Comparison of Responses to SF-36 Health Survey Questions with One-Week and Four-Week Recall Periods

Susan D. Keller, Martha S. Bayliss, John E. Ware, Jr., Ming-Ann Hsu, Anne M. Damiano, and Thomas F. Goss

Objective. To compare the measurement properties of acute (one-week recall) and standard (four-week recall) versions of SF-36 Health Survey (SF-36) scale scores.
Data Sources. SF-36 data collected from 142 participants (60% female, average age 39) in a clinical trial of an asthma medication: 74 patients randomized to the acute form and 68 to the standard.
Data Collection. The SF-36 was self-administered at the time of a clinic visit (before clinical examination) to synchronize with clinical measures of disease severity at three different time points during the clinical trial: −2 weeks (two weeks before randomization to treatment), baseline (week 0 or randomization), and +4 weeks (four weeks after baseline).
Principal Findings. The acute form yielded high-quality data; scales conformed to the assumptions of the summated ratings method used to score the standard SF-36; and scales had good distributional properties, were reliable, and had a factor content similar to the standard. The data indicated that while the acute form was more sensitive than the standard to change in health status associated with changes in acute symptoms, acute scale scores may not be comparable to national norms based on the standard, particularly for those scales that assess frequency of health events during a specified time period.
Conclusions. Results support the use of the acute form in its intended applications; however, further research is required to document the generalizability of greater sensitivity of the acute form to recent changes in health and to explore whether norms based on the standard can be used to interpret the acute scale scores.

Key Words. SF-36 Health Survey, recall period, asthma, reliability, validity

General health status surveys differ in the time frame (recall period) respondents are asked to consider when answering questions about their health

(Wilkin, Hallam, and Doggett 1992). Commonly used time frames instruct respondents to focus on *today* (e.g., the Sickness Impact Profile: see Bergner, Bobbitt, Kressel, et al. 1976); the *past week*, that is, during the week just previous to the interview (e.g., the Quality of Life Index: see Spitzer, Dobson, and Hall 1981); the *past four weeks* (e.g., the SF-36 Health Survey: see Ware and Sherbourne 1992); and the *past year* (e.g., questions from the National Health Interview Survey: see Patrick and Erickson 1993). Several considerations influence the choice of time frames. The time frame can follow the developers' definition of health. For example, the Sickness Impact Profile defines health as "a condition defined or perceived by the individual against which he evaluates his own behavior on the day of the interview" (Bergner, Bobbitt, Kressel, et al. 1976). The expected frequency of health events may also influence the choice of recall period. A period of up to a year may be preferred in monitoring relatively rare events like hospitalizations (Patrick and Erickson 1993), a stroke, or other complication. A recall period as short as one day will suffice for assessing daily events. Most measures in the Medical Outcomes Study (MOS) reference the previous four weeks in order to capture a more stable sample of recent health, not unduly affected by daily or momentary fluctuations (Fowler 1984; Stewart and Ware 1992). The four-week recall period was adopted for the six SF-36 Health Survey (SF-36) scales with an explicit time frame to maintain comparability with long-form MOS measures (Ware and Sherbourne 1992; Ware et al. 1993).

In theory, shorter recall periods should be more sensitive than longer recall periods to recent changes in health status; this was the rationale behind

---

The authors gratefully acknowledge Schering-Plough Corporation for funding the fielding of this study and the preparation of this manuscript. The preparation of this manuscript was also supported by the International Quality of Life Assessment (IQOLA) Project funding by Schering-Plough Corporation and Glaxo Wellcome, Inc.

Susan D. Keller, Ph.D. is Research Scientist, The Health Institute, New England Medical Center Hospitals, Boston; and Martha S. Bayliss, M.Sc. is Senior Project Director at The Health Institute. John E. Ware, Jr., Ph.D. is Senior Scientist and Director of Healthcare Assessment Laboratory; Research Professor of Psychiatry, Tufts University School of Medicine; and Adjunct Professor of Health and Social Behavior, Harvard University School of Public Health. Ming-Ann Hsu, M.P.H. is Senior Manager, Outcomes Research Division, Pfizer, Inc. Anne M. Damiano, Sc.D. and Thomas F. Goss, Pharm.D. are Principals at Covance Health Economics and Outcomes Services, Inc., Washington, DC. Address correspondence and requests for reprints to San Keller, Ph.D., New England Medical Center, The Health Institute, 750 Washington Street, Box 345, Boston, MA 02111, Fax: 617-636-8077, e-mail: san.keller@es.nemc.org. This article, submitted to *Health Services Research* on May 24, 1995, was revised and accepted for publication on September 18, 1996.

a one-week recall version (the "acute" form) of the SF-36 (Ware et al. 1993). Support for this hypothesis and the magnitude of any increase in sensitivity with a shorter recall period has not been demonstrated. Other important questions have also gone unanswered. Does changing the recall period, for instance, affect the psychometric properties of the questionnaire? Can data collected with the acute form be compared with normative data on the standard? Is the factor content of a scale and its interpretation robust across time frames?

Searches of MEDLINE (1976–present), PSYCHLIT (1974–present) and ERIC (1966–present), using the key words *health status, questionnaires, recall period, time period,* and *time frame,* yielded only a handful of articles that discussed the relationship of recall period to health survey results or to questionnaire scores in general. Evidence regarding whether a shorter recall period strengthens the relationship of health status scores to recent disease status has not been reported. This article reports studies of two equivalent samples of asthmatics designed to test whether responses to questions about health differ depending on whether they refer to a four-week recall period or a one-week period. Specifically, the following null hypotheses were tested with regard to the SF-36:

Hypothesis 1. The recall period will not affect whether scales conform to assumptions underlying their scoring and scaling.

Hypothesis 2. The physical and mental health constructs underlying the standard form will be replicated in the acute form.

Hypothesis 3. Mean scores for the SF-36 will be unaffected by the recall period.

Hypothesis 4. Usefulness of the SF-36 scales in detecting change in disease severity will be the same for the acute and the standard versions.

METHODS

SAMPLE

One hundred forty-two patients, with a documented history of asthma of at least six months duration, participated in a placebo-controlled, double-blind study to test the effect of an inhaled corticosteroid on the health-related quality of life of asthmatics (see Table 1). Except for their asthma, participants were required to be in good health and free from any clinically significant disease.

Table 1: Definition of Clinical Criterion Variables

Clinical Status at a Point in Time	
FEV_1%*	Measured at baseline and week 4 visits after sufficient medication washout time.† Operationalized as % of predicted normal for each patient.
Patient diary data	Report each day of the severity of all asthma symptoms‡ aggregated to average severity during the past week.
Change in Clinical Status During One Week§	
Generic health transition item	Patient evaluation of ***health in general*** now as compared to one week ago recorded at each visit (categorized as improved, stayed the same, or worsened during the past week).
Asthma-specific health transition item	Patient evaluation of ***overall asthma condition*** now as compared to one week ago recorded at each visit (categorized as improved, stayed the same, or worsened during the past week).
Patient diary data	Operationalized as the transition in daily diary scores during the seven days before health-related quality of life (HQL) assessment. Patient daily diary scores for the day before HQL assessment are subtracted from patient daily diary scores on the seventh day before HQL assessment. Daily diary scores are then categorized as improved, stayed the same, or worsened during the past week.

*A weekly average for FEV_1% was not available, so disease severity for this criterion was estimated from a measurement on the day of the week 4 visit.

† Washout times: Proventil®, 8 hours; oral beta-agonists, 12 hours; long-acting beta-agonists, 24 hours.

‡ Response choices were "none," "trivial or doubtful," "mild; clearly present, but causing little or no discomfort," "moderate; annoying, but not causing marked discomfort," "moderately severe; causing marked discomfort," "severe; some interference with sleep or activities, but not incapacitating," or "incapacitating."

§ FEV_1% and patient and physician ratings of disease severity were not assessed at weekly intervals and so could not be used in the analysis of clinical change.

Patients' average age was 39.4 years (s.d. = 15.0 years) with a range of age from 14 to 70. Approximately 60 percent of the sample was female. At baseline, the average pulmonary function, as measured by percent of normal forced expiratory volume (FEV_1%), was 80.1 (s.d. = 16.7) with a range of FEV_1% percent of normal of 44 to 118 for the entire sample. Baseline general health ratings (respondents rated their health on a scale of excellent to poor) for this sample (74.7) were comparable to those for age- and gender-adjusted general population norms (72.2).

MEASURES

Health Status

Health status was assessed using the SF-36, which yields eight multi-item scales measuring Physical Functioning (PF), role limitations due to physical health problems (Role Physical: RP), Bodily Pain (BP), General Health Perceptions (GH), Vitality (VT), Social Functioning (SF), role limitations due to emotional health problems (Role Emotional: RE), Mental Health (MH), a single-item evaluation of change in health (Health Transition) (Ware and Sherbourne 1992; Ware et al. 1993), and physical and mental summary component scales (PCS and MCS) (Ware, Kosinski, and Keller 1994). The conceptual development, reliability, validity, and other information about the SF-36 are documented in two user's manuals (Ware et al. 1993; Ware, Kosinski, and Keller 1994) and in over 200 publications (*Annotated Bibliography for the SF-36 Health Survey* 1996).

The acute form of the SF-36 was designed for applications in which health status would be measured weekly or biweekly (Ware et al. 1993). It was created by changing the recall period for six of the scales (RP, BP, VT, SF, RE, and MH) from "the past four weeks" to "the past week." For example, the standard instructions, "During the *past four weeks*, how much did *pain* interfere with your normal work . . . ?" were changed to: "During the *past week*, how much did *pain* interfere with your normal work . . . ?" Two SF-36 scales (PF and GH) do not have a recall period and so are identical across acute and standard forms. The time frame for the Health Transition item, which is not used in scoring any of the eight scales, was changed from "one year ago" to "one week ago."

Clinical Variables

Clinical variables included patient, physician, and objective (FEV_1%) measurements of clinical status. Substantial convergence was observed among these indicators; inter-correlations ranged from 0.67 to 0.71 for the patient and physician assessments of symptoms. Correlations between FEV_1% and the symptom severity indicators were notably lower, ranging from 0.13 to 0.20, indicating that FEV_1% and symptom assessments provide qualitatively different information about a patient's experience of asthma.

PROCEDURE

All patients received the study medication at enrollment, four weeks before randomization (−4 weeks). After four weeks, at baseline (week 0), half of the

patients were randomized to placebo for the next eight weeks. The effect of treatment on health scores is ignored here. A health status questionnaire for use in patients with asthma contained the SF-36 (either acute or standard forms) as the first module of questions. This questionnaire was self-administered at the time of a clinic visit (before clinical examination) to synchronize with clinical measures of disease severity.

Out of 142 patients in the clinical trial, 74 were randomized to the acute form and 68 to the standard form. The two forms were self-administered at three different time points during the clinical trial: −2 weeks (two weeks before randomization to treatment), baseline (week 0 or randomization), and +4 weeks (four weeks after baseline).

STATISTICAL ANALYSES

Data were analyzed to test four null hypotheses presented earlier. Data quality, results of tests of scaling and scoring assumptions, construct validity, and clinical validity were evaluated and compared across forms and with published findings for the standard form (Ware et al. 1993; McHorney, Kosinski, and Ware 1994). Baseline data (week 0) were used to test data quality and scaling and scoring assumptions, and these tests were replicated at weeks −2 and +4. The effect of recall period on mean scale scores was tested at week −2, baseline, and week 4. The analyses of change over time were conducted between baseline and week 4.

Data Quality

Data quality was evaluated in two ways. First, the percentage of patients who completed all items within each scale and of those who had computable HQL scales (i.e., completed at least half of the items) were estimated, and are shown further on, in Table 4. Second, the Response Consistency Index (RCI) was calculated (Ware et al. 1993). The RCI is a count of the frequency of inconsistent responses across 15 pairs of SF-36 items. For example, a report of being able to walk a mile, but not a block, without limitation is considered an inconsistent response. These analyses were conducted to answer the question of whether the different recall period of the acute form would affect the number of missing or inconsistent responses.

Tests of Scoring and Scaling Assumptions

To test hypothesis 1, the two forms were evaluated and compared according to four major psychometric criteria underlying the construction, scoring,

and interpretation of scales (McHorney et al. 1994): (1) tests of assumptions underlying a summated ratings scale, (2) tests of item-discriminant validity underlying scale groupings of items, (3) scale score reliability, and (4) features of score distributions. These analyses were conducted to see if the scoring algorithms used for the standard SF-36 form were appropriate for the acute form.

The method of summated ratings assumes that items in the same scale can be aggregated without score standardization or item weighting (Likert 1932). To avoid standardization, items should have roughly equivalent means and standard deviations. To avoid weighting, items should be equally representative of (that is, have roughly equivalent relationships to) the underlying scale dimension. Items should also correlate (greater than 0.40: Helmstadter [1964] corrected for overlap: Howard and Forehand [1962]) with their hypothesized scales.

Item-discriminant validity is supported when the correlation between each item and its hypothesized scale is larger than its correlation with competing scales. (Differences between correlations of two standard errors were considered significant.) Tests of item-discriminant validity were summarized into item-scaling success rates that indicate the percentage of successful tests.

Scale level analyses included evaluation of scale score distributions for the percentage of people achieving either the highest score (ceiling effect) or the lowest score (floor effect) as well as assessment of scale reliability. The internal consistency reliability of each scale score was estimated using Cronbach's Alpha (Cronbach 1951), and results were compared with those published for the standard SF-36 (Ware, Snow, Kosinski, et al. 1993).

Construct Validity

Evidence of construct validity was obtained on the basis of item convergent and discriminant validity tests to support the grouping of the SF-36 items into the eight scales that refer to eight health constructs, testing hypothesis 2. In addition, principal components analysis was used to test for consistency in the second-order factor structure across the two forms. Based on previous work (McHorney, Ware, and Raczek 1993; Ware, Kosinski, Bayliss, et al. 1995), physical and mental higher-order factors were predicted to explain the great majority of the covariance between SF-36 scale scores. Two components were extracted from the correlations among SF-36 scale scores and were rotated to orthogonal simple structure using the varimax method (Comrey and Lee 1992). The pattern of scale-factor correlations for acute and standard

forms was compared with patterns of previous studies (McHorney, Ware, and Raczek 1993; Ware, Kosinski, and Keller 1994; Ware, Kosinski, Bayliss, et al. 1995). These analyses were conducted to determine whether the eight scales in the acute form had interpretations similar to those in the standard. These analyses answer questions such as: "Is the acute social functioning scale primarily an indicator of mental health as it is in the standard form?" If the results of these analyses are similar to those of the standard form, the standard mental and physical component summary measures used for scoring algorithms may be used to score data from the acute form as well.

Recall Period Effects on Group Means

Tests of hypothesis 3 addressed whether average scores differed as a function of recall period. Repeated-measures multivariate analysis of variance (MANOVA) was used to test for differences in mean scale scores due to the time frame across three time periods (weeks −2, 0, and 4). Null results for recall period effects would support the use of norms based on the standard form to interpret acute scores.

Sensitivity in Relation to Clinical Criteria

Change in severity was defined by categorizing all patients as improved, stable, or worsened during the past week according to each of three criteria (see Table 1). To test hypothesis 4, the logic of "known groups" validity (Kerlinger 1973) was used to assess the relationship of SF-36 scale change scores to clinical variables by comparing SF-36 change scores across groups known to differ in change in clinical status. SF-36 scores were available for clinical change between baseline and four weeks only; thus, SF-36 change scores were calculated by subtracting baseline scores from four-week scores. Clinical change was defined as patients' perceptions of their change in general health and asthma condition over the past week and change in daily diary scores over the past week. Thus, four-week SF-36 change scores for acute and standard scales were compared for sensitivity to change in condition over the past week. Unfortunately, it was not possible to compare the sensitivity of these change scores to change in condition over the past four weeks (because patient perceptions of their change in condition over the past four weeks were not assessed). MANOVA models were fit to the data initially to test for overall effects, followed by univariate models. No corrections were made for multiple comparisons in the analysis because of the conservative nature of these tests in small groups.

RESULTS

Equivalence of Groups Compared

The equivalence of the groups that completed standard and acute forms was established by comparing age, gender composition, height, weight, $FEV_1\%$, and general health rating. No significant differences between the groups were detected for any of these variables, and no noteworthy trends were apparent.

Data Quality

Rates of complete items and computable scales were uniformly high for all scales, and no significant differences in these rates were found between groups administered standard versus acute forms. While the response consistency for the standard form in this sample was comparable to that of the U.S. general population (91.2 percent and 90.3 percent, respectively), the response consistency for the acute SF-36 was lower (86.5 percent) *for the baseline administration only.* Most of these inconsistent responses occurred for the GH scale, and for the MH scale positive well-being items. Three patients were responsible for over 50 percent of the inconsistent responses to the baseline acute form.

Tests of Scaling and Scoring Assumptions

Results of analyses supported the first hypothesis: *Recall period will not affect whether scales conform to assumptions underlying their scoring and scaling.* These analyses were replicated for −2 week and +4 week time periods to check for consistency.

Item means, standard deviations, and correlations with scale scores were comparable across forms. Scaling success rates were high across all scales for both standard and acute forms, supporting the grouping of items into the eight scales. However, the RE acute form did consistently exhibit (at −2 weeks, baseline, and week 4) lower rates of scaling success relative to the standard form. Low item-scale correlations for RE items were due to lack of variability in responses, with most patients reporting no limitations.

With few exceptions, floor and ceiling effects for this sample did not differ across forms or from those observed for the standard form in the U.S. general population (see Table 2). The lower percentage of persons scoring at the ceiling of the PF scale for both forms (12 percent standard and 11 percent acute) than in the U.S. general population (35 percent) is consistent with the clinical picture of asthma as limiting the performance of strenuous

Table 2: Descriptive Statistics, Tests of Scaling Assumptions, and Reliability Estimates for Standard (St.) and Acute (Ac.) SF-36 Scales

	PF		*RP*		*BP*		*GH*		*VT*		*SF*		*RE*		*MH*	
	St.	*Ac.*	*St.*	*Ac.*	*St.*	*Ac.*	*St.*	*Ac.*	*St.*	*Ac.*	*St.*	*Ac.*	*St.*	*Ac.*	*St.*	*Ac.*
Mean	8.26	77.8	74.6	80.7	78.6	78.6	65.5	66.1	61.5	60.3	85.7	86.8	84.8	89.2	79.0	77.9
25th percentile	70	65	50	75	62	62	55	52	47.5	50	75	75	100	100	76	72
50th percentile	90	85	100	100	84	84	70	72	65	62.5	100	100	100	100	82	80
75th percentile	95	95	100	100	92	100	82	82	75	75	100	100	100	100	88	88
Std. deviation	14.2	19.6	37.3	32.2	17.3	22.3	21.1	20.1	16.4	17.8	19.2	20.3	30.2	22.8	14.4	12.8
Skewness	−0.7	−0.8	−1.1	−1.4	−0.5	−0.8	−0.8	−0.6	−0.3	−0.4	−1.4	−1.7	−1.8	−2.3	−1.3	−0/8
Range	45–100	35–100	0–100	0–100	32–100	12–100	5–100	17–95	25–90	20–95	25–100	12–100	0–100	0–100	32–100	40–100
% Ceiling	11.8	10.8	61.8	67.6	25.0	39.2	3.0	0	0	0	51.5	58.1	76.5	77.0	2.9	1.4
% Floor	0	0	13.2	6.8	0	0	0	0	0	0	0	0	5.9	2.7	0	0
% Complete items	99	96	99	100	100	100	99	99	100	100	100	100	100	100	99	97
% Computable items	100	100	100	100	100	100	99	99	100	100	100	100	100	100	100	100
% Scaling success (1 + 2)†	95.0	93.8	100	100	100	100	100	100	100	100	100	100	100	70.8	100	75.0
Range of Item Internal Consistency Correlations	.00–.75	.25–.78	.71–.79	.56–.75	.67	.78	.45–.74	.47–.75	.52–.74	.56–.63	.54	.76	.47–.76	.32–.56	.56–.70	.35–.51
Median Item Internal Consistency Correlation	.52	.71	.73	.71	.67	.78	.51	.60	.59	.61	.54	.76	.70	.43	.65	.37
Reliability (r_{kk})	.81	.89	.88	.84	.75	.85	.79	.81	.80	.79	.70	.86	.79	.59*	.83	.64*

* α Coefficients are not equal, $p < .05$.

Source: Multitrait Analysis Program results: acute $n = 73$; standard $n = 67$.

activities. Greater ceiling effects were found for the acute version of the BP scale (39.2 percent) compared with the standard version (25.0 percent) and the U.S. general population (23 percent): that is, bodily pain was reported less frequently over the past week than over the past month. Fewer floor effects were found for the acute RP and RE scales (6.8 percent and 2.7 percent, respectively) than for the standard scales (13.2 percent and 5.9 percent) in this sample or in the U.S. general population (14.1 percent and 6.3 percent). This suggests that people are less likely to experience role disability within one week than within one month.

Reliability of Scale Scores

Internal consistency reliability coefficients were satisfactory for group comparisons (well above 0.70) and did not differ between forms for six out of eight scales (see Table 4). Internal consistency reliability was significantly lower for the one-week than for the four-week versions of the RE (0.59 versus 0.79) and MH scales (0.64 versus 0.83). Lower internal consistency of some RE and MH items accounted for these results. Analyses conducted at week −2 and week 4 showed that while the reliability of the acute MH scale was higher at those two time points (0.78 and 0.77, respectively), the reliability of the acute RE scale was consistently lower (0.63 and 0.61, respectively) than that for the standard version.

Construct Validity

In support of hypothesis 2: *The physical and mental health constructs underlying the standard form will be replicated for the acute form* (Ware, Kosinski, and Keller 1994; Ware, Kosinski, Bayliss, et al. 1995), principal components analyses

Table 3: Correlations Between Scales and Rotated Physical and Mental Health Components

	Physical Health Component		*Mental Health Component*	
Scales	*Acute*	*Standard*	*Acute*	*Standard*
PF	0.78	0.75	0.26	0.12
RP	0.79	0.78	0.35	0.26
BP	0.60	0.58	0.27	0.21
GH	0.64	0.74	0.11	0.05
VT	0.48	0.45	0.67	0.46
SF	0.55	0.37	0.70	0.53
RE	0.27	0.05	0.79	0.76
MH	0.24	0.00	0.83	0.85

confirmed a two-factor higher-order structure of both forms (see Table 3). The components were interpreted as physical and mental health based on correlations with SF-36 scales (i.e., PF loaded highest on the "physical" component and MH loaded highest on the "mental" component). Further, the magnitude and pattern of scale-to-component correlations in the sample replicated results for the U.S. general population, with one exception. The correlation between the GH scale and the mental component was significantly lower ($p < .05$) for both acute ($r = 0.11$) and standard forms ($r = 0.05$) in the sample than in the U.S. general population ($r = 0.37$).

Scale Means and Normative Comparisons

Hypothesis 3, *Mean scale scores will be unaffected by recall period*, could not be rejected (see Table 4). However, because some differences approached significance and the confidence intervals were large, we are cautious about accepting the null hypothesis. A repeated measures MANOVA indicated that means did not differ across repeated administrations but the effect of form (standard versus acute) approached significance ($p = .08$). Univariate tests for scores indicated that differences across forms were largest for the role and social functioning scales (RP, RE, and SF). Compared to the standard scale mean scores, mean scores for the acute form averaged nearly five points higher (more favorable) for the RP scale, nearly seven points higher for the RE scale, and nearly three points higher on the SF scale. The difference in means between the two forms was significant at a conventional level ($p = .05$) for the RE scale (without adjustment for multiple comparisons).

Table 4: Acute versus Standard SF-36 Scale Scores, Average over Three Administrations (Week −2, Baseline, and Week 4)

			Differences Between Forms		
	Mean: Acute Form	*Mean: Standard Form*	*Mean: Difference*	*95% C.I.*	p-*Value*
PF	79.05	80.82	1.76	−1.62 - +5.15	.31
RP	82.26	77.40	−4.86	−11.17 - +1.45	.13
BP	79.09	77.96	−1.13	−4.87 - +2.61	.55
GH	66.85	66.04	−0.82	−4.88 - +3.25	.69
VT	60.48	60.78	0.30	−3.13 - +3.73	.86
SF	88.27	85.42	−2.86	−6.47 - +0.76	.12
RE	89.52	82.66	−6.86	−12.12 - −1.60	.01
MH	78.34	77.48	−0.86	−3.65 - +1.92	.54

Note: MANOVA F-statistic for difference between acute and standard form: $F(8,406) = 1.76$, $p < .0830$.

Table 5: Validity of Acute versus Standard SF-36 Scale Change Scores in Detecting Change in Clinical Status

HQL Concept: 4-Week change In	*Source: Generic HT Item*		*Source: Asthma-Specific HT Item*		*Source: Patient Diary*	
	F-*statistic*	p-*Value*	F-*statistic*	p-*Value*	F-*statistic*	p-*Value*
PF-Standard	1.85	.17	1.71	.19	0.70	.50
PF-Acute	2.37	.10	8.05	.00	8.63	.00
RP-Standard	0.85	.43	0.53	.59	1.21	.31
RP-Acute	14.04	.00	19.86	.00	4.63	.01
BP-Standard	0.76	.47	0.92	.40	0.85	.43
BP-Acute	8.04	.00	10.48	.00	2.44	.10
GH-Standard	0.18	.84	1.74	.19	0.91	.41
GH-Acute	2.08	.13	3.66	.03	0.18	.84
VT-Standard	0.71	.50	1.39	.26	3.00	.06
VT-Acute	4.37	.02	8.82	.00	2.04	.14
SF-Standard	0.32	.73	0.08	.92	1.05	.35
SF-Acute	1.65	.20	3.49	.04	2.16	.12
RE-Standard	1.11	.34	1.11	.34	1.95	.15
RE-Acute	1.30	.28	1.32	.27	1.57	.22
MH-Standard	5.23	.01	0.40	.67	0.63	.54
MH-Acute	1.68	.19	3.64	.03	3.75	.03
MANOVA F for 8 **standard** scales and individual criteria	1.37	.19	1.10	.38	1.07	.41
MANOVA F for 8 **acute** scales and individual criteria	2.99	.00	4.02	.00	1.91	.04

Validity in Relation to Clinical Criteria

We rejected the fourth hypothesis: *The usefulness of the SF-36 scales in detecting the impact of change in disease severity will be the same for acute and standard versions.* Table 5 shows that in comparison to the standard form, changes in scores for the acute form tended to be more responsive to recent changes in disease state. Changes in disease state were operationalized as improved, stayed the same, or declined according to patient-reported transitions in general health and asthma condition over the past week as well as changes in asthma-specific daily diary scores over the past week (see Table 1). MANOVA *F*-statistics favored the acute form four-week change scores for all three criteria. Out of

18 comparisons, 15 favored the acute form. In ten cases, the *F*-statistics for scales on the acute form were statistically significant while the standard form versions were not. Highly consistent results were seen for BP, RP, and SF in these analyses, concepts whose acute forms were more sensitive to change according to all three criteria; however, the acute RP scale was the only scale for which this effect was significant across all three criteria.

DISCUSSION

SUMMARY OF RESULTS

In general, recall period did not affect whether scales conformed to assumptions underlying their construction and scoring. The only consistent exception was lower internal consistency estimates for the RE scale, which was linked to lower variability in RE scores. Because patients in the study were free from major health problems other than their asthma and since asthma's main impact is on the physical dimensions of health, we neither anticipated, nor found limitations due to emotional health in this sample. This would be especially true during a short recall period, because previous research has shown that the frequency of reported health events is a function of length of recall period: longer recall periods permit more opportunities for events to occur (Cohen, Erickson, and Powell 1983).

Compared to changes for the standard scales, change scores (from baseline to week 4) for the acute scales were generally more highly related to one-week change in disease severity. The acute form may have been generally superior to the standard in detecting the impact of change in disease severity over the past week because the time frame for clinical change represented a greater proportion of the acute, relative to the standard, recall period. A limitation of the current study is that it was not specifically designed to compare the sensitivity of the acute and standard forms to health events that occurred outside the recall period of the acute form. Determination of whether any differences in sensitivity between acute and standard forms are noteworthy would require a study with a larger sample and both one-week and four-week intervals between data collections for SF-36 and clinical variables.

The physical and mental health constructs underlying the standard form were replicated in the acute form. In other words, SF-36 scales have the same factor content and interpretation regardless of whether respondents consider the previous week or the previous month. This result supports the creation of

two summary component scores for the acute form, which has been shown to decrease the eight outcome measures to two without substantial loss of information in studies of the standard form (Ware, Kosinski, and Keller 1994; Ware, Kosinski, Bayliss, et al. 1995).

While the physical and mental components were replicated in this sample, the relationship of the GH scale to the mental component was not: it was found that the correlation of GH to mental health (both acute and standard forms) was lower in this sample than in the U.S. general population. This suggests that asthma patients' perceptions of their general health may be influenced less by their mental health than are the health perceptions of a sample from the general U.S. population. This result is consistent with the clinical picture of asthma as a condition that primarily affects physical functioning; thus, evaluations of general health primarily reflect physical health among patients with asthma.

Our results suggest that a large difference in mean scores between acute and standard forms is unlikely; however, we are cautious about accepting the null hypothesis of no difference because observed differences approached significance and the confidence intervals around differences observed are substantial. Further, univariate analyses (not adjusted for multiple comparisons) indicated a significant difference between acute and standard forms for the RE scale. The higher acute means may be due to a lower prevalence of negative events during the shorter acute recall period, as noted above. The potential difference in mean scores by form has implications for norm-based interpretation of SF-36 scale scores (see Implications for Measurement and Further Research of HQL below).

LIMITATIONS

Restricted Disease Severity

This study was part of a larger clinical research trial that excluded some asthmatics. Studied patients were healthy except for their asthma condition and they represented a limited range of asthma severity (55–85% FEV_1%). As a result, their scores were less variable than those in the U.S. general population for some of the scales. Therefore, conclusions about differences in acute and standard forms may not be generalizable to a different population of asthma sufferers or to a well population in general. The greater drop-out rate among the patients with more severe asthma and with lower health scores restricted variability further.

Sample Size and Error Rate

Sample sizes were relatively small for groups where standard ($n = 68$) and acute ($n = 74$) forms were administered. Thus, statistical power to detect differences in correlations and group means as function of form was low. In addition, for the analyses of sensitivity to disease severity, patients were divided into groups according to whether they had improved, declined, or stayed the same. These subsamples were small and unequal in size and their variances were unequal as well, a phenomenon shown to affect the Type I error rate (Glass, Peckham, and Sanders 1972; Zimmerman 1987). MANOVA *F*s and the consideration of convergence in results across criteria were used to balance this effect. Also, when nonparametric statistics were used to replicate parametric tests of the fourth hypothesis, results were found to be consistent with the parametric results.

IMPLICATIONS FOR FURTHER RESEARCH

Before generalizing these results, the acute form should be tested in a sample from the general population and among other disease groups. Further tests of the sensitivity of acute and standard forms relative to acute changes in disease severity should employ a larger sample with greater variability in disease severity and in patterns of change in severity over time. For full understanding of the relative advantages and disadvantages of the acute and standard forms, the forms should also be compared in applications for which the standard form is likely to have an advantage. For example, the sensitivity of acute forms to change in condition should be compared to standard forms in measuring patients with chronic conditions.

These data indicate that scores based on the acute form may be more favorable particularly for those scales that would be less affected by change in condition over a short time period. Occurrence of such events is a function of the length of the recall period; longer recall periods permit more opportunities for events to occur (Cohen, Erickson, and Powell 1983). Comparisons of average scores suggest that the acute form yields more favorable mean scores, particularly for the two role-functioning scales. Implications for interpretation depend on the pattern of results, specifically: (1) when acute scores are ***higher than*** norms for the standard form, differences may have been *over*estimated; and (2) when acute scores are ***lower than*** or ***equal to*** norms for the standard form, differences may have been *under*estimated; however, in such cases, significant differences are likely to be real, given that the data reported here suggest that the acute form is likely to be favorably biased. Stronger recommendations regarding interpretations of acute scores await further research.

CONCLUSIONS

The results of this study support the use of the acute form in further research and suggest that it may achieve its desired objective. Specifically, these results demonstrate that the acute form yields high-quality data, that the scoring algorithms developed for the standard form are appropriate for the acute form, that the acute scales are reliable, and that the acute form measures the eight concepts and two summary concepts measured by the standard form. Thus, the results indicate that users of the acute form are employing an instrument with psychometric properties equal to the standard. In applications requiring weekly or biweekly assessments of health, the acute form may be preferred to the standard because the one-week recall period may make more sense to respondents. Evidence suggests that the acute form is at least as sensitive, and possibly more sensitive, than the standard version to recent changes in health. However, strong conclusions regarding the relative sensitivity of acute and standard forms await further research. In addition, comparability between standard and acute mean scores has not been definitively established. Thus, when norms based on the standard form are used to interpret scale scores from the acute form, the reader should be advised that comparability of these norms awaits further research.

REFERENCES

Annotated Bibliography for the SF-36 Health Survey. 1996. Boston, MA: New England Medical Center, The Health Institute.

Bergner, M., R. A. Bobbitt, S. Kressel, W. E. Pollard, B. S. Gilson, and J. R. Morris. 1976. "The Sickness Impact Profile: Conceptual Formulation and Methodology for the Development of a Health Status Index." *International Journal of Health Services* 6 (3): 393–415.

Cohen, B., P. Erickson, and A. Powell. 1983. "The Impact on Length of Recall Period on the Estimation of Health Events." *Survey Results and Methods Section Proceedings of the American Statistical Association.* 497–501.

Comrey, A. L., and H. B. Lee. 1992. *A First Course in Factor Analysis, Second Edition.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Cronbach, L. J. 1951. "Coefficient Alpha and the Internal Structure of Tests." *Psychometrika* 16 (3): 297–334.

Fowler, F. J. 1984. *Survey Research Methods.* Beverly Hills, CA: Sage Publications.

Glass, G. V., P. D. Peckham, and J. R. Sanders. 1972. "Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance." *Review of Educational Research* 42 (3): 237–88.

Helmstadter, G. C. 1964. *Principles of Psychological Measurement.* New York: Appleton-Century-Crofts, Inc.

Howard, K. I., and G. G. Forehand. 1962. "A Method for Correcting Item-Total Correlations for the Effect of Relevant Item Inclusion." *Educational and Psychological Measurement* 22 (4): 731–35.

Kerlinger, F. N. 1973. *Foundations of Behavioral Research.* New York: Holt, Rinehart and Winston, Inc.

Likert, R. 1932. "A Technique for the Measurement of Attitudes." *Archives of Psychology* 140 (1): 5–55.

McHorney, C. A., M. Kosinski, and J. E. Ware. 1994. "Comparisons of the Costs and Quality of Norms for the SF-36 Health Survey Collected by Mail versus Telephone Interview: Results from a National Survey." *Medical Care* 32 (6): 551–67.

McHorney, C. A., J. E. Ware, J. F. R. Lu, and C. D. Sherbourne. 1994. "The MOS 36-Item Short-Form Health Survey (SF-36): III. Tests of Data Quality, Scaling Assumptions and Reliability Across Diverse Patient Groups." *Medical Care* 32 (1): 40–66.

McHorney, C. A., J. E. Ware, and A. E. Raczek. 1993. "The MOS 36-Item Short-Form Health Status Survey (SF-36): II. Psychometric and Clinical Tests of Validity in Measuring Physical and Mental Health Constructs." *Medical Care* 31 (3): 247–63.

Patrick, D. L., and P. Erickson. 1993. *Health Status and Health Policy: Allocating Resources to Health Care.* New York: Oxford University Press.

Spitzer, W. O., A. J. Dobson, J. Hall, E. Chesterman, J. Levi, R. Shepherd, R. N. Battista, and B. R. Catchkive. 1981. "Measuring the Quality of Life of Cancer Patients: A Concise QL-Index for Use by Physicians." *Journal of Chronic Diseases* 34 (5): 585–97.

Stewart, A. L., and J. E. Ware. 1992. *Measuring Functioning and Well-Being: The Medical Outcomes Study Approach.* Durham, NC: Duke University Press.

Ware, J. E., M. Kosinski, M. S. Bayliss, C. A. McHorney, W. H. Rogers, and A. Raczek. 1995. "Comparison of Methods for the Scoring and Statistical Analysis of SF-36 Health Profiles and Summary Measures: Results from the Medical Outcomes Study." *Medical Care* 33 (Supplement): AS264–79.

Ware, J. E., M. Kosinski, and S. D. Keller. 1994. *SF-36 Physical and Mental Component Summary Measures: A User's Manual.* Boston: The Health Institute.

Ware, J. E., and C. D. Sherbourne. 1992. "The MOS 36-Item Short-Form Health Survey (SF-36): I. Conceptual Framework and Item Selection." *Medical Care* 30 (6): 473–83.

Ware, J. E., K. K. Snow, M. Kosinski, and B. Gandek. 1993. *SF-36 Health Survey Manual and Interpretation Guide.* Boston: New England Medical Center, The Health Institute.

Wilkin, D., L. Hallam, and M. A. Doggett. 1992. *Measures of Need and Outcomes for Primary Health Care.* Oxford: Oxford Medical Press.

Zimmerman, D. W. 1987. "Comparative Power of Student *t*-Test and Mann Whitney U Test for Unequal Sample Sizes and Variances." *Journal of Experimental Education* 55 (3): 171–74.