SPINE Volume 25, Number 24, pp 3130–3139

SF-36 Health Survey Update

John E. Ware, Jr, PhD

The SF-36 (Medical Outcomes Trust, Boston, MA) is a multipurpose, short-form health survey with only 36 questions. It yields an eight-scale profile of scores as well as physical and mental health summary measures. It is a generic measure, as opposed to one that targets a specific age, disease, or treatment group. Accordingly, the SF-36 has been useful in comparing general and specific populations, comparing the relative burden of diseases, differentiating the health benefits produced by a wide range of different treatments, and screening individual patients.[47] This report summarizes the steps in the construction of the SF-36; how it led to the development of an even shorter (one-page, 2-minute) survey form, the SF-12; the improvements reflected in version 2.0 of the SF-36, psychometric studies of assumptions underlying scale construction and scoring, how they have been translated in more than 40 countries as part of the International Quality of Life Assessment (IQOLA) project, and studies of reliability and validity.

SF-36 Literature

The experience to date with the SF-36 has been documented in more than 1000 publications. Those published in 1998 and before have been summarized in an annotated bibliography.[47] The most complete information about the history and development of the SF-36, its psychometric evaluation, studies of reliability and validity, and normative data are available in the first of three SF-36 user's manuals.[71] A second manual documents the[14] development and validation of the SF-36 summary measures and presents norms for those measures.[68] A third[67] presents similar information for the SF-12 Health Survey, an even shorter version constructed from a subset of 12 items, and compares that form with the SF-36. One of the most complete independent accounts of SF-36 development along with a critical commentary is offered by McDowell and Newell.[31] Additional publications are listed on the SF-36 Web page (http://www.sf-36.com).

The usefulness of the SF-36 in estimating disease burden is illustrated in articles describing more than 130 diseases and conditions. Among the most frequently studied conditions, with more than 20 SF-36 publications each, are arthritis, back pain, depression, diabetes, and hypertension.[47] Translation of the SF-36 is the subject of 148 publications, and one or more articles compare results from the SF-36 with those of 225 other generic and disease-specific instruments.[47]

Construction of the SF-36

The SF-36 was constructed to satisfy minimum psychometric standards necessary for group comparisons. The eight health concepts were selected from 40 concepts included in the Medical Outcomes Study (MOS).[50] Those chosen represent the most frequently measured concepts in widely used health surveys and those most affected by disease and treatment.[68,70] SF-36 items also represent multiple operational indicators of health, including behavioral function and dysfunction, distress and well-being, objective reports and subjective ratings, and both favorable and unfavorable self-evaluations of general health status.[68]

Most SF-36 items have their roots in instruments that have been in use since the 1970s and 1980s,[50] including the General Psychological Well-Being Inventory,[15] various physical and role functioning measures,[18,41,45,51] the Health Perceptions Questionnaire,[14] and other measures that were useful during the Health Insurance Experiment (HIE).[11] The MOS researchers selected and adapted questionnaire items from these and other sources and developed new measures for a 149-item Functioning and Well-Being Profile (FWBP).[50] The FWBP was the source for questionnaire items and instructions adapted for use in the SF-36. The SF-36 was first made available in a "developmental" form in 1988 and in "standard" form in 1990.[58,70] As documented elsewhere,[71] the standard form eliminated more than one fourth of the words contained in MOS versions of the 36 items and also reflected improvements in item wording, format and scoring.

Version 2.0

In 1996, version 2.0 of the SF-36 —the international version—was introduced, to improve the two role functioning scales and to achieve other objectives.[65] Compared with the standard SF-36 version 1.0, improvements in version 2.0 included simpler instructions and questionnaire items, an improved layout for questions and answers in the self-administered version, greater compara-

From *QualityMetric, Inc.; Lincoln, Rhode Island; and the Health Assessment Lab; Tufts Medical School, and Harvard School of Public Health, Boston, Massachusetts.
Development and validation of the SF-36 Health Survey was supported by a grant from the Henry J. Kaiser Family Foundation to The Health Institute, New England Medical Center (J. E. Ware, Jr., Principal Investigator). Development of the SF-36 PCS and MCS summary measures was supported by unrestricted research grants for the International Quality of Life Assessment (IQOLA) Project for the Glaxo Research Institute, Research Triangle Park, NC and Schering-Plough Corp., Kenilworth, NJ (J. E. Ware, Jr., Principal Investigator).
The IQOLA Project is sponsored by unrestricted research grants from GlaxoWellcome Inc., Research Triangle Park, NC, and Schering-Plough Corp., Kenilworth, NJ. Associate sponsors include Astra, Parke-Davis, Pharmacia & Upjohn, Proctor & Gamble Pharmaceuticals, Searle, Solvay Duphar B.V., and Synthelabo. Additional support has also been provided by more than 40 other pharmaceutical companies.

bility with widely used translations and cultural adaptations, and five-level response choices in place of dichotomous response choices for items in the two role functioning scales. These and other improvements are briefly explained in the next section.

Layout

All responses to questions in version 2.0 are printed in a left-to-right (also referred to as horizontal) format, rather than with the mixture of horizontal and vertical listings of response choices that were printed below questions in the MOS and in the original SF-36. Mixed formats of response choices confuse respondents and cause missing and inconsistent responses, particularly among the elderly. Other improvements include more consistent use of indentation, numbering of instructions, deletion of useless item labels, and a simpler formatting of boxes that are checked by respondents.

Type Size and Boldface Type

A larger type size has been adopted throughout. Only instructions, as opposed to response choices, are in bold typeface to simplify the look and feel of version 2.0. These and other refinements were adopted on the basis of lessons learned in health care and from surveys in other fields.

Wording Changes

Evidence from numerous focus group studies, formal cognitive tests, and empirical studies in more than a dozen countries support the improvements in item wording and the changes in some terms used to identify health concepts adopted in version 2.0. These improvements make the English-language SF-36 easier to understand and administer and make it more objective. Version 2.0 is also more comparable with translations of the SF-36. Because most of the improvements in item wording were developed during the process of translating and adapting the SF-36 for use in other countries during the International Quality of Life Assessment (IQOLA) Project, version 2.0 is sometimes referred to as the international version.

Five-Choice Response Scales

There is considerable empirical evidence that the version 2.0 five-level response scales substantially improve the two SF-36 role functioning scales. Version 2.0 response scales extend the range measured and greatly increase score precision without increasing respondent burden. Specifically, version 2.0 achieves a fourfold increase in the number of levels defined by both role scales, a substantially smaller standard deviation, and substantially reduces the percentage of respondents who score at both the ceiling and floor for both role scales. The elimination of one of the six response choices ("a good bit of the time") from the mental health and vitality items was based on the finding that this response choice is not consistently ordered between adjacent categories in studies of item responses in version 1.0 or in translations of the SF-36. Eliminating this choice simplified the format of the form with little or no loss of information.

Scoring

Version 2.0 scoring uses norm-based scoring algorithms for all eight scales (T-score transformation with mean, 50 ± 10 [SD]) that has made the SF-36 summary measures much easier to interpret. Version 2.0 scoring software also achieves improved estimation of missing responses and provides respondent-specific data quality indicators.

Comparability of Results

To make version 1.0 easier to interpret and directly comparable to results based on version 2.0, cross-sectional and longitudinal norms for general and specific populations were re-estimated for version 1.0 using norm-based scoring for all eight scales and for the two summary measures. Further, national calibration studies were fielded in the United States in 1998 and 1999 to evaluate the effect of all improvements and to assure the comparability of average scores across versions 1.0 and 2.0.

■ Psychometric Considerations

SF-36 Measurement Model

Figure 1 illustrates the taxonomy of items and concepts underlying the construction of the SF-36 scales and summary measures. The taxonomy has three levels: (1) items, (2) eight scales that aggregate 2–10 items each, and (3) two summary measures that aggregate scales. All but one of the 36 items (self-reported health transition) are used to score the eight SF-36 scales. Each item is used in scoring only one scale.

The eight scales are hypothesized to form two distinct higher ordered clusters according to the physical and mental health variance that they have in common. Factor-analytic studies have confirmed physical and mental health factors that account for 80–85% of the reliable variance in the eight scales in the US general population,[68] among patients in the MOS,[34,68] and in general populations in Sweden[56] and the United Kingdom.[63] As of 1998, these studies had been replicated in more than a dozen countries.[12,68]

Three scales (Physical Functioning, Role-Physical, Bodily Pain) correlate most highly with the physical component and contribute most to the scoring of the Physical Component Summary (PCS) measure.[68] The mental component correlates most highly with the Mental Health, Role-Emotional, and Social Functioning scales, which also contribute most to the scoring of the Mental Component Summary (MCS) measure. Three of the scales (Vitality, General Health, and Social Functioning) have noteworthy correlations with both components.

The importance of these findings is illustrated in the discussion of empirical validity that follows. Specifically, scales that load highest on the physical component are most responsive to treatments that change physical morbidity, whereas scales loading highest on the mental

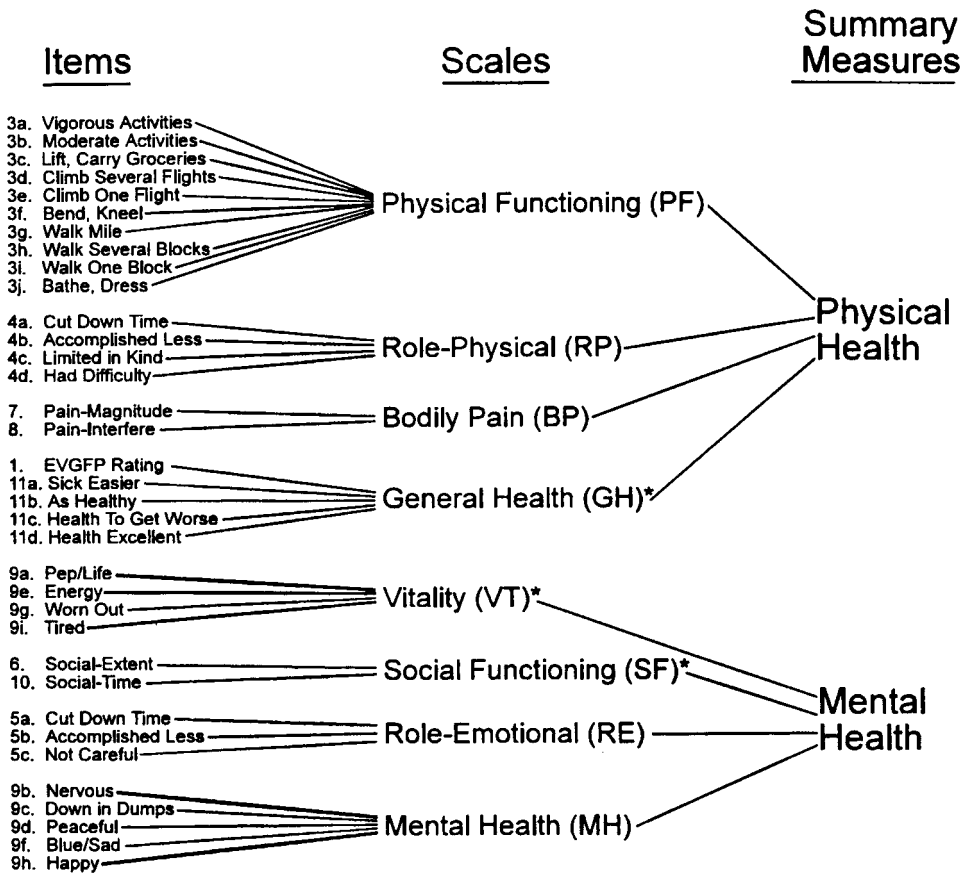

Figure 1. SF-36 measurement model. *Significant correlation with other summary measure.

component respond most to drugs and therapies that target mental health.

Scaling and Scoring Assumptions

A major objective in constructing the SF-36 was achievement of high psychometric standards. Guidelines for testing were derived from those recommended for use in validating psychological and educational measures by the American Psychological Association, the American Education Research Association, and the National Council on Measurement in Education.[3] Extensive psychometric testing has been conducted on the SF-36 in the United States[32–34,66,71] and other countries.[4,10,16,20,30,43,52,63] By using the same tests of scaling and scoring assumptions that were used in developing the SF-36, investigators have compared results across general population studies in 10 countries.[12]

On the strength of favorable results from tests to date, nearly all studies have used the method of summated ratings and standardized SF-36 scoring algorithms documented elsewhere.[68,71] This method assumes that items shown in the same scale in Figure 1 can be aggregated without score standardization or item weighing. Standardization of items within a scale was avoided by selecting or constructing items with roughly equivalent means and standard deviations. Weighing was avoided by using equally representative items (that is, items with roughly equivalent correlations to the underlying scale dimension). All items have been shown to correlate substantially (greater than 0.40, corrected for overlap) with their hypothesized scales with rare exceptions.[33,71]

Reliability and Confidence Intervals

The reliability of the eight scales and two summary measures has been estimated using both internal consistency and test–retest methods. With rare exceptions, published reliability statistics have exceeded the minimum standard of 0.70 recommended for measures used in group comparisons in more than 25 studies[53]; most have exceeded 0.80.[33,71] Reliability estimates for physical and mental summary scores usually exceed 0.90.[68] A review of the first 15 published studies revealed that the median reliability coefficients for each of the eight scales was equal or greater than 0.80 except for Social Functioning, which had a median reliability across studies of 0.76.[71] In addition, a reliability of 0.93 has been reported for the Mental Health scale, by using the alternate forms method, suggesting that the internal consistency method underestimated the reliability of that scale by about 3%.[32]

The trends in reliability coefficients for the SF-36 scales and summary measures summarized have also been replicated across 24 patient groups differing in sociodemographic characteristics and diagnoses.[33,68,71] Although studies of subgroups indicate slight declines in reliability for more disadvantaged respondents, reliability coefficients consistently exceeded recommended standards for group level analysis. Reliability estimates con-

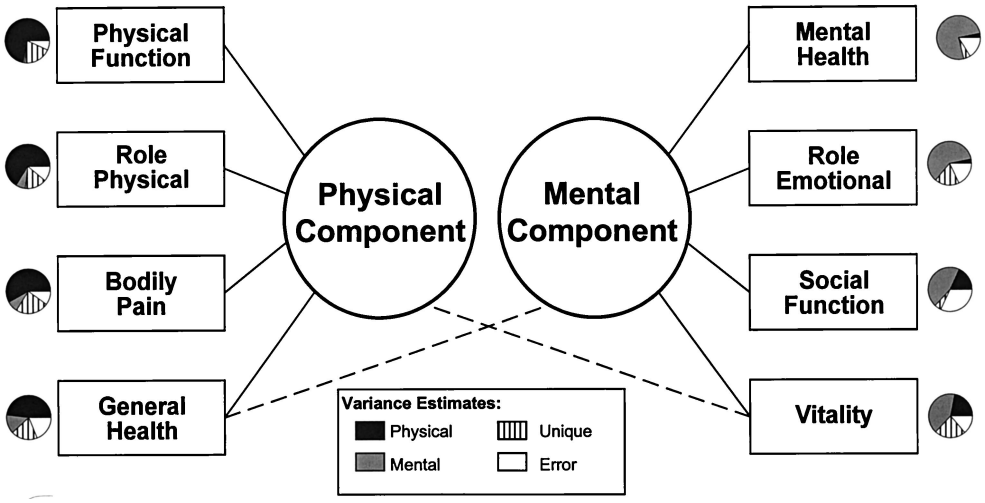

Figure 2. SF-36 scales measure physical and mental components of health. (Source: Ware, Kosinski, and Keller.[68])

sistent with these trends have been published in more than 200 studies, results from more than 30 test–retest studies have also been summarized.[47]

Standard errors of measurement, 95% confidence intervals for individual scores, and distributions of change scores from test–retest and 1-year stability studies have been published for the eight SF-36 scales and for the two summary scores.[10,68,71] Confidence intervals around individual scores are much smaller for the two summary measures than for the eight scales (±6–7 points *vs.* ±13–32 points, respectively)[68] Estimates of sample sizes required to detect differences in average scores of various magnitudes have been documented for five different study designs for each of the eight scales and for the two summary measures.[68,71]

Validity

Studies of validity generally support the intended meaning of high and low SF-36 scores as documented in the original user's manuals.[68,71] Because of the widespread use of the SF-36 across a variety of applications, evidence from many types of validity research is relevant to these interpretations. Studies to date have yielded content, concurrent, criterion, construct, and predictive evidence of validity.

The content validity of the SF-36 has been compared with that of other widely used generic health surveys.[68,71] Systematic comparisons indicate that the SF-36 includes eight of the most frequently measured health concepts. Among the content areas included in widely used surveys, but not included in the SF-36, are: sleep adequacy, cognitive functioning, sexual functioning, health distress, family functioning, self-esteem, eating, recreation and hobbies, communication, and symptoms and problems that are specific to one condition. Symptoms and problems that are specific to a particular condition are not included in the SF-36, because the SF-36 is a generic measure.

To facilitate the evaluation of concepts not included, the SF-36 users' manuals include tables of correlations between the eight scales and the two summary measures and 32 measures of other general concepts,[68,71] as well as 19 specific symptoms. SF-36 scales correlate substantially ($r = 0.40$ or greater) with most of the omitted general health concepts and with the frequency and severity of many specific symptoms and problems. A noteworthy exception is sexual functioning, which correlates relatively weakly with SF-36 scales and is a good candidate for inclusion in questionnaires that supplement the SF-36.

Because most SF-36 scales were constructed to reproduce longer scales, attention was initially given to how well the short-form versions perform in empirical tests relative to the full-length versions. Relative to the longer MOS measures they were constructed to reproduce, SF-36 scales have been shown to achieve approximately 80–90% of their empirical validity in studies involving physical and mental health criteria.[34]

The validity, and therefore the interpretation, of each of the eight scales and the two summary measures has been shown to differ markedly, as would be expected from factor-analytic studies of their construct validity (see Figure 2).[34,67,68] Specifically, the Mental Health, Role-Emotional, and Social Functioning scales and the MCS summary measure have been shown to be the most valid of the SF-36 scales as mental health measures. This pattern of results has been replicated in both cross-cultural and longitudinal tests using the method of known-groups validity. The Physical Functioning, Role-Physical, and Bodily Pain scales and the PCS summary have been shown to be the most valid SF-36 scales for measuring physical health. Criteria used in the known-groups validation of the SF-36, which include accepted clinical indicators of diagnosis and severity of depression, heart disease, and other conditions, are well-documented in peer-reviewed publications and in the two users' manuals.[25,34,66,68,71]

The Mental Health scale has been shown to be useful in screening for psychiatric disorders,[8,68] as has the MCS summary measure.[68] For example, using a cutoff score of 42, the MCS had a sensitivity of 74% and a specificity of

81% in detecting patients diagnosed with depressive disorder.[68]

Relative to other published measures, SF-36 scales have performed well in most tests published to date.[10,22,26,27,72] As documented in the SF-36 annotated bibliography,[47] studies have compared the SF-36 with 225 other measures. Results in predictive studies of validity have linked SF-36 scales and summary measures to utilization of health care services,[68] the clinical course of depression,[9,71] loss of job within 1 year,[68] 180-day survival,[49] and 5-year survival.[68]

Results from clinical studies comparing scores for patients before and after treatment have largely supported hypotheses about the validity of SF-36 scales based on results of psychometric studies. For example, clinical studies have shown that three of the scales (Physical Functioning, Role-Physical, and Bodily Pain) with the most physical factor content (Figure 2) tend to be most responsive to the benefits of knee replacement,[24] hip replacement,[24,29] and heart valve surgery,[42] In contrast, the three scales with the most mental factor content (Mental Health, Role-Emotional, and Social Functioning) in factor-analytic studies have been shown to be most responsive in comparisons of patients before and after recovery from depression,[66] change in the severity of depression,[9] and as well as drug treatment and interpersonal therapy for depression.[13]

The discovery that 80–85% of the reliable variance in the eight SF-36 scales led to the construction of psychometrically based physical and mental health summary measures. It was hoped that they would make it possible to reduce the number of statistical comparisons involved in analyzing the SF-36 (from eight to two) without substantial loss of information. In both cross-sectional and longitudinal studies reported to date, this appears to be the case.[66,68] The advantages and disadvantages of analyzing the eight-scale SF-36 profile *versus* the two summary measures are illustrated and discussed elsewhere.[66,68]

Finally, the SF-36 self-evaluated health transition item (five response categories ranging from "much better" to "much worse), which is not used in scoring the scales or summary measures, has been shown to be useful in estimating average changes in health status during the year before its administration. In the MOS, measured changes in health status during a 1-year follow-up period corresponded substantially, on average, to self-evaluated transitions at the end of the year. With the 0-100 GHRI scale[14] serving as a criterion, those who evaluated their health as much better improved an average of 13.2 points. The average change was 5.8 points for those who reported that they were somewhat better. An average decline of −10.8 was observed for those who reported that their health was somewhat worse and −34.4 for those reporting much worse. (It should be noted that the latter category had only 29 patients.) Change scores for those choosing the "about the same" category averaged 1.6 points. These results are encouraging with regard to the use and interpretation of self-evaluated transitions at the group level. Pending results from ongoing studies of the reliability of responses to the SF-36 self-evaluated transition item, it should be interpreted with caution at the individual level. Additional results and their implications are discussed elsewhere.[68,71]

■ Administration Methods and Scoring

The SF-36 is suitable for self-administration, computerized administration, or administration by a trained interviewer in person or by telephone, to persons aged 14 years and older. The SF-36 has been administered successfully in general population surveys in the United States and other countries[64] as well as to young and older adult patients with specific diseases.[71] It can be administered in 5–10 minutes with a high degree of acceptability and data quality.[71] Indicators of data quality that have yielded satisfactory results in studies to date include very high item completion rates and favorable results for a response consistency index based on 15 pairs of SF-36 items, which is scored at the individual level.[71] Computer-administered and telephone voice recognition interactive systems of administration are currently being evaluated.

Summary Measures

Table 1 summarizes information about the eight SF-36 scales and two summary measures that is important in their use and interpretation. The eight scales are ordered in terms of their factor content (*i.e.*, construct validity), because they are in the SF-36 profile to facilitate interpretation. The first scale is Physical Functioning, which has been shown to be the best all around measure of physical health; the last scale, Mental Health, is the most valid measure of mental health in studies to date.[34,68,71] Of note, Mental Health and Physical Functioning are the poorest measures of the physical and mental components, respectively. Scales between those are ordered according to their validity in measuring physical and mental health. The Vitality and General Health scales have substantial or moderate validity for both components of health status and should be interpreted accordingly.

The number of items and levels and the range of states defined by each scale are also shown in Table 1. These attributes have been linked to their empirical validity.[35] The most precise (least coarse) scales are those with 20 or more levels (Physical Functioning, General Health, Vitality, and Mental Health). They also define the widest range of health states and therefore usually produce the least skewed score distributions. The relatively coarse role disability scales (Role-Physical and Role-Emotional) each measure only four or five levels across a restricted range and therefore usually have the most problems with ceiling and floor effects.

Means and standard deviations for each of the eight scales in the general US adult population are also presented. These can be used to determine whether a group or individual in question scores above or below the US

Table 1. Summary of Information About SF36 Scales and Physical and Mental Component Summary Measures

Scales	Correlations PCS	Correlations MCS	Number of Items	Number of Levels	Mean§	SD	Reliability	CI*	Definition (% observed) Lowest Possible Score (Floor)‡	Definition (% observed) Highest Possible Score (Ceiling)‡
Physical Functioning (PF)	.85	.12	10	21	84.2	23.3	.93	12.3	Very limited in performing all physical activities, including bathing or dressing (0.8%)	Performs all types of physical activities including the most vigorous without limitations due to health (38.8%)
Role-Physical (RP)	.81	.27	4	5	80.9	34.0	.89	22.6	Problems with work or other daily activities as a result of physical health (10.3%)	No problems with work or other daily activities (70.9%)
Bodily Pain (BP)	.76	.28	2	11	75.2	23.7	.90	15.0	Very severe and extremely limiting pain (0.6%)	No pain or limitations due to pain (31.9%)
General Health (GH)	.69	.37	5	21	71.9	20.3	.81	17.6	Evaluates personal health as poor and believes it is likely to get worse (0.0%)	Evaluates personal health as excellent (7.4%)
Vitality (VT)	.47	.65	4	21	60.9	20.9	.86	15.6	Feels tired and worn out all of the time (0.5%)	Feels full of pep and energy all of the time (1.5%)
Social Functioning (SF)	.42	.67	2	9	83.3	22.7	.68	25.7	Extreme and frequent interference with normal social activities due to physical and emotional problems (0.6%)	Performs normal social activities without interference due to physical or emotional problems (52.3%)
Role-Emotional (RE)	.16	.78	3	4	81.3	33.0	.82	28.0	Problems with work or other daily activities as a result of emotional problems (9.6%)	No problems with work or other daily activities (71.0%)
Mental Health (MH)	.17	.87	5	26	74.7	18.1	.84	14.0	Feelings of nervousness and depression all of the time (0.0%)	Feels peaceful, happy, and calm all of the time (0.2%)
Physical Component Summary (PCS)			35	567†	50.0	10.0	.92	5.7	Limitations in self-care, physical, social, and role activities, severe bodily pain, frequent tiredness, health rated "poor" (0.0%)	No physical limitations, disabilities, or decrements in well-being, high energy level, health rated "excellent" (0.0%)
Mental Component Summary (MCS)			35	493†	50.0	10.0	.88	6.3	Frequent psychological distress, social and role disability due to emotional problems, health rated "poor" (0.0%)	Frequent positive affect, absence of psychological distress and limitations in usual social/role activities due to emotional problems, health rated "excellent" (0.0%)

Note: From Ware, Kosinski, and Keller.[68]
* CI = 95% confidence interval.
† Number of levels observed at baseline; scores rounded to the first decimal place (n = 2474).
‡ Percentage observed comes from general U.S. population sample.
§ Scores for eight scales are the percentage of the total possible score achieved for each of these scales. Scores for PCS and MCS are *T*-scores.
PCS = physical component summary; MCS = mental component summary.

average. Detailed normative data including frequency distributions of scores and percentile ranks are documented in the two users' manuals.[68,71]

Table 1 illustrates the practical implications of a number of theoretical advantages of the PCS and MCS summary measures, including reliability and the number and range of levels covered.

Norm-Based Scoring and Interpretation

The interpretation of results has been made much easier with the standardization of mean scores and standard deviations for all SF-36 scales. Specifically, norm-based scoring has been very useful when interpreting differences across scales in the SF-36 profile and for monitoring disease groups over time. As documented elsewhere,[68] linear transformations were performed to transform scores to a mean of 50 and standard deviations of 10, in the general US population. This transformation achieves the same mean and standard deviation for all eight scales and for the PCS and MCS measures.

The advantages of norm-based scoring can be illustrated by comparing the SF-36 profile scored using the original 0–100 scoring algorithms based on the summated ratings method and the norm-based scoring algorithms for a sample of asthmatic patients who participated in a clinical trial.[39] The original SF-36 0–100 scoring produced the profile shown in Figure 3. The shape of this profile—the peaks and valleys due to higher and lower scores across scales—reflect both the impact

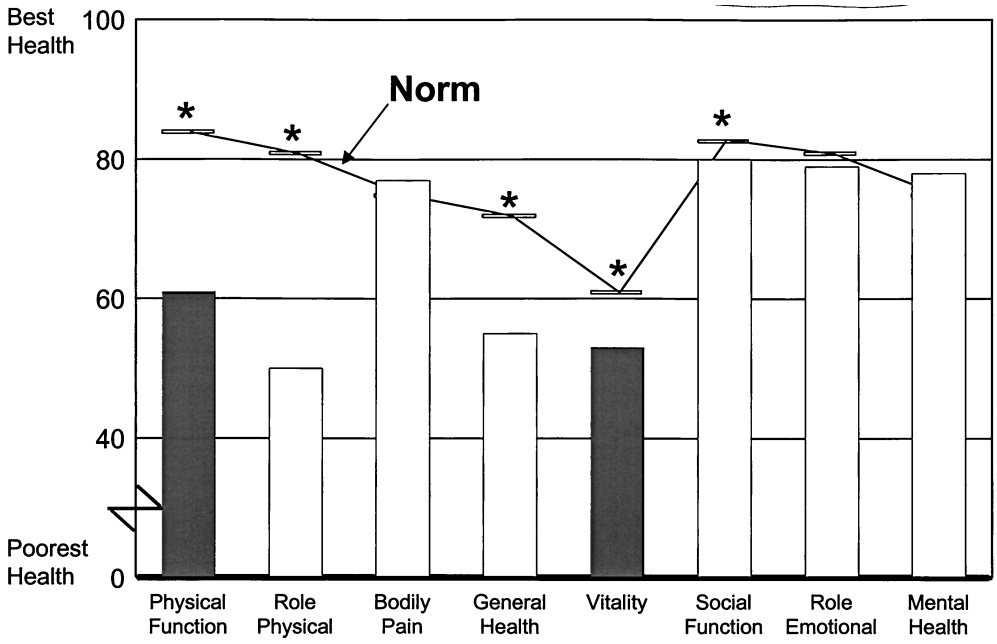

Figure 3. SF-36 health profile: adults with asthma compared with U.S. norm. (Source: Okamoto.[39]) *Norm significantly higher.

of asthma on SF-36 health concepts and arbitrary differences in the ceilings and floors of the SF-36 scales. Three scales, namely, General Health, Vitality, and Mental Health, measure relatively wide score ranges and set the ceiling relatively high by measuring very favorable levels of those health concepts.[71] Other scales, such as Physical Functioning, and Role-Physical, assess a narrower range. The most favorable levels (scored 100 using the original SF-36 algorithms) for physical functioning and Role-Physical represent the absence of limitations and do not extend the range into well-being. Thus, the average score for each scale differs substantially across the profile for reasons that have nothing to do with asthma when the original SF-36 0–100 scoring is used. The inference from the profile in Figure 3, that asthma has a greater impact on Physical Functioning than on Vitality, is incorrect.

General population norms provide a much better basis for comparisons across scales (see Figure 3). For example, the Physical Functioning scale averages between 80 and 90, whereas the Vitality average score is below 60 (on the 100-point score range) in the general population. In relation to these norms, the impact of asthma appears much larger on the Physical Functioning scale than on the Vitality scale, although both are statistically significant. When using the original 0–100 scoring, these differences in norms must be kept in mind when interpreting a profile. Differences in standard deviations, which are also substantial across some scales, must also be considered for this purpose.

In norm-based scoring, each scale was scored to have same average (50) and the same standard deviation (10 points). Without referring to norms, it is clear that anytime a scale score is below 50, health status is below average, and each point is one tenth of a standard deviation. As shown in Figure 4, with norm-based scoring, differences in scale scores much more clearly reflect the impact of the disease, in this example the impact of asthma. Clinicians can more quickly and appropriately interpret the effect of asthma on a SF-36 health profile. Because the PCS and MCS measures take into account the correlation among the eight SF-36 scales, it is clear that asthma impacts on the physical component of health, and (from the profile with five significant differences) impacts very broadly.

The application of norm-based scoring to a clinical trial of treatment effects is illustrated in Figure 5. Patients treated using an inhaler showed statistically significant improvements relative to baseline after 16 weeks of treatment on three of the eight SF-36 scales, those most closely associated with physical functioning.

■ Translations

The International Quality of Life Assessment (IQOLA) Project is translating, validating, and norming the SF-36 Health Survey for use in multinational clinical trials and other international studies.[1,17,62–64] Based at the Health Assessment Laboratory at New England Medical Center, the project began in 1991 with sponsored investigators from 14 countries: Australia, Belgium, Canada, Denmark, France, Germany, Italy, Japan, The Netherlands, Norway, Spain, Sweden, the United Kingdom (English version), and the United States (English and Spanish versions). In addition, researchers from more than 30 other countries are translating and validating the SF-36 using IQOLA Project methods, including: Argentina, Bangladesh, Brazil, Bulgaria, Cambodia, China, Croatia, Czech Republic, Estonia, Finland, Greece, Hong Kong, Hungary, Iceland, Indonesia, Israel, Korea, Mexico, New Zealand, Poland, Portugal, Romania, Russia, Singapore, Slovak Republic, South Africa, Taiwan, Tanzania, Turkey, the United Kingdom (Welsh), the United States (Chinese, Japanese, Vietnamese), and Yugoslavia.

Four major stages of activity are included. First, translation follows a standard protocol, including multiple forward and backward translations. Qualitative and quantitative methods are used to evaluate the quality of a translation and its conceptual equivalence with the original survey. Second, formal psychometric tests of scaling assumptions and scoring assumptions are conducted before publication of a translation. Third, data from clinical trials and other studies are being analyzed to address issues of validity and comparability across countries. Normative data are being collected in general population surveys in eleven countries for purposes of norm-based interpretation. Published norms will soon be available for 10 countries. User's manuals in English, Swedish, and Italian are available, and others are forthcoming.

Published IQOLA Project SF-36 translations and English-language adaptations are distributed royalty-free by the Health Assessment Laboratory. Currently, published forms include the German,[12] Spanish,[2] Swedish,[52] and Italian[4] translations and English-language adaptations for use in Australia and New Zealand, Canada, and the United Kingdom. Information about the availability of SF-36 translations can be accessed on the Internet at http://www.SF-36.com.

■ Discussion

McDowell and Newell[31] attribute the "meteoric rise to prominence" observed for the SF-36 Health Survey to a variety of factors. The widespread adoption of the SF-36 in general population surveys and clinical trials is evidence that more practical measurement tools are more likely to be used. The standardization of measurement across studies is producing considerable information about norms and benchmarks useful in comparing "well" and "sick" populations and for estimating the burden of specific conditions.

Although many studies appear to be relying on the SF-36 as the principal measure of health outcome, among the most useful studies are those that use it as a "generic core." A generic core battery of measures makes it possible to compare results across studies and populations and accelerates the accumulation of interpretation guidelines that are essential to determining the clinical, economic, and social relevance of differences in health status and outcomes. Because it is short, the SF-36 can be reproduced in a questionnaire with ample room for other more precise general and specific measures. Numerous studies[22,38,56] have adopted this strategy and have illustrated the advantages of supplementing it.

How useful is the SF-36 for purposes of comparing general and specific population groups, compared with longer surveys? Some SF-36 scales have been shown to have 10–20% less precision than the long-form MOS measures that SF-36 scales were constructed to reproduce.[35] This disadvantage of the SF-36 should be weighed against the fact that some of these long-form measures place a 5–10 times greater burden on the respondent. Findings in empirical studies of this trade-off indicate that the SF-36 provides a practical alternative to longer measures and that the eight scales and two summary scales rarely miss a noteworthy difference in physical or mental health status in group comparisons.[24,68,71] Regardless, that the SF-36 represents a documented compromise in measurement precision (relative to longer MOS measures) leading to a reduction in the statistical power of hypothesis testing should be taken into account in planning clinical trials and other studies. To facilitate such planning, tables of the sample sizes required for conventional statistical tests are published in the two SF-36 users' manuals.[68,71] Compared with longer non-MOS measures, such as the Sickness Impact Profile, the SF-36 has performed equally well or better in detecting differences in health in two studies.[5,24]

The value of general and specific population norms, which was demonstrated well for the Sickness Impact Profile[7] and later for the MOS SF-20[48,49] and other measures, has also been demonstrated for the SF-36. In addition to the 20 medical conditions described in the MOS and 14 conditions described in the US population norming survey,[68] other publications have reported descriptive data for patients with cardiac disease[21,26]; depressive disorders[13]; epilepsy[54,56]; diabetes mellitus[19,38]; migraine headache[40]; heart transplantation[44]; ischemic heart disease[42]; ischemic stroke[23]; low back pain[16,29]; lung disease[55]; menorrhagia[16]; orthopedic conditions leading to knee replacement,[22] knee surgery,[24] and hip replacement[24,29]; and renal disease.[6,28,37] Whereas some of the initial descriptive studies using the SF-36 were performed primarily to validate scale scores,[35] on the strength of validation studies to date, SF-36 scales appear to be increasingly accepted as valid health measures for purposes of documenting disease burden.

Much remains to be discovered about population health in comprehensive terms of functional health and well-being, the relative burden of disease, and the relative benefits of alternative treatments. One reason has been the unavailability of practical measurement tools appropriate for widespread use across diverse populations. The SF-36 was constructed to provide a basis for such comparisons of results.

As predicted when it was first published,[70] the SF-36 has been widely adopted because of its brevity and its comprehensiveness. Although these two measurement goals are competing, the SF-36 appears to have achieved a psychometrically sound compromise between them. Population and large-group descriptive studies and clinical trials to date demonstrate that the SF-36 is very useful for descriptive purposes such as documenting differences between sick and well patients and for estimating the relative burden of different medical conditions. Although its usefulness in clinical trials was doubted by many, experience to date from more than 250 longitudinal studies suggests that the SF-36 is also a useful tool for evaluating the benefits of alternative treatments.[47]

References

1. Aaronson NK, Acquadro C, Alonso J, et al. International Quality of Life Assessment (IQOLA) Project. Qual Life Res 1992;1:349–51.
2. Alonso J, Prieto L, Antó JM. La versión española del "SF-36 Health Survey" (Cuestionairo de Salud SF-36): un instrumento para la medida de los resultados clínicos. Med Clin (Barc) 1995;104:771–6.
3. American Psychological Association. Standards for Educational and Psychological Tests. Washington, DC: American Psychological Association, 1974.
4. Apolone G, Cifani S, Liberati MC, et al. Questionario sullo stato di salute SF-36. Traduzione e validazione della versione italiana: risultati del progetto IQOLA. Metodologia e Didattica Clinica 1997;5:86–94.
5. Beaton DE, Bombardier C, Hogg-Johnson S. Choose your tool: a comparison of the psychometric properties of five generic health status instruments in workers with soft tissue injuries. Qual Life Res 1994;3:50–6.
6. Benedetti E, Matas AJ, Hakim N, et al. Renal transplantation for patients 60 years or older: a single-institution experience. Ann Surg 1994;220:445–60.
7. Bergner M, Bobbitt RA, Carter WB, et al. Med Care 1981;19:787–805.
8. Berwick DM, Murphy JM, Goldman PA, et al. Performance of a five-item mental health screening test. Medical Care 1991;29(2):169–76.
9. Beusterien KM, Steinwald B, Ware JE. Usefulness of the SF-36 Health Survey in measuring health outcomes in the depressed elderly. J Geriatr Psychiatry Neurol 1996;9.
10. Brazier JE, Harper R, Jones NMB, et al. Validating the SF-36 Health Survey Questionnaire: new outcome measure for primary care. BMJ 1992;305:160–4.
11. Brook RH, Ware JE, Davies-Avery A, et al. Overview of adult health status measures fielded in RAND's Health Insurance Study. Med Care 1979;17(suppl): 1–131.
12. Bullinger M. German translation and psychometric testing of the SF-36 Health Survey: preliminary results from the IQOLA Project. Soc Sci Med 1995; 41:1359–66.
13. Coulehan JL, Schulberg HC, Block MR, et al. Treating depressed primary care patients improves their physical, mental, and social functioning. Arch Intern Med 1997;157:1113–20.
14. Davies AR, Ware JE. Measuring Health Perceptions in the Health Insurance Experiment. Health Insurance Experiment Series. Santa Monica, CA: Rand Corp., 1981; R-2711-HHS.
15. Dupuy HJ. The Psychological General Well-Being (PGWB) index. In: Wenger NK, Mattson ME, Furberg CD, et al., Eds. Assessment of quality of life in clinical trials of cardiovascular disease. new York: Le Jacq Publishing, Inc., 1984:170–83.
16. Garratt AM, Ruta DA, Abdalla MI, et al. The SF-36 Health Survey Questionnaire: an outcome measure suitable for routine use within the NHS? BMJ 1993;306:1440–4.
17. Garratt AM, Ruta DA, Abdalla MI, et al. SF-36 Health Survey Questionnaire: II. responsiveness to changes in health status in four common clinical conditions. Qual Health Care 1994;3:186–92.
18. Hulka, BS, Cassel JC. The AAFP-UNC study of the organization, utilization and assessment of primary medical care. Am J Public Health 1973;63:494–501.
19. Jacobson AM, de Groot M, Samson JA. The evaluation of two measures of quality of life in patients with type I and type II diabetes. Diabetes Care 1994; 17:267–74.
20. Jenkinson C, Coulter A, Wright L. Short Form 36 (SF-36) Health Survey Questionnaire: normative data for adults of working age. BMJ 1993;306:1437–40.
21. Jette DU, Downing J. Health status of individuals entering a cardiac rehabilitation program as measured by the Medical Outcomes Study 36-Item Short-Form Survey (SF-36). Phys Ther 1994;74:521–7.
22. Kantz ME, Harris WJ, Levitsky K, et al. Methods for assessing condition-specific and generic functional status outcomes after total knee replacement. Med Care 1992;30(suppl):MS240–52.
23. Kappelle LJ, Adams HP, Heffner ML, et al. Prognosis of young adults with ischemic stroke: a long-term follow-up study assessing recurrent vascular events and functional outcome in the Iowa Registry of Stroke in young adults. Stroke 1994;25:1360–5.
24. Katz JN, Larson MG, Phillips CB, et al. Comparative measurement sensitivity of short and longer health status instruments. Med Care 1992;30:917–25.
25. Kravitz RL, Greenfield S, Rogers WH, et al. Differences in the mix of patients among medical specialties and systems of care: results from the Medical Outcomes Study. JAMA 1992;267:1617–23.
26. Krousel-Wood MA, Re RN. Health status assessment in a hypertension section of an internal medicine clinic. Am J Med Sci 1994;308:211–7.
27. Krousel-Wood MA, McCune TW, Abdoh A, et al. Predicting work status for patients in an occupational medicine setting who report back pain. Arch Fam Med 1994;3:349–55.
28. Kurtin PS, Davies AR, Meyer KB, et al. Patient-based health status measures in outpatient dialysis: early experiences in developing an outcomes assessment program. Med Care 1992;30(suppl):MS136–49.
29. Lansky D, Butler JBV, Waller FT. Using health status measures in the hospital setting: from acute care to "outcomes management." Med Care 1992; 30(suppl):MS57–73. Manocchia, Bayliss, Connor et al, 1998
30. McCallum J. The SF-36 in an Australian sample: validating a new, generic health status measure. Aust J Public Health 1995;19:160–6.
31. McDowell I, Newell C. Measuring Health: A Guide to Rating Scales and Questionnaires. 2nd ed. New York: Oxford University Press, 1996.
32. McHorney CA, Ware JE. Construction and validation of an alternate form general mental health scale for the Medical Outcomes Study Short Form 36-Item Health Survey. Med Care 1995;33:15–28.
33. McHorney CA, Ware JE, Lu JFR, Sherbourne CD. The MOS 36-Item Short-Form Health Survey (SF-36): III. tests of data quality, scaling assumptions and reliability across diverse patient groups. Med Care 1994;32:40–66.
34. McHorney CA, Ware JE, Raczek AE. The MOS 36-Item Short-Form Health Survey (SF-36): II. psychometric and clinical tests of validity in measuring physical and mental health constructs. Med Care 1993;31:247–63.
35. McHorney CA, Ware JE, Rogers W, et al. The validity and relative precision of MOS short- and long-form health status scales and Dartmouth COOP charts: results from the Medical Outcomes Study. Med Care 1992;30(suppl):MS253–65.
36. Medical Outcomes Trust. Medical Outcomes Trust: Improving Medical Outcomes from the Patient's Point of View. Boston, MA: Medical Outcomes Trust, 1991.
37. Meyer KB, Espindle DM, DeGiacomo JM, et al. Monitoring dialysis patients' health status. Am J Kidney Dis 1994;24:267–79.
38. Nerenz DR, Repasky DP, Whitehouse FW, et al. Ongoing assessment of health status in patients with diabetes mellitus. Med Care 1992;30(suppl): MS112–24.
39. Okamoto LJ, Noonan M, DeBoisblanc BP, et al. Fluticasone propionate improves quality of life in patients with asthma requiring oral corticosteroids. Ann Allergy Asthma Immunol 1996;76:455–61.
40. Osterhaus JT, Townsend RJ, Gandek B, et al. Measuring the functional status and well-being of patients with migraine headaches. Headache 1994;34: 337–43.
41. Patrick DL, Bush JW, Chen MM. Toward an operational definition of health. J Health Soc Behav 1973;14:6–21.
42. Phillips RC, Lansky DJ. Outcomes management in heart valve replacement surgery: early experience. J Heart Valve Dis 1992;1:42–50.
43. Rampal P, Martin C, Marquis P, et al. A quality of life study in five hundred and eighty-one duodenal ulcer patients. Scand J Gastroenterol 1994;29:44–51.
44. Rector TS, Ormaza SM, Kubo SM. Health status of heart transplant recipients versus patients awaiting heart transplantation: a preliminary evaluation of the SF-36 Questionnaire. J Heart Lung Transplant 1993;12:983–6.
45. Reynolds WJ, Rushing WA, Miles DL. The validation of a functional status index. J Health Soc Behav 1974;15:271–89.
46. Rumsfield JS, MaWhinney S, McCarthy M, et al. Health-related quality of life as a predictor of mortality following coronary artery bypass graft surgery. JAMA 1999;218:1298–303.
47. Shiely J-C, Bayliss MS, Keller SD, et al. SF-36 Health Survey Annotated Bibliography: The First Edition (1988–1995). Boston, MA: The Health Institute, New England Medical Center, 1996.
48. Stewart AL, Greenfield S, Hays RD, et al. Functional status and well-being of patients with chronic conditions: results from the Medical Outcomes Study. JAMA 1989;262:907–13.
49. Stewart AL, Hays RD, Ware JE. The MOS Short-form General Health Survey: reliability and validity in a patient population. Med Care 1988;26:724–35.
50. Stewart AL, Ware JE. Measuring Functioning and Well-Being: The Medical Outcomes Study Approach. Durham, NC: Duke University Press, 1992.
51. Stewart AL, Ware JE, Brook RH. Advances in the measurement of functional status: construction of aggregate indexes. Med Care 1981;19:473–88.
52. Sullivan M, Karlsson J, Ware JE. The Swedish SF-36 Health Survey: I. evaluation of data quality, scaling assumption, reliability and construct validity across general populations in Sweden. Social Science Medical 1995;41:1349–58.
53. Tsai C, Bayliss MS, Ware JE. SF-36 Health Survey Annotated Bibliography: Second Edition (1988–1996). Boston, MA: Health Assessment Lab, New England Medical Center, 1997.
54. Vickrey BG, Hays RD, Graber J, Rausch R, Engel J, Brook RH. A health-related quality of life instrument for patients evaluated for epilepsy surgery. Med Care 1992;30:299–319.
55. Viramontes JL, O'Brien B. Relationship between symptoms and health-related quality of life in chronic lung disease. J Gen Intern Med 1994;9:46–8.
56. Wagner AK, Keller SD, Kosinski M, et al. Advances in methods for assessing the impact of epilepsy and antiepileptic drug therapy on patients' health-related quality of life. Qual Life Res 1995;4:115–34.

57. Wagner PJ, Phillips W, Radford M, et al. Frequent use of medical services: patient reports of intentions to seek care. Arch Fam Med 1995;4:594–9.
58. Ware JE. How to Score the Revised MOS Short-Form Health Scale (SF-36). Boston, MA: The Health Institute, New England Medical Center Hospitals, 1988.
59. Ware JE. Scales for measuring general health perceptions. Health Serv Res 1976;11:396–415.
60. Ware JE. The status of health assessment 1994. Annu Rev Public Health 1995;16:327–54.
61. Ware JE. Tech Notes: Confidence intervals for individual scores. Medical Outcomes Trust Bulletin 1994;2:3.
62. Ware JE, Gandek B, the IQOLA Project Group. The SF-36 Health Survey: development and use in mental health research and the IQOLA Project. Int J Ment Health 1994;23:49–73.
63. Ware JE, Gandek B, Keller SD, the IQOLA Project Group. Evaluating instruments used cross-nationally: methods from the IQOLA Project. In: Spilker B, ed. Quality of Life and Pharmacoeconomics in clinical trials. 2nd ed. New York: Raven Press; 1996:681–92.
64. Ware JE, Keller SD, Gandek B, et al., and the IQOLA Project Group. Evaluating translations of health status questionnaires: methods from the IQOLA Project. Int J Technol Assess Health Care 1995;11:525–51.
65. Ware JE, Kosinski M. The SF-36 Health Survey (Version 2.0) Technical Note. Boston, MA: Health Assessment Lab, September 20, 1996 (updates September 27, 1997).
66. Ware JE, Kosinski M, Bayliss MS, et al. Comparison of methods for the scoring and statistical analysis of SF-36 health profiles and summary measures: summary of results from the Medical Outcomes Study. Med Care 1995; 33(suppl):AS264–79.
67. Ware JE, Kosinski M, Keller SD. SF-12: How to Score the SF-12 Physical and Mental Health Summary Scales. 2nd ed. Boston, MA: The Health Institute, New England Medical Center; 1995.
68. Ware JE, Kosinski M, Keller SK. SF-36 Physical and Mental Health Summary Scales: A User's Manual. Boston, MA: The Health Institute, 1994.
69. Ware JE, Kosinski M, Gandek BG, et al. The factor structure and content of the SF-36 Health Survey in 10 countries: Results from the International Quality of Life Assessment (IQOLA) Project. J Clin Epidemiol 1998;1159–65.
70. Ware JE, Sherbourne CD. The MOS 36-Item Short-Form Health Survey (SF-36): I. Conceptual framework and item selection. Med Care 1992;30:473–83.
71. Ware JE, Snow KK, Kosinski M, et al. SF-36 Health Survey Manual and Interpretation Guide. Boston, MA: New England Medical Center, The Health Institute, 1993.
72. Weinberger M, Samsa GP, Hanlon JT, et al. An evaluation of a brief health status measure in elderly veterans. J Am Geriatr Soc 1991;39:691–4.
73. Wells KB, Burnam MA, Rogers W, et al. The course of depression in adult outpatients: results from the Medical Outcomes Study. Arch Gen Psychiatry 1992;49:788–94.

Address reprint requests to

John E. Ware, Jr., PhD
QualityMetric, Inc.
640 George Washington Highway
Lincoln, RI 02865
E-mail: jware@qmetric.com