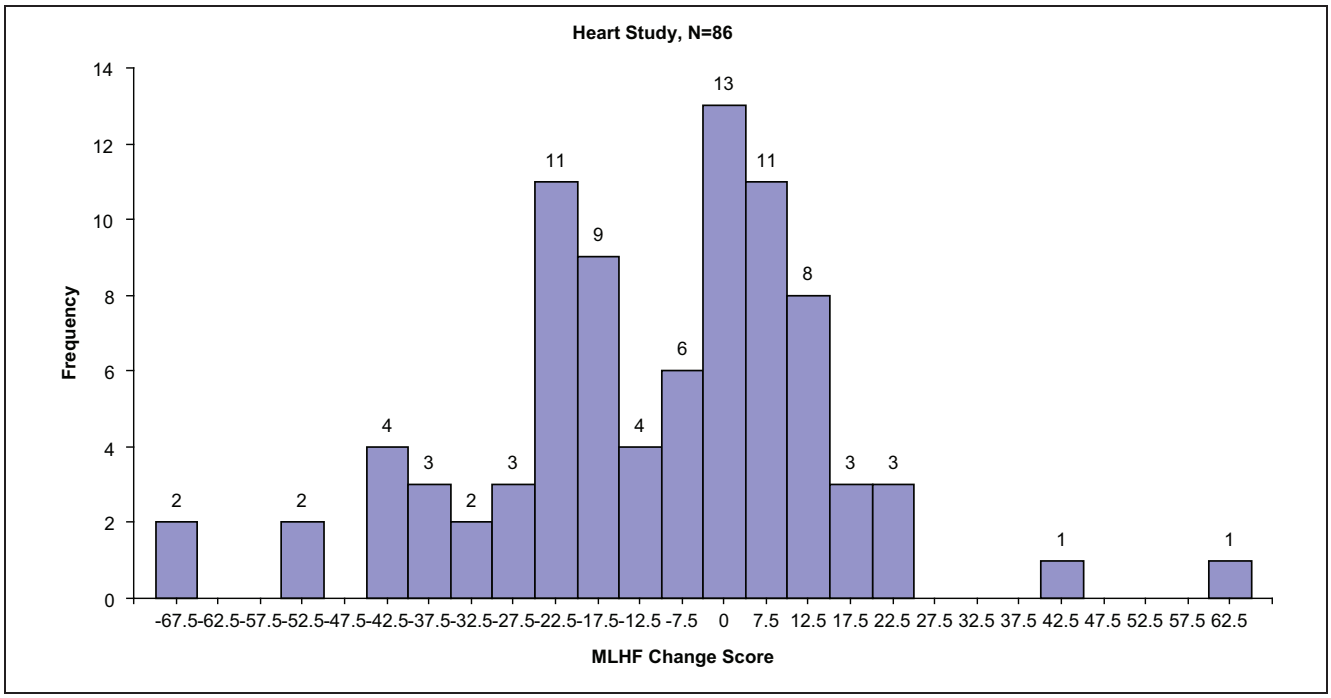

Figure 2 Distribution of change in Minnesota Living with Heart Failure Questionnaire total scores.

measures, on which patients changed and which did not.

DISCUSSION

There is very little pair-wise agreement between the disease-targeted measures and the 5 preference-based measures about which patients improved, were stable, or deteriorated. In general, agreement for the cataract cohort was poor and for the heart failure cohort was negligible. For the cataract cohort, the agreements between the relevant HUI single-attribute (''disease-targeted'') scores and the NEI-VFQ-25, those for HUI2 sensation and HUI3 vision, were the exceptions; agreement was fair. Agreement was also fair between the utility scored and conventional versions of the NEI-VFQ-25. Given that both of these measures are based on the same questionnaire, it is perhaps surprising that the agreement is only fair and is not clearly higher than agreement between the NEI-VFQ-25 and HUI2 sensation and HUI3 vision. In the heart failure cohort, fair agreement was observed only for the SRH. On the basis of the ROC analyses, the results reported here appear to be robust to the choice of cut points for a clinically important change.

The agreement analyses treat the disease-targeted measure as the gold standard. Yet, even though there is evidence of cross-sectional and longitudinal construct validity for the 2 disease-targeted measures, neither can be regarded as a true gold standard. Furthermore, that vision-related or heart-related HRQL improved does not necessarily imply that overall HRQL improved. It is possible that the side effects of interventions could more than offset the gains and therefore overall HRQL might not improve. It is also possible that even though vision- or heart-related HRQL improved, overall HRQL did not due to the burdens associated with comorbidities. The NEI-VFQ-25 asks subjects about a wide variety of difficulties that they might experience due to limited vision, including reading, hobbies, navigating, driving, going up and down stairs, interacting with others, dressing, and the amount of assistance the subject needs from others. Similarly, the MLHF asks about limitations in/problems with mobility, sexual activity, interacting with others, fatigue, hobbies, worry, concentration, memory, and depression that the subject experiences due to the subject's heart condition. Although the breadth of coverage of these disease-targeted measures probably reduces the scope for a discrepancy between trends in vision- or

Table 4 Agreement among 10 Measures in Cataract Cohort (n = 210)

Pair	% Agreement	κ Statistic	95% Confidence Interval	Weighted κ Statistic	95% Confidence Interval
VFQ_t and EQ-5D	39	0.08	−0.01 to 0.16	0.10	0.01 to 0.18
VFQ_t and HUI2	48	0.11	0.01 to 0.21	0.14	0.04 to 0.25
VFQ_t and HUI2 sensation	55	0.22	0.11 to 0.32	0.22	0.12 to 0.32
VFQ_t and HUI3	44	0.07	−0.02 to 0.17	0.11	0.01 to 0.21
VFQ_t and HUI3 vision	57	0.25	0.15 to 0.35	0.25	0.15 to 0.36
VFQ_t and QWB-SA	40	0.06	−0.02 to 0.15	0.12	0.03 to 0.21
VFQ_t and SF-6D	39	0.09	0.00 to 0.17	0.09	0.00 to 0.17
VFQ_t and SRH	33	0.01	−0.06 to 0.08	−0.05	−0.12 to 0.02
VFQ_t and VFQ_{ui}	56	0.26	0.17 to 0.35	0.33	0.24 to 0.42

Note: EQ-5D = 5-dimension Euro QOL measure; HUI2 = Health Utilities Index Mark 2; HUI3 = Health Utilities Index Mark 3; QWB-SA = Quality of Well-Being–Self-Administered scale; SF-6D = Short-Form 6D; VFQ_t = total score of National Eye Institute Visual Function Questionnaire (NEI-VFQ-25); VFQ_{ui} = preference-based score based on NEI-VFQ-25. Self-rated health (SRH) was scored as 1 = poor, 2 = fair, 3 = good, 4 = very good, and 5 = excellent.

Table 5 Agreement among 7 Measures in Heart Failure Cohort (n = 86)

Pair	% Agreement	κ Statistic	95% Confidence Interval	Weighted κ Statistic	95% Confidence Interval
MLHF and EQ-5D	19	−0.25	−0.37 to –0.13	−0.30	−0.45 to –0.15
MLHF and HUI2	29	−0.10	−0.26 to 0.05	−0.19	−0.36 to –0.02
MLHF and HUI3	26	−0.17	−0.32 to –0.02	−0.23	−0.40 to –0.06
MLHF and QWB-SA	22	−0.22	−0.37 to –0.07	−0.30	−0.46 to –0.14
MLHF and SF-6D	26	−0.11	−0.25 to 0.04	−0.22	−0.38 to –0.06
MLHF and SRH	49	0.25	0.12 to 0.39	0.34	0.19 to 0.49

Note: EQ-5D = 5-dimension Euro QOL measure; HUI2 = Health Utilities Index Mark 2; HUI3 = Health Utilities Index Mark 3; MLHF = Minnesota Living with Heart Failure; QWB-SA = Quality of Well-Being–Self-Administered scale; SF-6D = Short-Form 6D. Self-rated health (SRH) was scored as 1 = poor, 2 = fair, 3 = good, 4 = very good, and 5 = excellent.

heart-related HRQL and overall HRQL, it does not eliminate the possibility for such discrepancies.

The results on overall change in measures underscore that scores from these 5 preference-based measures are not interchangeable (Table 2). In the cataract cohort, using published guidelines[2,57] on clinically important differences/changes, clearly clinically important change was detected by the NEI-VFQ-25 and HUI3. Given that vision is included in HUI3 and that the NEI-VFQ-25 is a vision-targeted measure, this result is not surprising. But vision is included in the QWB-SA, and the overall score did not reflect the gain in HRQL that was measured by the NEI-VFQ-25 and HUI3. Of course, it should be noted that only the ''worst'' symptom for that subject in the QWB-SA is used to compute the overall score and further, that in a relatively elderly cohort, it is likely that many subjects were experiencing symptoms that are more burdensome than impaired vision, and thus, the vision item frequently did not affect the calculation of the QWB-SA score. Others have noted the lack of responsiveness of the EQ-5D to the effects of cataract surgery.[60]

For the heart failure cohort, using published guidelines[23–25,46,47,51] on clinically important differences/changes, the MLHF, HUI3, QWB-SA, and SF-6D recorded clinically important change. Fatigue and shortness of breath symptoms on the QWB-SA as well as the mobility, physical, and social activity scales may have captured some of the effects of heart failure on HRQL. Similarly, the physical functioning, vitality, and role attributes on SF-6D may have registered some of the effects. HUI3 ambulation may have performed similarly.[61]

It should be noted that reliability is less than perfect for each of the measures used in the study,[62] so disagreement between change scores is also influenced by measurement error and short-term fluctuations in health that are unrelated to the conditions of primary interest. Change over time is measured with even less precision than absolute scores at a point in time. Jones and Feeny[63] and Pickard and others[64] found lower levels of agreement between proxy and self-report for change scores than were evident for cross-sectional comparisons of baseline and follow-up scores. Other investigators have pointed

Table 6 Comparisons of Change among Measures of Health-Related Quality of Life from Baseline to 1 Month in Cataract Surgery Cohort

	Got Worse (n = 15)	Stayed Same (n = 62)	Showed Improvement (n = 133)	Total
EQ-5D				
−	4	13	21	38
0	7	39	73	119
+	4	10	39	53
HUI2				
−	8	16	22	46
0	2	20	39	61
+	5	26	72	103
HUI3				
−	8	20	30	58
0	3	14	32	49
+	4	28	71	103
QWB-SA				
−	11	27	34	72
0	2	12	39	53
+	2	23	60	85
SF-6D				
−	10	20	43	73
0	3	23	42	68
+	2	19	48	69
SRH				
−	1	7	27	35
0	7	46	84	137
+	7	9	22	38

Note: EQ-5D = 5-dimension Euro QOL measure; HUI2 = Health Utilities Index Mark 2; HUI3 = Health Utilities Index Mark 3; QWB-SA = Quality of Well-Being–Self-Administered scale; SF-6D = Short-Form 6D; SRH = self-rated health; − = got worse; 0 = stayed same; + = showed improvement; Underlined numbers = agreement.

Table 7 Comparisons of Change among Measures of Health-Related Quality of Life from Baseline to 6 Months in Heart Failure Cohort

	Got Worse (n = 46)	Stayed Same (n = 13)	Showed Improvement (n = 27)	Total
EQ-5D				
−	11	6	16	33
0	16	2	8	26
+	19	5	3	27
HUI2				
−	11	6	13	30
0	8	5	5	18
+	27	2	9	38
HUI3				
−	10	8	14	32
0	10	3	4	17
+	26	2	9	37
QWB-SA				
−	9	6	15	30
0	7	3	5	15
+	30	4	7	41
SF-6D				
−	9	2	12	23
0	8	7	9	24
+	29	4	6	39
SRH				
−	24	3	4	31
0	20	8	13	41
+	2	2	10	14

Note: EQ-5D = 5-dimension Euro QOL measure; HUI2 = Health Utilities Index Mark 2; HUI3 = Health Utilities Index Mark 3; QWB-SA = Quality of Well-Being–Self-Administered scale; SF-6D = Short-Form 6D; SRH = self-rated health; − = got worse; 0 = stayed same; + = showed improvement; Underlined numbers = agreement.

out that due to the size of measurement error typically found in HRQL measures, change must often be quite substantial for measures to agree.[62] Our results were not sensitive to the magnitude of change that was considered clinically important, but the magnitude of true underlying changes does influence the agreement that can be expected. Hence, some measures of change may have better agreement than found in our studies, when reflecting interventions with larger overall effects on HRQL.

Some limitations of the study should be noted. Because the analyses are based on subjects for whom both baseline and the designated follow-up assessments were available, the results are not necessarily representative of the experience of the entire inception cohorts. In the cataract cohort, those with and without complete data were similar. In the heart failure cohort, those with missing data were older than those without missing data. If older patients experienced less improvement in HRQL than younger patients, it is possible that the estimate of change based on subjects for whom we had complete data is biased upwards. As noted in the Results, 7% of subjects had help in completing questionnaires, so responses could have been influenced by others. Another limitation is that while the scoring functions for the QWB-SA and EQ-5D are based on preference scores from random samples of community-dwelling adults in the United States, the scoring functions for HUI2 and HUI3 are based on preferences from random samples of the Canadian population, and the function for the SF-6D is based on United Kingdom preferences. There is evidence of the generalizability of the QWB scoring function,[30,65,66] the HUI2 scoring function,[28,67] and the

HUI3 scoring function.[29,68–70] In contrast, there is considerable variability across ''national'' EQ-5D scoring functions. Nonetheless, having not relied exclusively on US-based scoring functions is unlikely to be an important factor influencing the results. Finally, we classified cataract and heart failure patients as changed if the absolute value of their change score was $\geq$5.0 whether or not the difference was statistically significant. Hays and others[71] note that changes that are statistically significant at the level of an individual subject will typically exceed the guideline for a clinically important difference.[72]

CONCLUSIONS

The results underscore the lack of interchangeability of scores among these 5 widely used preference-based measures. Not only are the absolute scores not necessarily interchangeable, but in these results, the change scores were also not interchangeable.[12] The results also point to a lack of precision in estimating the magnitude of change in HRQL.

In making choices about which preference-based measure(s) to use in a study, investigators need to consider carefully the coverage of the health status classification systems and the relevance of those systems to their clinical or population health application, evidence on the cross-sectional construct validity of the measures in that application, and evidence of the responsiveness (longitudinal construct validity) of the measures in that context. Further, users of the results of studies that have employed preference-based measures to assess HRQL need to interpret those results carefully.

ACKNOWLEDGMENTS

The authors acknowledge the contributions of Steven Tally, University of California, San Diego (UCSD), to the work reported here. The authors also appreciate the help of Barbara Brody, MPH, and Denise Herman, MD, from UCSD; Nancy Sweitzer, MD, PhD, and Neal Barney, MD, from the University of Wisconsin; and Greg Fonerow, MD, and John Bartlett, MD, from the University of California, Los Angeles (UCLA) for their collaboration on subject acquisition.

REFERENCES

1. Rabin R, de Charro F. EQ-5D: a measure of health status from the EuroQol Group. Ann Med. 2001;33(5):337–43.

2. Horsman J, Furlong W, Feeny D, Torrance G. The Health Utilities Index (HUI®): concepts, measurement properties and applications. Health Qual Life Outcomes. 2003;1(1):54.

3. Kaplan RM, Anderson JP. The general health policy model: an integrated approach. In: Spilker B, ed. Quality of Life and Pharmacoeconomics in Clinical Trials. 2nd ed. Philadelphia: Lippincott-Raven Press; 1996. p 309–22.

4. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ. 2002; 21(2):271–92.

5. Brazier JE, Roberts J. The estimation of a preference-based measure of health from the SF-12. Med Care. 2004;42(9):851–9.

6. Fryback DG, Palta M, Cherepanov D, Bolt D, Kim JS. Comparison of 5 health-related quality-of-life indexes using item response theory analysis. Med Decis Making. 2010;30(1):5–15.

7. Cherepanov D, Palta M, Fryback DG. Underlying dimensions of the five health-related quality-of-life measures used in utility assessment: evidence from the National Health Measurement Study. Med Care. 2010;48(8):718–25.

8. Feeny DH. Preference-based measures: utility and quality-adjusted life years. In: Fayers P, Hays R, eds. Assessing Quality of Life in Clinical Trials. 2nd ed. Oxford: Oxford University Press; 2005. p 405–29.

9. Marra CA, Esdaile JM, Guh D, et al. A comparison of four indirect methods of assessing utility values in rheumatoid arthritis. Med Care. 2004;42(11):1125–31.

10. Marra CA, Marion SA, Guh DP, et al. Not all ''quality-adjusted life years'' are equal. J Clin Epidemiol. 2007;60(6):616–24.

11. Marra CA, Woolcott JC, Kopec JA, et al. A comparison of generic, indirect utility measures (the HUI2, HUI3, SF-6D, and the EQ-5D) and disease-specific instruments (the RAQoL and the HAQ) in rheumatoid arthritis. Soc Sci Med. 2005;60(7): 1571–82.

12. Feeny DH, Wu L, Eng K. Comparing Short Form 6D, standard gamble, and Health Utilities Index Mark 2 and Mark 3 utility scores: results from total hip arthroplasty patients. Qual Life Res. 2004;13(10):1659–70.

13. Luo N, Johnson JA, Shaw JW, Feeny D, Coons SJ. Self-reported health status of the general adult U.S. population as assessed by the EQ-5D and Health Utilities Index. Med Care. 2005;43(11):1078–86.

14. Fryback DG, Dunham NC, Palta M, et al. U.S. norms for six generic health-related quality of life indexes from the National Health Measurement Study. Med Care. 2007;45(12):1162–70.

15. Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PM. On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. Qual Life Res. 2003;12(4):349–62.

16. Marra CA, Rashidi AA, Guh D, et al. Are indirect utility measures reliable and responsive in rheumatoid arthritis patients? Qual Life Res. 2005;14(5):1333–44.

17. Hatoum HT, Brazier JE, Akhras KS. Comparison of the HUI3 with the SF-36 preference based SF-6D in a clinical trial setting. Value Health. 2004;7(5):602–9.

18. McDonough CM, Tosteson TD, Tosteson AN, Jette AM, Grove MR, Weinstein JN. A longitudinal comparison of 5 preference-weighted health state classification systems in persons with intervertebral disk herniation. Med Decis Making. 2011;31(2):270–80.

19. Kaplan RM, Tally S, Hays RD, et al. Five preference-based indexes in cataract and heart-failure patients were not equally responsive to change. J Clin Epidemiol. 2011;64(5):497–506.

20. Mangione CM, Lee PP, Gutierrez PR, Spritzer K, Berry S, Hays RD. Development of the 25-item National Eye Institute Visual Function Questionnaire. Arch Ophthalmol. 2001;119(7):1050–8.

21. Varma R, Wu J, Chong K, Azen SP, Hays RD. Impact of severity and bilaterality of visual impairment on health-related quality of life. Ophthalmology. 2006;113(10):1846–53.

22. McDonnell PJ, Mangione C, Lee P, et al. Responsiveness of the National Eye Institute Refractive Error Quality of Life instrument to surgical correction of refractive error. Ophthalmology. 2003;110(12):2302–9.

23. Rector TS. A conceptual model of quality of life in relation to heart failure. J Card Fail. 2005;11(3):173–6.

24. Rector TS, Kubo SH, Cohn JN. Patients' self-assessment of their congestive heart failure: part II. Content, reliability and validity of a new measure, The Minnesota Living with Heart Failure Questionnaire. Heart Fail. 1987;(3):198–209.

25. Rector TS, Francis GS, Cohn JN. Patients' self-assessment of their congestive heart failure: part 1. Patient perceived dysfunction and its poor correlation with maximal exercise tests. Heart Fail. 1987;(3):192–6.

26. Garin O, Ferrer M, Pont A, et al. Disease-specific health-related quality of life questionnaires for heart failure: a systematic review with meta-analyses. Qual Life Res. 2009;18(1):71–85.

27. Shaw JW, Johnson JA, Coons SJ. US valuation of the EQ-5D health states: development and testing of the D1 valuation model. Med Care. 2005;43(3):203–20.

28. Torrance GW, Feeny DH, Furlong WJ, Barr RD, Zhang Y, Wang Q. Multi-attribute utility function for a comprehensive health status classification system: Health Utilities Index Mark 2. Med Care. 1996;34(7):702–22.

29. Feeny DH, 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(2):113–28.

30. Kaplan RM, Bush JW, Berry CC. Health status: types of validity and the index of well-being. Health Serv Res. 1976;11(4):478–507.

31. Kaplan RM, Frosch DL. Decision making in medicine and health care. Annu Rev Clin Psychol. 2005;1:525–56.

32. Kaplan RM, Anderson JP, Patterson TL, et al. Validity of the Quality of Well-Being Scale for persons with human immunodeficiency virus infection. HNRC Group. HIV Neurobehavioral Research Center. Psychosom Med. 1995;57(2):138–47.

33. Idler EL, Benyamini Y. Self-rated health and mortality: a review of twenty-seven community studies. J Health Soc Behav. 1997;38(1):21–37.

34. Ware J Jr., Kosinski M, Keller SD. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. Med Care. 1996;34(3):220–33.

35. Revicki D, Rentz AM, Kowalski JW, Chen WH. Assessment of unidimensionality for the Visual Function Questionnaire-Utility Index (VFQ-UI) items in patients with central vision loss. Qual Life Res. 2010;19(Suppl 1):49–50.

36. Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR. Methods to explain the clinical significance of health status measures. Mayo Clin Proc. 2002;77(4):371–83.

37. Harrison MJ, Davies LM, Bansback NJ, et al. The comparative responsiveness of the EQ-5D and SF-6D to change in patients with inflammatory arthritis. Qual Life Res. 2009;18(9):1195–205.

38. Revicki DA, Osoba D, Fairclough D, et al. Recommendations on health-related quality of life research to support labeling and promotional claims in the United States. Qual Life Res. 2000; 9(8):887–900.

39. Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol. 2008; 61(2):102–9.

40. Hays RD, Farivar SS, Liu H. Approaches and recommendations for estimating minimally important differences for health-related quality of life measures. COPD. 2005;2(1):63–7.

41. Leidy NK, Wyrwich KW. Bridging the gap: using triangulation methodology to estimate minimal clinically important differences (MCIDs). COPD. 2005;2(1):157–65.

42. Terwee CB, Roorda LD, Dekker J, et al. Mind the MIC: large variation among populations and methods. J Clin Epidemiol. 2010;63(5):524–34.

43. Beaton DE, van Eerd D, Smith P, et al. Minimal change is sensitive, less specific to recovery: a diagnostic testing approach to interpretability. J Clin Epidemiol. 2011;64(5):487–96.

44. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.

45. Drummond M. Introducing economic and quality of life measurements into clinical studies. Ann Med. 2001;33(5):344–9.

46. Walters SJ, Brazier JE. What is the relationship between the minimally important difference and health state utility values? The case of the SF-6D. Health Qual Life Outcomes. 2003;1:4.

47. Walters SJ, Brazier JE. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. Qual Life Res. 2005;14(6):1523–32.

48. Pickard AS, Neary MP, Cella D. Estimation of minimally important differences in EQ-5D utility and VAS scores in cancer. Health Qual Life Outcomes. 2007;5:70.

49. Majumdar SR, Johnson JA, Bowker SL, et al. A Canadian consensus for the standardized evaluation of quality improvement interventions in type 2 diabetes. Can J Diabetes. 2005;29(3):220–9.

50. Samsa G, Edelman D, Rothman ML, Williams GR, Lipscomb J, Matchar D. Determining clinically important differences in health status measures: a general approach with illustration to the Health Utilities Index Mark II. Pharmacoeconomics. 1999;15(2):141–55.

51. Kaplan RM. The minimally clinically important difference in generic utility-based measures. COPD. 2005;2(1):91–7.

52. 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(3):290–9.

53. Groessl EJ, Kaplan RM, Barrett-Connor E, Ganiats TG. Body mass index and quality of well-being in a community of older adults. Am J Prev Med. 2004;26(2):126–9.

54. Kontodimopoulos N, Pappa E, Papadopoulos AA, Tountas Y, Niakas D. Comparing SF-6D and EQ-5D utilities across groups differing in health status. Qual Life Res. 2009;18(1):87–97.

55. Sullivan PW, Lawrence WF, Ghushchyan V. A national catalog of preference-based scores for chronic conditions in the United States. Med Care. 2005;43(7):736–49.

56. Khanna D, Furst DE, Wong WK, et al. Reliability, validity, and minimally important differences of the SF-6D in systemic sclerosis. Qual Life Res. 2007;16(6):1083–92.

57. Globe DR, Wu J, Azen SP, Varma R. The impact of visual impairment on self-reported visual functioning in Latinos: the Los Angeles Latino Eye Study. Ophthalmology. 2004;111(6): 1141–9.

58. Andres AM, Marzo PF. Chance-corrected measures of reliability and validity in K x K tables. Stat Methods Med Res. 2005;14(5):473–92.

59. Altman DG. Practical Statistics for Medical Research. London: Chapman & Hall; 1991.

60. Browne JP, van der Meulen JH, Lewsey JD, Lamping DL, Black N. Mathematical coupling may account for the association between baseline severity and minimally important difference values. J Clin Epidemiol. 2010;63(8):865–74.

61. Pressler SJ, Eckert GJ, Morrison GC, Murray MD, Oldridge NB. Evaluation of the Health Utilities Index Mark-3 in heart failure. J Card Fail. 2011;17(2):143–50.

62. Palta M, Chen HY, Kaplan RM, Feeny D, Cherepanov D, Fryback DG. Standard error of measurement of five health utility indexes across the range of health for use in estimating reliability and responsiveness. Med Decis Making. 2011;31(2):260–9.

63. Jones CA, Feeny DH. Agreement between patient and proxy responses of health-related quality of life after hip fracture. J Am Geriatr Soc. 2005;53(7):1227–33.

64. Pickard AS, Johnson JA, Feeny DH, Shuaib A, Carriere KC, Nasser AM. Agreement between patient and proxy assessments of health-related quality of life after stroke using the EQ-5D and Health Utilities Index. Stroke. 2004;35(2):607–12.

65. Balaban DJ, Sagi PC, Goldfarb NI, Nettler S. Weights for scoring the quality of well-being instrument among rheumatoid arthritics: a comparison to general population weights. Med Care. 1986;24(11):973–80.

66. Hector RD Sr., Anderson JP, Paul RC, Weiss RE, Hays RD, Kaplan RM. Health state preferences are equivalent in the United States and Trinidad and Tobago. Qual Life Res. 2010;19(5):729–38.

67. Wang Q, Furlong W, Feeny D, Torrance G, Barr R. How robust is the Health Utilities Index Mark 2 utility function? Med Decis Making. 2002;22(4):350–8.

68. Le Gales C, Buron C, Costet N, Rosman S, Slama PR. Development of a preference-weighted health status classification system in France: the Health Utilities Index 3. Health Care Manag Sci. 2002;5(1):41–51.

69. Raat H, Bonsel GJ, Hoogeveen WC, Essink-Bot ML. Feasibility and reliability of a mailed questionnaire to obtain visual analogue scale valuations for health states defined by the Health Utilities Index Mark 3. Med Care. 2004;42(1):13–8.

70. Ruiz M, Rejas J, Soto J, Pardo A, Rebollo I. [Adaptation and validation of the Health Utilities Index Mark 3 into Spanish and correction norms for Spanish population]. Med Clin (Barc). 2003;120(3):89–96.

71. Hays RD, Brodsky M, Johnston MF, Spritzer KL, Hui KK. Evaluating the statistical significance of health-related quality-of-life change in individual patients. Eval Health Prof. 2005; 28(2):160–71.

72. McLeod LD, Coon CD, Martin SA, Fehnel SE, Hays RD. Interpreting patient-reported outcome results: US FDA guidance and emerging methods. Expert Rev Pharmacoecon Outcomes Res. 2011;11(2):163–9.