Health Status and Outcomes Assessment Tools

John E. Ware, Jr, PhD[1], and James E. Dewey, PhD[2]

[1]Dr. Ware is the Director of the Health Assessment Lab (HAL) in Boston and is on the faculty at Tufts University School of Medicine and the Harvard School of Public Health. He is a member of the Institute of Medicine and serves on the Board of Directors of the Medical Outcomes Trust. He is also the founding President of QualityMetric, Inc., a Lincoln, Rhode Island, company that is using computer technology and the latest psychometric advances to develop the next generation of patient-based tools for monitoring health outcomes. Quantifying health outcomes has been Dr. Ware's research focus for nearly 30 years, including 14 years at the RAND Corporation as Principal Investigator of the Medical Outcomes Study, and ten years as Senior Scientist at The Health Institute at New England Medical Center. He is also Principal Investigator of the International Quality of Life Assessment (IQOLA) Project and received the Association for Health Services Research (AHSR) "Article of the Year" Award in 1993 and "Distinguished Investigator" award in 1994.

[2]Dr. Dewey is a co-founder of QualityMetric, Inc. and past president of the Society of Prospective Medicine. He founded Response Technologies, Inc., a two-time **Inc.** *magazine 500 award winning company. His doctoral research at Purdue University validated the Centers for Disease Control HRA in the early 1980s. He then administered the Rhode Island Department of Health's Wellness Check program, the nation's first automated health risk appraisal. He holds a US patent for the development of an expert data collection analysis and response system and method. He has currently joined forces with Dr. John Ware in working on a breakthrough that uses modern psychometric principles and computer technology to create very brief health assessments with the precision necessary for individual patient-level decision-making.*

(Portions of this chapter were taken from Ware JE, Davies AR. **Monitoring Health Outcomes from the Patients' Point of View: A Primer.** *Kenilworth, NJ: Integrated Therapeutics Group; 1996, reprinted with permission.)*

Corresponding author: John Ware, Executive Director, The Health Assessment Lab, President an d CEO, QualityMetric, Inc., 640 George Washington Highway, Lincoln, RI 02865; phone: 401.334.8800; email: JWARE@QMETRIC.COM

Introduction

In the era of managed care and increasing health care consumerism, more and more emphasis is being placed on viewing health, health status, and outcomes from the patient point-of-view. "Health status" and "outcomes" have different meanings depending upon people's perspective. For clinicians, both would likely be defined using various clinical, biological, or physiological terms. When asking patients, however, we get entirely different descriptions. This chapter will provide a brief overview of this perspective.

Quantity and quality are two dimensions that characterize life. Average life expectancy, mortality rates, and other such indicators represent the quantity of life. Yet in developed countries, these indicators make little contribution to understanding quality of life—how well people live.

It has become fashionable to consider all indicators of health, not merely those of biologic functioning, as reflecting quality of life. This practice offers a shorthand method of referring to a collection of concepts both broader and more qualitative than the usual clinical endpoints. As traditionally defined, however, quality of life is a much broader concept than health.[1,2] Quality of life encompasses such factors as standard of living, quality of housing and neighborhood, job satisfaction, and health.[1]

Most contemporary definitions of health mention both functioning and well-being, which together reflect health-related quality of life.[3] The World Health Organization defines health as a "state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity."[4] Dictionary definitions of health emphasize its physical and mental dimensions, and refer to the body and bodily needs and to emotional and intellectual status. Health also connotes completeness—nothing is missing from the person—and proper function—all is working efficiently. Well-being, including soundness and vitality, also appears in dictionary definitions,[3] as does the concept of disability in social and role functioning.

A Comprehensive Definition of Health Outcomes

In the late 1990s, a comprehensive definition of health includes three types of measures: biological functioning, general health, and disease-specific symptoms and problems.

Biological measures of health status—at the center of Figure 1—focus on the physiology and functioning of organs and organ systems (or subsystems); these measures are commonly used during diagnosis and in monitoring treatment effects. In many cases, a given biological measure is closely associated with a particular disease or condition (e.g., blood pressure and hypertension, glycosylated hemoglobin and diabetes, serum creatinine and kidney disease).

General measures assess health-related aspects of quality of life that are relevant regardless of individual characteristics (e.g., age, gender, disease, or condition). Rather than being disease-, condition-, or procedure-specific, most current general measures of physical and mental health and social and role functioning (ability to do normal work activities) are generic, and reflect the full range of health states, from limitations and disability to well-being. State-of-the-art general health measures capture at least four concepts: physical function, mental health, social and role function, and general health perceptions.

Measures of physical function commonly focus on limitations, disabilities, capacities, and abilities in those bodily behaviors common to everyday life (e.g., self-care, walking, running). Others reflect bodily pain. Measures of mental health focus chiefly on frequency and intensity of psychological distress; increasingly, they also include assessments of psychological well-being and cognitive functioning. Social and Role Function measures capture the frequency and nature of social contacts and relationships, and the capacity to engage in activities common to a given role (e.g., employment, school). Increasingly, these measures capture the impact of physical and mental health problems on social functioning and role performance.

A comprehensive and valid health survey also must reflect the values or preferences of the individual. Who else, after all, is more qualified to evaluate current health status or expectations for health in the future? General health perception measures are the most generic of all health status measures, reflecting personal beliefs and perceptions about health overall, rather than its distinct physical, mental, social, or role aspects. Such evaluations provide a good overall health summary and reflect the impact of specific symptoms and other health states experienced but not captured explicitly by measures in the other three categories. They can be thought of as direct measures of health-related quality of life.

While generic measures of health status provide an important "common denominator" for defining health outcomes across diseases (and in terms of particular relevance to the individual patient), they do not always provide enough detail or reflect all the aspects of health affected by a given disease or condition. Because most outcomes measurement focuses by design on defined clinical groupings, generic measures typically are supplemented by ***disease- or condition-specific measures***. Like biological measures, these latter measures are commonly specific to a single disease or condition. Unlike them, they capture the patient's perspective regarding some aspect of that condition or its effect on general health. For example, disease-specific measures may focus on particular functions affected by the disease (e.g., finger mobility or pain in arthritic patients), or the experience of symptoms associated with a disease or its treatment (e.g., nausea and vomiting in cancer patients). Others may be measures of the general health concepts identified above, attributing any limitation, problem, or disability reported to the particular disease, condition, or procedure under study (e.g., bodily pain due to back problems, limitations in role functioning due to dialysis.).

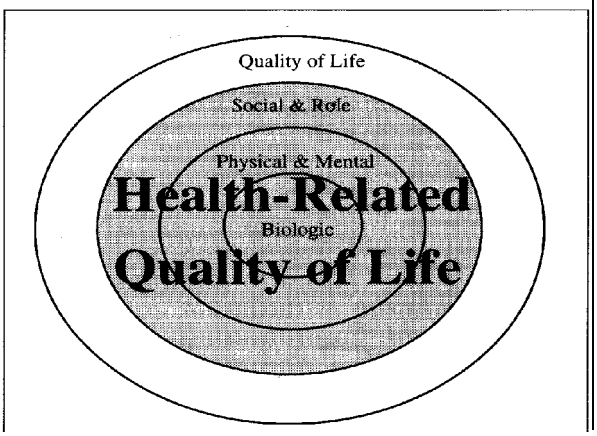
Measuring General Health Outcomes

General health outcomes are measured by giving patients carefully constructed surveys. Questions in these surveys ask patients about different aspects of their health, and offer them defined response options. The standardization of questions about health status, the response choices offered, and the scoring of these responses make the interpretation of survey results possible. Until the late 1980s and early 1990s, surveys had not been used widely in clinical practice or clinical research for two reasons: 1) their length made it impractical for patients to complete them in most clinical settings, and 2) their results could not be easily interpreted and used by clinicians.

Short-Form Surveys

In a busy practice setting, the length of the health survey used must take into account the amount of time the patient has to complete the questionnaire. Results from experience and experimentation suggest that an "ideal" survey for clinical use can be completed in 10

Figure 1. Quality of Life Concepts



minutes or less.[5,6] For the average patient, this means a survey of some 40 to 60 items. In addition to length, the actual number of items used and the distribution of items across health concepts are critical in determining the acceptability and usefulness of a health survey in clinical practice.

During the past several years, considerable strides have been made in developing and testing short-form health surveys. Most prominent among the short-forms currently used in clinical practice settings and clinical research are the 9-item Dartmouth COOP Charts,[5] the Duke Health Profile,[7] and the Medical Outcomes Study SF-20, SF-36 and SF-12 Health Surveys, which are, respectively, 20-item, 36-item, and 12-item short-forms.[6–9]

Work on the development of short-form health surveys also has examined the tradeoffs involved in using fewer vs. more items to assess each health concept. Analyses have demonstrated that a well-constructed multi-item scale, even with as few as 5 to 10 items, is more useful in detecting differences between groups of patients, measuring change over time, and predicting subsequent medical expenditures—than the best single-item measure of the same concept.[10,11]

Depending on the purpose, however, even single-item health measures can provide valuable information. In studies of the Dartmouth COOP Charts, which are single-item measures, primary care physicians who used them reported that they provided new, clinically relevant information for 30% of patients, and led to changes in patient management for 10-15% of patients.[5] When used to identify mental health problems in primary care settings, the best single item from the 5-item Mental Health scale in the SF-36 detected "nearly three-fourths of those with a diagnosable psychiatric disorder with a false-positive rate of only about 5%."[12]

Features of Measures Important to Interpretation

While many attributes of scale scores—the quantitative results of a health status survey—affect interpretation of those results, four attributes are particularly important and should be considered prerequisites to their interpretation and use: 1) reliability; 2) range of measurement; 3) number of levels of measurement; and 4) confidence intervals.

Reliability. Reliability has to do with how confident we can be that an observed score is the "true" score. Before a scale score can be interpreted, its reliability must be established. It would be inefficient to try to understand the meaning of a number that cannot be reproduced. Reliability is an issue of consistency upon repetition. We can be confident that a score is the true score if we obtain the same score again and again upon repeated assessment of an unchanging patient (e.g., when repeat sphygmomanometric (blood pressure) measurements during a visit are averaged.

Several methods are used to estimate the reliability, or repeatability, of health status scores. The most common are: a) treating the survey questions (items) in a scale as repeated measures of the same concept and estimating reliability from the relationships among them (internal-consistency method); b) determining the association between scores collected from a clinically stable sample at two times (test-retest correlation); and c) assessing the proportion of scores at follow-up that are statistically different from those at baseline (confidence interval method). In all cases, the reliability estimate is expressed as a proportion ranging from 0.0 (i.e., providing a completely "noisy" signal, or score) to 1.0 (a "true" score).

Range of Measurement. Health states range from very poor, including disability and dysfunction, to very good, including high levels of behavioral functioning and well-being. The wider the range of health states captured by the items in a survey, the broader the range of measurement. Restrictions on the range of measurement represented across items in a general health survey, and across a multi-item scale of a single health dimension, will set limits on the ability of scores to make distinctions among patients and to detect changes within and among patients over time.

Specifically, the range determines how many patients are concentrated at the top ("ceiling") and at the bottom ("floor") of the score distribution. On any given scale, those at the top cannot improve and those at the bottom cannot get worse; thus, important distinctions or changes can be missed. For example, measures of mortality have a very low "ceiling" as outcome indicators for cardiac surgery, because 98% of patients are placed in the same survival category. In the past, health status surveys focused on the lower end of the range—on disability, dysfunction, and limitations. Thus, in general populations and in many primary care settings, many respondents will receive perfect scores. As one solution to the problem, most modern health status surveys extend the range of measurement to include positive health states.

Levels of Measurement. The number of levels (or categories or score values) into which a health status measure can classify an individual or group also influences the scale's ability to distinguish differences and changes over time. For example, consider a measure of walking that has only two levels: 1) can walk: and 2) cannot walk. Most people in a general population and many patients would endorse the first statement and would receive the same score. Any differences in their walking ability would be missed. Only when someone became unable to walk would a change be detected. By contrast, consider a measure of walking with four levels: 1) can walk, with no difficulty; 2) can walk, with some difficulty; 3) can walk, with great difficulty; and 4) cannot walk. Those who scored "can walk" on the two-level scale would be divided among the first three levels of this more refined scale. Thus, the more levels of measurement available, the greater the possibility of distinguishing among patients and detecting change over time.

The coarseness of a health status scale refers to the size of the differences between the scale levels and the number and range of levels. These levels define categories (or scale levels) into which patients are classified. Range and level of measurement together determine the coarseness of the scale; the coarseness, in turn, affects the scale's usefulness in measuring health.

Confidence Intervals. As with all health status scales, the interpretation of individual patient scores must take into account the amount of "noise" in the scores they yield. The "noise level" can be quantified and displayed visually as a confidence interval (CI) around a patient's score. The size of the CI is a function both of the reliability of a score and the standard deviation of the score distribution in the population of interest. Because reliability most affects the size of confidence intervals when individual scores are interpreted, a much higher standard of reliability is required for measures to be interpreted for an individual patient as opposed to interpreting average scores for large groups of patients.[13] For example, reducing the reliability of a multi-item physical health summary measure from an actual 0.93 to a hypothetical 0.50 would increase the 95% confidence interval for an individual patient score by more than 250%.[8]

The SF-36: A Short-Form Health Status Survey

Work on shorter forms and attention to the important features of health status measurement discussed above led to the development of the SF-36 Health Status Survey—SF referring to short form, 36 to the number of items.[7,8,14] The SF-36 is one of the most widely used surveys in the world for assessing health status from the patient's point of view in clinical research and practice. It has been the subject of more than 1,000 publications and has been translated into more than 45 languages (which are spoken by nearly three-fourths of the world population).

SF-36: Health Concepts Measured

The SF-36 measures eight health concepts:

- limitations in physical functioning due to health
- limitations in usual role activities due to physical health
- bodily pain
- perception of health in general
- energy and fatigue
- limitations in social activities due to physical or emotional health
- limitations in usual role activities due to personal or emotional problems
- psychological distress and well-being

Eight SF-36 scales are scored using from two to ten items each. As discussed earlier, multi-item scales assess a broader range of attributes or activities related to the health concepts, and yield more reliable scores. Respondents answer items in each scale by selecting standardized response choices. Response choices associated with each item represent a graduated range of options (e.g., not at all, slightly, moderately, quite a

bit, extremely). The use of graduated response options in conjunction with multiple items increases the number of achievable levels in each scale, and hence the precision of measurement. In addition, one SF-36 item assesses change in health over the past year.

SF-36: Scale Descriptions

A brief description of each of the SF-36 scales appears below. Other sources provide far more information regarding the origin and selection of SF-36 items, as well as empirical evidence regarding reliability, precision, and interpretation of scores.[7,8,14]

Physical Functioning (PF). The ten items in the Physical Functioning scale ask respondents to indicate the extent to which their health limits them in performing physical activities. These ten items reflect a broad range of activities and allow detection of relatively severe to minor limitations. A three-level response continuum for each item captures both the presence and extent of physical limitation.

Role-Physical (RP). The four items in the Role-Physical scale ask respondents to what degree their physical health limits them in the kind of work or other usual activities they perform, causes them to cut down on the amount of time they spend on work or other usual activities, and causes difficulty in performing work or other usual activities. Because the items refer to work and "other regular daily activities," they are also applicable to students, to those who are retired, and to those who have more than one usual role.

Bodily Pain (BP). The two items in the Bodily Pain scale obtain assessments of the frequency of pain or discomfort and the extent of interference with normal activities due to pain.

General Health (GH). The five items in the General Health scale obtain respondents' assessments of their current health status overall, susceptibility to illness, and their expectations for health in the future. Scores from this scale provide a good summary of health status overall, and reflect the impact of specific symptoms and other health states experienced but not captured explicitly by the other scales.

Vitality (VT). The four items in the vitality scale capture changes in subjective well-being by asking respondents to indicate how frequently they experience feelings of energy and fatigue.

Social Functioning (SF). The two items in the Social Functioning scale ask respondents about the impact of either physical health or emotional problems on normal or usual social activities. Respondents are asked to indicate limitations in social function due specifically to *health*; this minimizes variation in scores that may be attributable to non-health related factors.

Role-Emotional (RE). The three items in the Role-Emotional scale ask respondents to what degree emotional problems have caused them to accomplish less in their work or other usual activities, cut down on the amount of time spent on work or other usual activities, and perform work or other activities less carefully. As with the Role-Physical scale, the items here refer both to work and "other regular daily activities," and thus are also applicable to many different roles in life.

Mental Health (MH). The five items in the Mental Health scale ask respondents to indicate how frequently they experience feelings representing the four major mental health dimensions: anxiety, depression, loss of behavioral/emotional control, and psychological well-being.

Change in Health. A single item asks respondents to rate their health in general now compared to one year ago. The accuracy of self-reported transitions in response to this item is currently under investigation.

Physical and Mental Summary Measures. The summary measures were constructed on the basis of factor analyses of correlations among the eight SF-36 scales in patient and in the general US populations. Two components were identified: physical and mental health factors accounted for 82.4% of the reliable variance in the eight scales and are easily interpreted in the general population. All eight multi-item scales are used to score each of the summaries. The three best (those with highest beta weights in the factor analysis) physical scales (PF, BP, RP) receive more weight in the physical summary, as do the three best mental scales (MH, RE, SF) in the mental summary. Because the summaries are derived by principle components analysis, they are referred to as the physical component summary (PCS) score and the mental component summary (MCS) score.

The validity of the SF-36 scales and summary measures in relation to such clinical indicators as presence or absence of disease, severity within disease category, and changes in disease-related symptoms over time has been studied extensively.[6,12,15_19]

Summaries of this evidence, pertinent to both general and patient populations, appear in interpretation manuals for the SF-36 scales[7] and summary measures.[8]

Scoring

Original scoring methods for the SF-36 scales resulted in ranges from 0 to 100, with 100 representing the best score on each scale, or the best health. Thus, functioning scales were scored so that higher scores indicated better functioning; the mental health scale was scored so that higher scores indicated greater psychological well-being; and the pain scale was scored so that higher scores indicated freedom from pain.

Widespread acceptance and use of the SF-36 led to a change in the original 0-100 scoring to improve interpretation of the scale scores. Because the scales have different population means and standard deviations, comparison between scale scores could be misleading. As such, current recommended practice is to convert the 0-100 scores, using US population means and standard deviations (SD) and linear T-score transformations, resulting in mean of 50 and SD of 10 for each of the scale scores. This norm-based approach has greatly improved the interpretation of the scoring of the SF-36, and brings the presentation of results for the eight-scale profile in line with the two summary measures.

Norm-based methods were also used to standardize the physical and mental composite summary scores. In that way, PCS and MCS also have a mean of 50 and a SD of 10 in the general US population, making direct comparison between scale and summary scores possible.

Administration

In most clinical applications, the SF-36 is self-administered by patients at the time of a physician, clinic, or hospital visit. In some clinical research applications, the SF-36 is self-administered at home using mail-out/mail-back questionnaires or through telephone interviews. While the majority of respondents can self-administer the survey in one of its formats, others may require interview administration, whether face-to-face or by telephone. The SF-36 can be used alone, or included as one part of a longer questionnaire, interview, or other protocol, depending on the purpose of data collection. To illustrate, the Medical Outcomes Study used a 70-item patient assessment survey, which included the SF-36. Self-administered at the time of an office visit, this survey required only about 10-20 minutes to complete for more than 22,000 patients in some 500 practices.[6] In these practices, physicians and staff were highly motivated to have all patients seen during a two-week period complete the assessment. Completion rates were 65% in solo practices and 74% in the better-staffed and better organized group practices. In another example, patients self-administering only the SF-36 during an outpatient hemodialysis session took as long as 20 minutes to complete the survey. This length of time reflected the fact that patients had approximately 3.5 hours (during a 4-hour dialysis session) to complete the form and few other demands on their time. Physicians asked each patient to complete the SF-36; completion rates were above 90%.[20]

In clinical settings, the SF-36 should be administered before the patient sees a provider so that the interaction between the patient and provider will not influence answers to the questionnaire. Ideally, the questionnaire also should be administered before the patient is asked other questions about health, symptoms, and concurrent illnesses. Detailed guidelines regarding administration of the survey in clinical settings can be found elsewhere.[7,8]

Available Forms

Two forms of the SF-36—standard" and "acute"—have been developed. The standard form instructs respondents to think *about the last four weeks* when answering most items; the acute form instructs respondents to think about *the last week* when answering those items. The forms are identical in all other respects. Because the standard version has been used more extensively to date, more information is available regarding its reliability and validity among different patient groups.[7,8] Normative data is available for the standard version,[7,8] and will be available for the acute version in early 1999. As the SF-36 is used increasingly in clinical trials and clinical practice, the acute version is gaining in popularity.

In 1991, the International Quality of Life Assessment Project was launched to translate, norm, and validate the SF-36 in other languages. Included in this work are translations into Danish, French, Flemish, German, Italian, Japanese, Norwegian, Spanish, and Swedish.[21] English-language adaptations also were developed for use in Australia, Canada, and the UK. Of particular interest to clinicians and clinical

investigators in the US are Chinese, Japanese, and Spanish adaptations for use in the US.

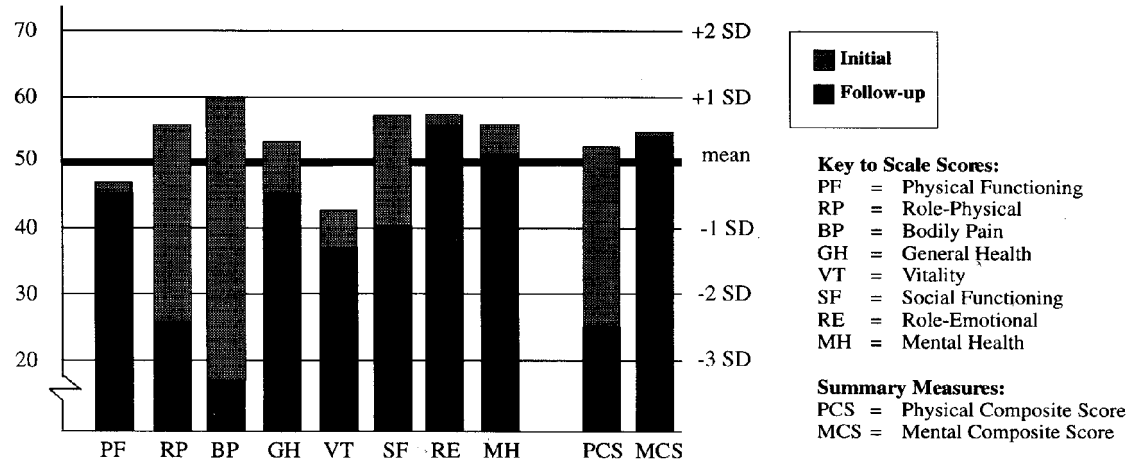
Comments from patients and clinicians using the SF-36 in various clinical research and practice settings have resulted in the development of several different formats. These include regular and large-type printed survey booklets, which recognize differences in visual acuity; various scannable formats, which permit rapid scoring and feedback; and more recently, computer-assisted administrations using touch-screen entry, touch-tone telephone or interactive voice recognition (IVR) technologies, and the Internet.

SF-36: Feedback and Interpretation

For the forms to be widely accepted by physicians, they must be shorter, use processing systems that rapidly enter and score them with a high degree of accuracy, and display their results in a user-friendly format. Figure 2 illustrates a feedback format for displaying SF-36 scale and summary scores for individual patients that is currently the standard for results reporting. Other feedback formats are illustrated and discussed elsewhere.[7,8]

Figure 2 shows norm-based scores on the eight SF-36 scales—termed the SF-36 Health Status Profile—and the two summary measures for a middle-aged employed male patient receiving outpatient hemodialysis for end-stage renal disease. The patient's first and second visits during his seventh year on dialysis are here labeled "Initial" and "Follow-up." The bar graphs for the SF-36 profile and for the PCS and MCS reveal that this patient had deteriorated substantially between visits. Significant declines (larger than the 95% confidence interval) were observed for four of the scales (Role-Physical, Bodily Pain, Social Functioning, and Mental Health)

Figure 2. SF-36 Health Status Profile: Initial and Follow-up Profiles for ESRD Patient



and for the PCS. The decline of 24 points (from 52 to 28) in PCS scores represents a decline from just below the 50th percentile to well below the 25th percentile for a 38-year-old male (based on norms for a general population).

Clinical Uses of Patient-Based Health Surveys

Standardized health surveys have the potential to become the new "laboratory tests" of medical practice.[22] Without them, it appears that patient functioning and well-being are far less likely to be discussed during a typical medical visit. In a recent survey, two-thirds to three-fourths of US adults reported that physicians rarely or never ask about the extent of the patient's limitations in performing everyday activities, even in the presence of chronic conditions.[23] As a result, practicing physicians are unaware of relatively concrete impairment manifested by observable limitations in physical, social, and role functioning.[24] Differences in severity of psychological

distress also are often not apparent to treating physicians.[25] Severely psychologically distressed patients suffering from psychiatric disorders often go unrecognized and untreated even when mental health treatment is covered by health insurance.[8] It has been suggested that more widespread use of standardized health measures may improve clinical practice.[12,26]

Standardization of longitudinal, patient-based health status assessments in both practice and research will be useful in a variety of ways:

- Ensuring that all important dimensions of functional status and well-being are considered consistently (in addition to more traditional measures of clinical endpoints);
- Detecting, tracking, and explaining changes in functional capacity and well-being over time;
- Making it possible to consider the patient's total functioning when choosing among therapies and treatment options;
- Predicting more accurately the burden and course of chronic diseases;
- Guiding the efficient use of community resources and social services.

Future Directions

Simple, reliable, inexpensive, and precise methods for measuring patients' assessments of health related quality of life have implications for clinical practice, research, health care reform, and other objectives of medicine. Several trends pointing to future directions in the use of such measures can be identified.

Increasingly important will be standardization of health outcomes measurement. Standardization—of item content, response choices, instructions, administration methods, and scoring algorithms—is necessary to achieve reproducible results that can be interpreted and compared meaningfully. In working with clinicians who are using patient-based health assessments, we have found that the need for standardization applies both to the general health surveys and to the more familiar clinical indicators of outcomes. For example, in studying their patients' outcomes following total knee replacement, three orthopedic surgeons had comparative SF-36 data but could not compare clinical outcomes because of noteworthy differences in the operational definitions of knee function used by each. After discussing and resolving these differences, they standardized their clinical outcomes assessment protocols, and can now compare outcomes and performance routinely across the patient groups.

To meet future needs, information about general health outcomes must be routinely collected and made available in the nation's health care databases. Minimum standards of comprehensive measurement should be adopted to monitor the health of the general population, evaluate the impact of health care policies, and monitor results of episodes of care.

Comprehensive monitoring is critical, as it produces information about the burden of acute and chronic disease in physical, mental, and social terms, as well as in terms of what people are able to do and how they feel. Normative data from comprehensive assessments are critical for interpretation of general health measures in clinical research and practice.

In addition to the availability of norms, several other issues will affect the use of health status assessments in everyday clinical practice. Practicality is essential. Data collection and data processing methods must permit their smooth integration into busy clinical settings, providing added value rather than disrupting care delivery. Already computers are becoming more commonplace in physician examination rooms, allowing the busy clinician to document information obtained from the patient in the form of symptoms and clinical information. It is only a matter of time before technology (e.g., the Internet, interactive voice response, touch screen computer terminals) allows direct input and reporting of health related quality of life information as well.

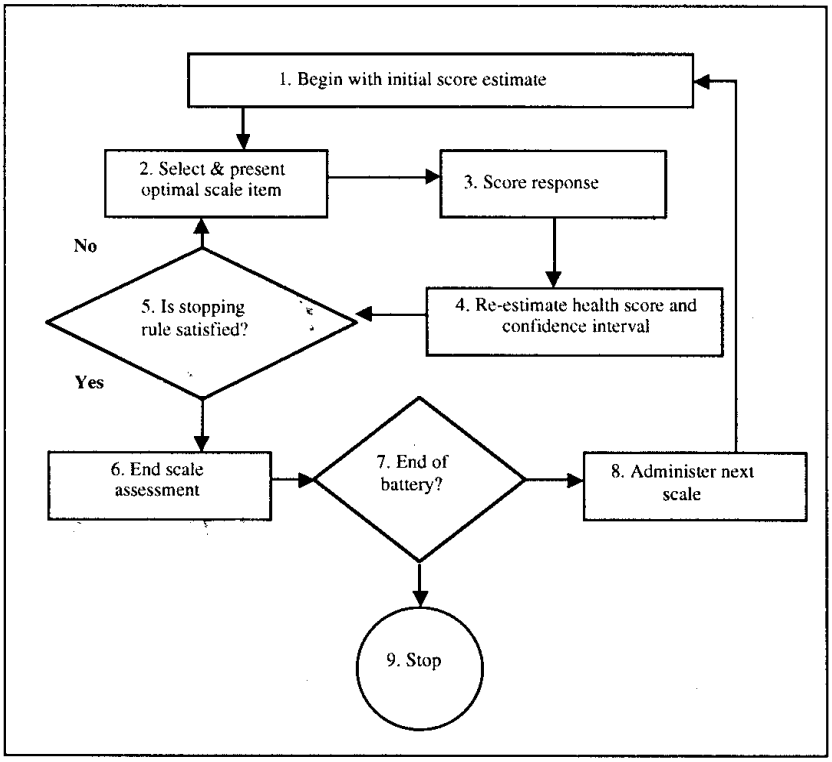
Interestingly, the reduction in length that has made general health status surveys more practical for administration in clinical settings also sets certain limits on their usefulness. The short-form health status survey was intended as a population measure, suitable for comparisons between groups, but lacking in the precision necessary for individual patient decision-making. Furthermore, current fixed-length, fixed-format instruments often represent health as the absence of limitations. As such, they suffer from "ceiling" effects with most of the general population classified at the top of the health scale.

Computer Adaptive Health Status Assessment

Precision assessment using "classical" psychometric principles and methods requires a large number of items, spanning the complete range of a particular health concept. The time and complexity to complete such a process makes wide-scale adoption unlikely. In the early 1950s, new statistical methods such as Rasch Models and Item Response Theory (IRT) evolved. These "modern" psychometrics, when combined with computer technology, have great potential for providing valid, precise, and efficient health status assessment at the individual patient level.[27]

The basic idea of this precision assessment method is that after an initial question is posed (selected with any available knowledge of the health state of the respondent), the response given triggers the selection of the next best item from a comprehensive item pool that can be used to estimate the individual's score on the health concept being measured. The response to the second item is determined, and a score and confidence interval are calculated. This process of item selection, response, score and confidence interval calculation is repeated until a pre-determined level of precision is reached (see Figure 3). In most instances, with consistent responses, this method would pinpoint the individual's score using a fraction of the items of the larger item pool. The result will be health status measurement that meets the standard of precision necessary for decision making at the individual patient level, and can be completed in an efficient and cost effective manner.

Fig 3. Logic of Computerized Adaptive Health Assessment



Adapted from Wainer, et al.[28]

The Use of Health Status Assessment in Prospective Medicine

A major advantage of a health status measure such as the SF-36, is that it has been thoroughly researched and validated, and is widely accepted by the medical community for use in assessing patient outcomes. A disadvantage, from a prospective medicine perspective, is that for the most part, the SF-36 has been used retrospectively, with little effort made to provide timely feedback to the individual or clinician. While clinician use of the SF-36 at the point-of-care has been increasing,[29] its widespread in-office use on a real-time basis remains somewhat in the future.

With a growing number of medical procedures and treatments aimed at improving quality of life, it is incredible that a process is not currently in place to accurately measure, monitor, and track progress toward that goal. To make health status assessment a standard part of every medical encounter, the availability of a very brief, yet individually precise measure, such as the computer adaptive health assessment described above, is required. With such a tool, one can easily picture a time in the not-too-distant future when no medical care or treatment decision is made without first determining the patient's health status and quality of life.

The traditional health risk appraisal (HRA) has long been used at the individual level. An entire industry of communicating health risks has grown up over the

years around the use of this tool. Effective and ethical use of HRA is the topic of several chapters in this handbook. The time has come for a blending of the best features of health status and health risk assessment. The opportunity to use standard, reliable, valid, and universally accepted health status assessment in such a manner that provides actionable feedback to both clinician and patient is the essence of what prospective medicine can and should be in the future.

References

1. Campbell A, Converse PE, Rodgers WL. *The Quality of American Life: Perceptions, Evaluations, and Satisfactions*. New York, NY: Russell Sage Foundation; 1976.
2. Ware JE. Comments on the use of health status assessment in clinical settings. *Medical Care.* 1992;30(5): MS205-MS209.
3. Ware JE. Standards for validating health measures: definition and content. *Journal of Chronic Diseases.* 1987;40(6):473-480.
4. World Health Organization. World Health Organization constitution. *Basic Documents*. Geneva, Switzerland: World Health Organization; 1948.
5. Nelson EC, Wasson J, Kirk J, et al. Assessment of function in routine clinical practice: description of the COOP chart method and preliminary findings. *Journal of Chronic Disease.* 1987;40(Suppl 1):55S-63S.
6. Stewart AL, Ware JE. *Measuring Functioning and Well-Being: The Medical Outcomes Study Approach*. Durham, NC: Duke University Press; 1992.
7. Parkerson GR, Broadhead WE, Tse C-KJ. The Duke Health Profile: a 17-item measure of health and dysfunction. *Medical Care*. 1990;28(11):1056-1072.
8. Ware JE, Snow KK, Kosinski M, Gandek B. *SF-36 Health Survey Manual and Interpretation Guide*. Boston, MA: The Health Institute; 1993.
9. Ware JE, Kosinski M, Keller SK. *SF-36 Physical and Mental Health Summary Scales: A User's Manual*. Boston, MA: The Health Institute; 1994.
10. Manning WG, Newhouse JP, Ware JE. The status of health in demand estimation: beyond excellent, good, fair, and poor. In: Fuchs VR, ed. *Economic Aspects of Health*. Chicago, IL: University of Chicago Press; 1982.
11. McHorney CA, Ware JE, Rogers W, Raczek A, Lu JFR. The validity and relative precision of MOS short- and long-form health status scales and Dartmouth COOP charts: results from the Medical Outcomes Study. *Medical Care.* 1992;30(Suppl 5):MS253-MS265.
12. Berwick DM, Murphy JM, Goldman PA, Ware JE, Barsky AJ, Weinstein MC. Performance of a five-item mental health screening test. *Medical Care.* 1991;29(2):169-176.
13. Guilford JP. *Psychometric Methods*. New York, NY: McGraw-Hill; 1954.
14. Ware JE, Sherbourne CD. The MOS 36-Item Short-Form Health Survey (SF-36): I. conceptual framework and item selection. *Medical Care.* 1992;30(6):473-483.
15. Stewart AL, Hays RD, Ware JE. The MOS Short-form General Health Survey: reliability and validity in a patient population. *Medical Care.* 1988;26(7):724-735.
16. Nelson EC, Landgraf JM, Hays RD, Wasson JH, Kirk JW. The functional status of patients: how can it be measured in physicians' offices? *Medical Care.* 1990; 28(12):1111-1126.
17. 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. *Medical Care.* 1993;31(3):247-263.
18. Phillips RC, Lansky DJ. Outcomes management in heart valve replacement surgery: early experience. *Journal of Heart Valve Disease*. 1992;1(1):42-50.
19. Wells KB, Burnam MA, Rogers W, Hays R, Camp P. The course of depression in adult outpatients: results from the Medical Outcomes Study. *Archives of General Psychiatry.* 1992;49:788-794.
20. Kurtin PS, Davies AR, Meyer KB, DeGiacomo JM, Kantz ME. Patient-based health status measures in outpatient dialysis: early experiences in developing an outcomes assessment program. *Medical Care.* 1992;30(Suppl 5): MS136-MS149.
21. Ware JE, Gandek B, the IQOLA Project Group. The SF-36 Health Survey: development and use in mental health research and the IQOLA Project.

International Journal of Mental Health. 1994;23(2):49-73.
22. Deyo RA, Carter WB. Strategies for improving and expanding the application of health status measures in clinical settings: a researcher-developer viewpoint. *Medical Care.* 1992;30(Suppl 5):MS176-MS186.
23. Schor EL, Lerner DJ, Malspeis S. Physician's assessment of functional health status and well-being: the patient's perspective. *Archives of Internal Medicine.* 1995; 155(3):309-14.
24. Rubenstein LV, Calkins DR, Young RT, et al. Improving patient function: a randomized trial of functional disability screening. *Annals of Internal Medicine.* 1989; 111(10):836-842.
25. Wells KB, Stewart A, Hays RD, et al. The functioning and well-being of depressed patients: results from the Medical Outcomes Study. *Journal of the American Medical Association.* 1989;262(7):914-919
26. Health and Public Policy Committee, American College of Physicians. Comprehensive functional assessment for elderly patients. *Annals of Internal Medicine.* 1988; 109:70-72.
27. Bjorner JB, Ware JE. Using modern psychometric methods to measure health outcomes. *Medical Outcomes Trust Monitor*. 1998;3(2):12-16.
28. Wainer H, Dorans NJ, Flaugher R, et al. *Computerized Adaptive Testing: A Primer.* Hillsdale, NJ: Lawrence Erlbaum Associates; 1990.
29. Bush D. The use of health status assessment in a clinical practice. *Proceedings of the 33rd Annual Meeting of the Society of Prospective Medicine.* Sewickley, PA: The Society of Prospective Medicine; 1997.