Using Patients' Reports To Evaluate Medical Outcomes

Background Paper 1

SUMMARY

Most medical treatment is intended to improve patients' ability to function and their sense of well-being. Information about these outcomes can usually be supplied most accurately by the patients themselves.

Traditionally, medical conditions have been defined and treatments evaluated primarily through the results of diagnostic tests and clinical observation, but many studies of the outcomes of medical treatments now also routinely include protocols for asking patients questions about their health and well-being. Some outstanding examples of this phenomenon are:

- *the Medical Outcomes Study, which used a single set of measures to assess functioning and well-being of patients with any of six medical conditions; and*
- *the Patient Outcomes Research Teams and similar efforts, in which the researchers study a single medical condition and the outcomes of its treatment in considerable detail, including patient-reported health and well-being.*

The instruments used to measure these characteristics may be specifically tailored to the particular medical condition of interest. Or, they may be general measures of health-related quality of life that provide comprehensive views of the states of the patients' health at various points during the course of the treatment.

Health-related quality of life refers to those aspects of living that are affected by patients' medical conditions and to their finetioning and perceived well-being. Most survey instruments designed to measure health-related quality of life include questions related to four aspects: functional ability, perceived health, psychological well-being, and role limitations.

by

Floyd J. Fowler

University of

Massachusetts

Boston, MA

One way to analyze the effects of a particular treatment on patients' health is to describe the treatment results separately with respect to each of these aspects. The Sickness Impact Profile (SIP) and the RAND 36-item Health Survey (SF-36) are examples of instruments to measure health-related quality of life and at present are probably the major tools used to gather data from patients for this purpose in the United States. The SIP is perhaps the most comprehensive instrument for describing the effects of medical conditions on people, while the SF-36 attempts to strike a balance between comprehensiveness, validity, and parsimony.

Another approach to measuring health-related quality of life is to combine the ratings for all the various components of well-being into a single number that serves as a summary of the overall quality of life. The Quality of Well-Being Scale is probably the best-known example in the United States of the effort to produce quantitative summaries of people's health status.

Properly designed measures using patients' reports have proved as reliable and valid for describing the effects of medical conditions and treatments on patients as many other commonly accepted tests. Although the results of diagnostic tests and results based on patients' reports are difficult to compare directly, studies of diagnostic tests and of measurements taken in clinical settings almost always reveal considerable error across tests. The reliability and validity of measures based on patients' reports also vary, but the evidence is clear that measurement of medical conditions or health status, based on properly designed and evaluated questions, can be as reliable and valid as other measurements done in the clinical sciences.

At least three major conceptual and methodological challenges remain for researchers and users of patients' reports on medical outcomes:

- *How should prospective and retrospective designs be modified to ensure accurate measurements of the effects of treatment?*
- *How should researchers collect information about the results that would have been expected had a particular treatment not been given?*
- *How should the effects of treatment be calibrated to facilitate comparisons across treatments and conditions?*

There is a clear need for better understanding of how best to conduct studies based on patients' reports so that they lead to valid conclusions, and how best to assess the significance of the results. Nonetheless, this tool is already a very useful one that produces considerable knowledge that neither patients nor researchers have had before. ■

Although the saving of lives may provide evidence of the value of treatments for serious ailments such as strokes and heart attacks, only a moderate amount of the medical care delivered in the United States is intended to prevent death. Most treatment is meant to improve patients' functioning or well-being.

Even the reason for performing most common hospital surgical procedures is not to save lives, at least in the short term. The conditions treated by back surgery, for instance, are virtually never life-threatening. Fewer than perhaps 10 percent of hysterectomies and a similar proportion of surgical procedures to treat benign prostate disease are performed on patients whose lives are at risk (48,55,78). Outpatient surgical procedures such as cataract surgery, the most common procedure covered by Medicare, fall into the same category. Even such a major procedure as coronary artery bypass graft surgery is performed as often to reduce angina symptoms as to save lives (2).

Compared with surgery, ambulatory care is sought even less often for life-threatening conditions. Most patients visit doctors for checkups, for acute but self-limiting conditions (such as respiratory infections), or for other nonfatal conditions (such as back pain and arthritis) (79).

Relieving symptoms is the goal of most common surgical procedures; ruling out more serious conditions and providing diagnoses that can lead to symptomatic relief are the goals of most ambulatory visits. Thus, to a large extent, the criteria justifying medical treatments rest on how the treatments affect the way patients feel or what patients can do. Ascertaining the value of the treatment, then, requires information about patients' perceptions of their well-being.

Patients' own reports of their well-being are vital to studies of medical care for at least two reasons. First, the studies often require information that only the patients can report well. When comparisons have been made, physicians have usually been found to be poor reporters of patients' symptoms or experiences in such diverse cases as enlarged prostates and toxic reactions to cancer treatment (1 1,62).

Second, some of the key information needed from patients—their perceptions, emotional responses, preferences, and values—is subjective. There is widespread agreement that no one can reliably report such things for another person (75). Studies comparing reports by individuals themselves with reports from proxies suggested that the less observable the characteristic, the less likely others can report it accurately (24,50,66,76). Thus, although good studies of the outcomes of medical treatment gather data from various sources, many also rely on accurate measurements from patients' reports.

Prior to the mid-1980s, few studies were designed to document the benefits of treatment from the patient point of view. Studies of medical outcomes tended to focus primarily on short-term risks, such as death and strokes, and on rehospitalization. If the broader benefit to patients was assessed at all, it was usually based on ratings by physicians.

Researchers conducting a meta-analysis of the literature published between 1964 and 1990 on surgery for lumbar spinal stenosis, for example, found 74 journal articles that purported to provide information about outcomes from laminectomy, but only 61 percent of the articles reported the prevalence of leg or back pain, the main reason for which the surgery is done (74). The analysis found no randomized trials comparing surgery and conservative treatment, and almost nothing was published on the response of spinal stenosis patients to conservative treatment. Most important, the researchers found that it was often impossible to tell for certain who rated the outcomes, but that it almost always appeared to be the surgeon, not the patient, who was describing the benefits of the surgery.

The poor quality of the data undoubtedly reflected the low priority placed on documenting the value of medical treatment. Medical treatment was presumed to be worthwhile if physicians, based on their training and clinical experience, thought it would be of value. Studies were concerned chiefly with whether complications arose and with how they could be minimized. A further limitation derived from the fact that studies using survival and short-term complications as measures of outcomes document only the risks, not the benefits, of treatment. To document the benefits of treatment, accurate information about patients' health and well-being must be collected from them in standardized ways.

CURRENT APPLICATIONS

Examples of the Use of Patients' Reports in Research

To fill the gaps in the medical literature, several recent studies have attempted to measure the effect of medical treatment by asking patients questions. Of particular note are the Medical Outcomes Study (69) and the work of the Patient Outcome Research Teams (PORTS, described below) and other related research studies that focus on the health outcomes associated with particular medical conditions.

Medical Outcomes Study

The-Medical Outcomes Study (69) is a good example of the current approach to studying the effects of treatments. In that study, patients who had any of six different conditions were recruited in physicians' offices throughout the United States. The patients filled out questionnaires about their

health at the time they were first contacted; they then provided comparable data periodically so that changes could be measured. One of the important distinctive features of the Medical Outcomes Study was that a single set of measures (a predecessor to the SF-36, described below) was used in assessing overall functioning and well-being across all conditions (70). As a result, researchers could make three kinds of comparisons, each with its own value:

- how patients with the same health condition fared over time under different treatment protocols,
- how the lives of patients with different conditions were affected by those conditions, and
- how the benefits of treatments compared across conditions.

The Medical Outcomes Study was probably the first study that permitted all three of these types of analyses.

The PORTS: The Example of Lower Back Pain

The PORTS, interdisciplinary research teams funded by the Agency for Health Care Policy and Research, have been significant contributors to the development of measures of patient functioning and well-being. Each PORT studies the outcomes of medical care and treatment for one of 14 common medical conditions. The PORT study of lower back pain provides an example of how these research efforts are using a patient-oriented approach to evaluating medical outcomes.

Because the main goal in treating lower back problems is to reduce pain in the back and legs, patients are asked to describe the frequency and intensity with which they experience pain in these areas. One simple, straightforward analysis using these data is to determine the extent to which reports of back pain changed for better or worse over time with and without treatments (23). Back pain is significant because it can affect functioning, self-perceived health, and psychological well-being.

Deyo and his associates are conducting a study of back pain in cooperation with orthopedists and neurosurgeons in Maine. Patients who are being treated for back pain are asked by their physicians to participate. The physicians and patients together decide on the treatment to be used; the study does not affect the decision. Regardless of whether patients opt for surgery or nonsurgical treatment, the results are monitored.

Patients complete baseline questionnaires at the time of enrollment. The questionnaires cover the character and frequency of the back pain, the effect of the back pain, and the overall functioning and well-being of the patient. Physicians also complete forms describing the results of initial tests and the details of the treatment, but the patients' answers to the questions at 3, 6, and 12 months after enrollment are the main measures of the outcomes.

Other Outcomes Studies: *Indications for Hysterectomy*

In addition to the PORTS, other health researchers have been studying the outcomes associated with particular medical conditions and procedures using patients' reports.

Researchers recently completed a similar study of women who had conditions—such as excessive bleeding, abnormal pain, or large fibroids-that would make them candidates for hysterectomy (15,16). Whether they elected to be treated surgically or nonsurgically, the women completed questionnaires regarding their symptoms, including the frequency and intensity of their pain and bleeding. Both the conditions and the treatments have been reported to affect energy, sexual functioning, bowel functioning, frequency of urination, hot flashes, and anxiety level, so specific questions were included (either adapted from other survey instruments or newly designed) to monitor the patients' experiences in each of these areas. Other questions measured the women's general well-being, psychological well-being, perception of their health, and role limitations.

The protocol was very similar to that of the back study. Patients filled out questionnaires over the course of a year. Analyses evaluated the progress of the initial symptoms, the appearance of new problems, and the reported general function-

ing and well-being of patients who had been treated with or without surgery.

Six general characteristics of the Medical Outcomes Study and the studies of lower back pain and indications for hysterectomy mark important departures from most previous studies on the effects of medical treatments:

1. Combinations of data from patients and from medical records or physicians were used to describe the patients' initial condition and treatment.
2. Patients' reports were the primary measures of the effects of treatment.
3. Measures of patients' status were comprehensive, including changes in condition, various possible complications of treatment, and multiple measures of overall functioning and well-being.
4. Patients were followed for relatively long periods of time—a year in the lower back pain and hysterectomy studies, and several years in the Medical Outcomes Study.
5. Although patients were not assigned to different treatments as part of the protocol, the studies included patients treated in various ways, so that there was a context within which to evaluate the results of individual treatment approaches.
6. Numerous physicians who were in general community practice participated, thereby making results more likely to be representative than if the studies had been done only in university medical centers.

Role in Evaluating Effects of Medical Treatments

The role of patients' reports in evaluating the outcomes of medical care for a particular condition is to better understand the treatment effects on that condition, and to gain a broader understanding of the effects of care on patients' functioning and health-related quality of life overall.

Better understanding a treatment's effects on a medical condition has three components:

1. Assessing the characteristics of the condition. Traditionally, medical conditions have been defined through diagnostic tests and clinical observation. In some cases, however, patients' reports are needed in order to calibrate the severity of a condition. In other instances, patients' reports actually form the basis for defining the condition and its severity.
2. Measuring treatment complications. The value of treatment depends in part on whether the treatment has any negative consequences. Even when the treatment is aimed at saving lives or reducing strokes or heart attacks, the benefits must often be weighed against the risks of complications from the treatment.
3. Understanding how the condition affects patients' lives. Assessing the full value of a medical treatment requires understanding not only how a treatment affects a condition and what unwanted complications the treatment causes, but how much the condition affects patients' lives.

Assessing the Characteristics of the Condition

An example of a condition that is best measured using patients' reports is benign prostatic hyperplasia (BPH). Men's prostates tend to enlarge with age. As a consequence of this condition, some men experience a narrowing of the urethra, which obstructs urinary flow and produces such symptoms as frequent urination and difficulty in starting urination. Physicians can determine the size of the prostate gland through palpation and imaging; they can observe evidence of obstruction with cystoscopy; they can ascertain the rate at which urine flows and measure the extent to which the bladder completely empties after voiding. None of these physiological or clinical measures, however, correlates well with how patients experience symptoms or with the frequency of their symptoms (1,3,5,60).

From a medical point of view, there is no intrinsic reason to improve the rate at which urine flows, to reduce the obstruction that appears in a cystoscopy, or to make a prostate smaller. Although large post-voiding residual volumes of urine can lead to urinary-tract infections or to upper-urinary-tract pressure, which can cause deteri-

oration of the bladder or affect renal function, such problems probably affect no more than 10 percent of men who undergo prostate surgery. Of the 350,000 men who have prostate surgery each year, about a quarter do so because of acute retention, whereas well over half do so to reduce their symptoms (55). For the latter group, the best indicators of the condition's severity are the patients' reports about their symptoms, and the goal of the treatment is to reduce the symptoms and their effect on the patients' quality of life (33).

The treatment of back pain is analogous. Image studies are commonly used to diagnose the cause of lower back pain. Among persons over the age of 40, the backs of as many as half appear on x-ray or other image studies to have serious problems, such as ruptured disks or stenosis, although the patients themselves experience no pain or disability (12,89). At the same time, image studies reveal no anomalies in other people who report experiencing pain in their lower backs and down their legs—pain that physicians are confident stem from stenosis or problem disks. Studies comparing symptomatic and asymptomatic patients consistently show they cannot be distinguished on the basis of images (10,61).

Patients' reports and the results of image studies are often complementary: the image study shows a ruptured disk or stenosis that corresponds well with the symptoms reported by a patient. When the two do not coincide, however, it is by no means clear that the clinical indicator should take precedence. To operate on a back when an image study indicated problems but the patient reported none would usually be inappropriate (40,68). The pain and dysfunction patients experience and report define whether the patients have back problems and are critical components of the indications for treatment; the relief of those symptoms and the restoration of functioning constitute the standard by which to evaluate whether medical care is effective or not. As is the case with BPH, the presence or severity of the condition is best defined by the patients' reports, not by clinical studies.

Patients' reports do not standalone in decisions about medical treatment. Although relieving symptoms is the focus of BPH treatment, the diagnosis of the reason for the symptoms and the likelihood that treatment will be effective depend on direct clinical evidence that the prostate is obstructing urination. If surgery is to be an effective treatment for back pain, a physiological problem that can be repaired by surgery must be identified. And some medical conditions are almost always defined by clinical examination and by test results. Patients' reports play little role in defining the presence of malignancies or hypertension, for example. Many common conditions, however, are best described by a combination of clinical observation, diagnostic tests, and patients' reports. Cataracts, arthritis, angina, and diseases of the uterus are particularly clear cases in which patients' reports play critical roles in defining the presence or severity of the conditions. Although the treatment for these conditions is physiological, the indications for treatment and the benefits of the treatment require assessing the status of the condition, in part, by asking the patients questions.

Measuring Treatment Complications

Comprehensive studies of treatments systematically estimate the frequency and severity of complications as well as their effects on the treated conditions. The presence or severity of many common complications cannot be characterized without patients' reports.

Accounting for the risks of complications is particularly important when the likelihood of a life-saving benefit of a treatment is relatively low. The treatment of mild hypertension, for example, is effective in preventing stroke; it reduces the probability that an otherwise healthy 50-year-old will have a stroke during the next five years from about 15 to about nine strokes per 1,000 men (19,56). The low overall probability of stroke means, however, that the great majority of men with mild hypertension would not have had strokes even without treatment. Because the medications used to lower blood pressure can reduce energy and sexual functioning and can produce depression, sleep disorders, anxiety, fainting, dizziness, and fatigue (21), a full evaluation of treat-

ment for mild hypertension must include both the likelihood of stroke reduction and the rates at which patients report the various side effects.

Surgery is widely performed in cases of prostate cancer, but such surgery produces high rates of sexual impotence and significant incontinence (29). Furthermore, no studies have shown surgery to be more likely than less invasive procedures to save lives (88). Thus, the net value of the surgery cannot be assessed without taking these complications into account.

Ascertaining the Effects of a Condition on Patients' Lives

A condition or symptom that constitutes a major problem for one patient may be only a small problem for another (31 ,33). Differences in patients' roles and responsibilities account for some of this variation. A person whose job entails heavy physical labor, for example, may be affected by lower back pain to a greater extent than an office worker is. Even if the pain is the same, the office worker may be better able to avoid putting stress on his or her back and may be better able to perform despite the pain. In contrast, a physical laborer maybe unable to work at all if the back problem is severe.

The significance of health conditions also depends on the individuals' feelings or response styles, which may have nothing to do with roles. Women's responses to options regarding surgery for breast cancer demonstrate this concept. For the majority of patients, the probabilities of survival are the same whether they choose to have lumpectomy with radiation or to undergo mastectomy (28,83). The perceived cosmetic advantages of lumpectomy make that a clear choice for some women, whereas others choose more radical surgery (90) because they feel more secure with a more aggressive—though equally effective—treatment.

Thus, assessing the significance of a condition and the benefit of any treatment requires information about how much the condition matters to the patient. Because the answer to this question generally varies from one person to the next, the patients' own reports are crucial. To address this need, researchers have developed methods of measuring patients' health-related quality of life.

MEASURING HEALTH-RELATED QUALITY OF LIFE

I Concepts and Components

Studies of the outcomes of medical care often use condition-specific measures for describing the patients' medical conditions, the complications of the treatments, and the patients' perceptions of how the conditions and treatments have affected their lives. It is increasingly recognized, however, that medical outcomes cannot be fully determined without ascertaining the treatments' effects on the patients' quality of life. Thus, many studies now also include general measures of health-related quality of life to provide comprehensive views of the patients' health at various times during the course of the treatment.

In this context, *quality of life* refers to the aspects of living that are affected by patients' medical conditions and to their functioning and perceived well-being. As defined by Patrick and Erickson, "Health-related quality of life is the value assigned to duration of life as modified by the impairments, functional states, perceptions and social opportunities that are influenced by disease, injury, treatment, or policy" (64).

Experts do not entirely agree on exactly what constitutes health-related quality of life, but most surveys designed to measure it include questions related to four basic aspects of functioning and well-being (58,64,7 1):

1. Functional ability. Questions aimed at discovering functional ability ask what people can do. The most common questions inquire about such physical activities as walking across a room, climbing a flight of stairs, or walking around a block. Other questions may cover such things as the patients' abilities to read a newspaper, to watch television, to hear well enough to talk on the telephone, or to hold a pen. All such questions are independent of patients' role expectations, resources, or responsibilities.

2. Perceived health. The simplest question about self-perceived health asks people to rate how healthy they think they are. Such a question has been a staple of the National Health Interview Survey for many years, and is perhaps the most widely used measure of health status (52). Other commonly measured aspects of self-perceived health are the degrees to which patients worry about their health and to which they are satisfied with their health.
3. Psychological well-being. Measures of psychological well-being usually focus on the extent to which patients see themselves as distressed-where they would place themselves on an emotional continuum with depression at one end and happiness at the other end, or with anxiety at one end and calmness at the other (9,85). Although they disagree about what specific questions should be asked, most researchers accept psychological well-being as fundamental to the issue of quality of life.
4. Role functioning. How health conditions affect people's lives depends on their roles—what is expected of them, what kind of work they perform, what resources they possess, and what they must do on a day-to-day basis (e.g., 82). Questions about role functioning are self-adjusting. A condition that would seriously limit a young professional athlete might not limit a retired person at all. A condition's effect on mobility might be severe for a person who must ride buses, moderate for a person with a car, and minimal for a person with a chauffeur. Common questions about role functioning in measures of quality of life address patients' abilities to work, to take care of themselves, to maintain their households, and to participate in society. Patients often are also asked about their abilities to take care of business, to get around, and to participate in the recreational activities of their choice.

Calculating Effects on Overall Quality of Life

There are two distinct approaches to calculating the effects of a particular treatment on a patient overall health-related quality of life. One way is to describe the results separately with respect to each component. Under this approach, the patient responses to questions in the instrument measuring quality of life might suggest, for example, that for a particular treatment the patient's physical functioning improved but that his or her perceived health did not change.

Another approach is to combine the ratings for all the various components into a single number that serves as a summary of the overall quality of life. Researchers following this approach must first determine how much weight to give to function, psychological distress, and measurements of other aspects of patients' lives, so that ratings for those different components can be combined quantitatively. The methods used to assign weights to different aspects of quality of life include statistical models, ratings by physicians, average ratings by patients, and ratings by samples of the general public. Perhaps the most obvious method is to ask people how they value their quality of life overall (4,35,36,37,38,59,64).

Measuring Condition-Specific vs. General Effects on Quality of Life

As described above, studies of medical outcomes usually require condition-specific measures aimed at describing the status of the patients' conditions, complications of common treatments, and perceptions of how the conditions and treatments have affected patients' lives. In addition, most studies now include general measures of health-related quality of life that provide comprehensive views of the patients' health at various points during the course of the treatment.[1]

[1]The many strategies for measuring health status and health-related quality of life have been extensively described and reviewed. McDowell and Newell (58) describe and review 50 measurement schemes based on subjective judgments and ratings. Patrick and Erickson (64) provide an excellent discussion of the conceptual underpinnings of the major efforts to measure perceptions of health. as well as a more detailed description of the development, uses, and limits of some of the most important approaches. Froberg and Kane (35,36,37,38) and Stewart and Ware (7 I) also provide excellent reviews of issues related to various aspects of the measurement of functioning, well-being, and health status.

The question of how much a patient is limited because of a particular condition, such as lower back pain, contains two components: to what degree is the patient limited, and to what extent is the limitation tied to the lower back pain? If the patient has only one condition that affects functioning, any limitation may be attributable to that condition. A person who has multiple conditions, however, may have difficulty attributing any particular effect to a specific condition or health problem. Indeed, as people age, many of their physical and intellectual capabilities decline, which may make it increasingly hard to report on the effects of each specific health condition. As a result, the effect of treatments may be ascertained more accurately by asking patients to assess their functioning and well-being over time, with and without treatments, than by asking them to attribute their deviations from perfect health to particular health problems.

General measures of health-related quality of life have another advantage as well. Fixing one condition, even a troublesome condition, may do only a little to benefit the overall quality of life of a patient with multiple conditions. Measuring overall quality of life in a way that reflects the effects of all the patient's health problems can demonstrate the true value of the treatment to the patient. Overall measures of quality of life also enable *re*searchers to take into account both the benefits and the downsides of treatments for a particular condition.

Patients' Satisfaction with Care

Patients' satisfaction with care is often mentioned as part of assessing medical outcomes (18,87). Satisfaction with the results of treatment reflects how patients rate their post-treatment states of health. Satisfaction with the process of care, however, depends on physicians' personal styles and how patients have been treated. A patient's assessment of the quality of care, therefore, doesn't necessarily indicate whether the treatment improved a medical condition (17).

Nonetheless, satisfaction with care is sometimes important for assessing medical services. Tests or examinations may be used, for example, simply to assuage patients' fears and worries. In such cases, the patients' satisfaction with the fact that procedures have been performed may be important. In assessing how a treatment has affected a medical condition or health status, however, patients' satisfaction with how the process itself was carried out is usually irrelevant.

I Instruments for Measuring Health-Related Quality of Life

Instruments to measure peoples' health status have been in use for decades (box 1-l). Attempts to measure health-related quality of life in a broader sense using survey instruments that ask detailed questions of the patients themselves, however, is a much newer development.

There are now numerous instruments used around the world to measure health-related quality of life, although not all of them rely on patients' reports. The Nottingham Health Profile is widely used in the United Kingdom (41), for example, and the EuroQol has been used in a 14-country study in Europe (25). The Arthritis Input Measurement Scale (57) and the OARS* Multi-dimensional Functional Assessment Questionnaire (27) are two of the more frequently cited instruments that rely on self-reporting.

In the United States, several programs for developing general measures of patients' well-being for use in clinical studies have been particularly important in influencing research on the outcomes of medical care. These programs include the Sickness Impact Profile research, the Medical Outcomes Study, and the Quality of Well-Being Scale.

[2] **OARS is the abbreviation for the Older** American's Resources and **Services Schedule.**

BOX 1-1: The Historical Development of Instruments To Measure Health Status

Early efforts to clinically measure patients' functioning and the severity of conditions include the development of the widely used Karnofsky Index for patients with cancer in the 1940s and the development of scales for the activities of daily living in the 1950s (22). Neither of these approaches based its ratings on the patients' own reports, however, and neither attempted to assess health status across wide ranges of patients or the general population.

Still, these early measures greatly influenced later survey instruments. The early rating schemes for how well people could take care of the basic activities of daily living (ADLs) (such as bathing, dressing, eating, and toileting) and the instrumental activities of daily living (IADLs) (such as housekeeping, getting around, and participating in social events) have been the basis for numerous scales using reports from patients and experts (e.g., 46,51 ,58,63). Moreover, ADLs and IADLs usually are part of more comprehensive strategies for assessing health status.

Another important influence on current strategies for measuring health has been the National Health Interview Survey (NHIS), which was established in the late 1950s to characterize and monitor the health of the nation (97). The NHIS pioneered the concept of asking people to rate their own health, using three main approaches. First, to detect the presence of health conditions, interviewers read lists of diagnoses to respondents and ask them whether they have or have had the conditions. Second, to ascertain the effects of illnesses, the NHIS asks respondents about the extent to which illness has caused any loss of work, absences from school, or days in which normal activities have been restricted. Third, since its inception the NHIS has asked respondents the following widely used health status question, which has proven valuable for many purposes: "Overall how *would you* rate your health--exce//ertt, very *good, good, fair, or poor?"*

Measuring the effects of illnesses by measuring the resulting disabilities or restrictions in activities has allowed researchers to evaluate the costs and other consequences of illness at a population level. The extent to which illness causes people to restrict their activities is also a functional measure that shows up in many studies of the outcomes of medical care. Although the NHIS was not designed for such studies, it is one of the most pervasive sources of questions used to assess health status and medical treatment.

SOURCE: **F.J. Fowler, 1995**

Sickness Impact Profile

In the 1970s, the National Center for Health Services Research (the predecessor of the Agency for Health Care Policy and Research) funded a program to develop a comprehensive instrument to measure the effect of sickness on people (8). This instrument, the Sickness Impact Profile (SIP), includes 136 statements about people's functioning and activities, such as:

= "I am not doing heavy work around the house."
= "I laugh or cry suddenly."
= "I walk shorter distances or stop to rest often."

These questions are grouped into three broad categories ("indices"), each of which has several subcategories. Peoples' responses to these statements thus produce measures of how illness affects 12 different aspects of patients' lives (box 1-2).

The entire SIP takes about 30 minutes to administer and is perhaps the most comprehensive and detailed inventory in common use. It has been subjected to extensive psychometric evaluation to assess its reliability, its stability over time, its ability to differentiate well people from sick people,

BOX 1-2: Aspects of Health-Related Quality of Life Measured by the SIP and the SF-36

The Sickness Impact Profile

The SIP measures 12 aspects of health-related quality of life, grouped in three categories, as follows:

/dependent categories	*Physical*	*Psychosocial/*
. sleep	• ambulation	• social interaction
■ eating	■ mobility	. alertness behavior
■ work	• body care and movement	• emotional behavior
. home management		• communication
• recreation and pastimes		

RAND **36-Item Health Survey**

The aspects of health-related quality of life covered in the SF-36 include:

- current perception of health
- psychological well-being
- role limitations due to physical health problems
- role limitations due to mental health problems
- physical function
- social relations
- pain
- fatigue

SOURCE M Bergner, R A Bobbttt, W B Carter, etal,, "The Sickness Impact Profile: Development and Final Revision of a Health Status Measure," Medical Cara 19(8) 787-805, 1981, and J,E. Ware and C.D. Sherbourne, "The MOS 36-item Short-Form Health Survey (SF-36)–1 Conceptual Framework and Item Selection," Medical Care 30(6):473-483, 1992

and its capacity to reflect positive effects of treatment, as well as to verify the internal consistency of its scales. The aspects of living reflected in the profile's 12 subindices tend to mirror those in current assessments of medical outcomes and patients' functioning. The basic approach developed in the SIP has had a major influence on subsequent efforts to develop better methods to evaluate medical outcomes. Moreover, all or part of the SIP is often used today in studies of medical outcomes.

The RAND 36-Item Health Survey

The SF-36 survey is probably the nation's most widely used generic instrument for measuring patients' assessments of health-related quality of life. The origins of the SF-36 lie in a health survey developed for the Health Insurance Experiment (HIE),[3] one of the major health research efforts of the 1970s (13,84). The 20-item questionnaire that emerged from the HIE later became a key instrument for collecting data in the Medical Outcomes Study, undertaken in the 1980s by some of the researchers who had worked on the HIE. The primary goal of the Medical Outcomes Study was to describe the health status of patients before and after medical treatment. The questionnaire later evolved into a 36-item, eight-index set of questions measuring various aspects of health, functioning, and quality of life (86) (box 1-2).

A primary goal in the development of the SF-36 was to identify a minimum set of health status dimensions that would cover most of the gen-

[3] The HIE focused chiefly on how various insurance packages with different deductibles affected utilization and cost. An important part of the study was an assessment of how people's health and well-being were affected by the different programs. Participants in the HIE filled out numerous questionnaires, which contained items from most of the major health survey measures available at the time, including the SIP. The results were analyzed to identify redundancy and independent dimensions of health status and functioning (13).

eral medical outcomes that researchers would want to measure, and to ask the minimum number of questions that would reliably and validly measure each dimension. The measure of psychological well-being, for example, consists of five items that have proved to be the measurement equal (for assessing aggregate outcomes) of as many as 30 of the items frequently used in other instruments to assess mental distress (9). Like the SIP, the SF-36 was envisioned as a generic instrument that would be appropriate for use in studying the treatment of virtually any health condition.

Although the developers of the SF-36 encourage researchers to use it as a complete package, the individual indices included in the SF-36 may be used by themselves, as can subsets of the SIP (71).

Quality of We//-Being Scale (QWB)

The Quality of Well-Being Scale (QWB), which emerged from the work of James Bush and his associates (44), uses a different approach. The SIP and the SF-36 rely on patients' reporting alone and were designed to be analyzed by looking at scores on individual subscales, which produce markedly different profiles depending on the type of illness. A summary SIP score can be calculated for overall functioning and for each of the three subdomains, and work is underway to derive a total score for the SF-36, but neither questionnaire was designed primarily to produce a single summary of well-being. In designing the QWB, however, Bush and his associates focused specifically on producing a quantitative measure of overall well-being.

To do so, these researchers created a list of deviations from perfect health. The list includes symptoms (such as headaches, sore throats, and trouble sleeping), conditions (such as hernias, overweight, and blindness), and activity or role limitations (such as missing work and being unable to drive a car). The respondent is asked whether any of these problems occurred during the preceding four days, and he or she rates each problem numerically according to the degree to which each of the problems reduced his or her well-being (64). These ratings are then combined by the researchers to produce a single number representing the overall well-being of the person.

In a variation on this approach, Torrance has developed the Health Utilities Index (73), which identifies nine health domains (vision, hearing, speech, the ability to get around, the use of hands and fingers, feelings, memory, thinking, and pain and discomfort). As with the QWB Scale, the Health Utilities Index entails calculating a weighted score that reflects the existence and seriousness of the problems reported in each domain.

RELIABILITY AND VALIDITY

For those schooled in the physical and biological sciences, the notion that good measurement can come from asking people questions seems somewhat implausible. Nonetheless, the criteria for evaluating questions as measures of health status are the same as those for evaluating measures used in laboratories, physicians' offices, or anywhere else. When consistent standards are applied, measurements based on asking people questions stand up very well.

I Reliability

Reliability means that measurement is consistent: when two patients are in the same situation, their answers to the questions should be the same. To the extent that there is inconsistency among patients, or at different times with respect to the same patient (when the patient's circumstances have not changed), the measurement is unreliable and imprecise.

The most commonly used measure of reliability in medical science is test-retest reliability, in which researchers compare two readings from the same person at different points in time. When no change in the patient's condition is thought to have occurred, the readings should be consistent. Researchers assessing the subscales used in the SIP, the SF-36, and other similar questionnaires

routinely reported test-retest reliabilities of 0.85 and above.[4]

Validity

Assessing the validity of patients' reports for the purpose of evaluating medical outcomes is often difficult. Where there is a standard, a measure that everyone agrees is an accurate measure, validity can be assessed simply by comparing the results to the standard. Because no generally accepted standard for measuring patients' functioning or well-being exists, however, the evidence for the validity of patients' reports must come from the predictability of relationships.

Clinical measures are evaluated by examining the extent to which they discriminate between known groups and the extent to which they are responsive to treatments thought to be effective (47). Thus, a valid measure of symptoms of prostate disease, for example, should show higher levels among patients diagnosed with BPH than among the general population and higher levels in patients before they are surgically treated than after they are treated. The general approach of looking at patterns of association, how well the measures correlate with things with which they ought to be correlated, is the primary basis on which validity is assessed.

Many survey instruments used in clinical work ask patients multiple questions that cover the same general area. The study of associations between the answers to similar questions constitutes an important strategy for validating questions as measures. Questions about pain should correlate positively with other measures of discomfort, and they should be less correlated with measures of fatigue. The measurement can be strengthened by combining the answers to several questions to form an index. The reasons for using multi-item scales is that, all things being equal, multi-question scales are better than a single question at measuring what those questions have in common. (The extent to which multi-item scales provide a consistent and reliable measure of what they have in common is calculated by a statistic called Cronbach's alpha [20].)

Another issue *is face validity,* which means that the answers to questions mean what a reader of the wording of the question would most likely think they would mean. On the one hand, having questions that clinicians agree adequately cover what needs to be covered is critical to the acceptance of the results. On the other hand, questions cannot be presumed to be good measures just because they sound like the right questions. A requirement for any scientific enterprise is that the quality of measurement be documented through experiment and observation.

A good example was set by the researchers responsible for developing the SIP and those developing the SF-36 and related measures (8, 13,71). In the course of these programs of research, the investigators uniformly reported the ability of the measures to discriminate among clinical groups, the internal consistency of multi-index measures, the responsiveness to treatment, and the patterns of association with other measures with which they should be correlated. In all these respects, measures of the subscales in the SIP and of the various scales used in the Medical Outcomes Study meet high standards. The Cronbach's alpha rates routinely exceed 0.80, and correlations among related concepts are also very high.

The same kind of standards can be applied to more specific measures aimed at particular conditions and symptoms. A recent effort by the Measurement Committee of the American Urological Association to measure symptoms of BPH, comparing alternative measures of symptoms, demonstrates the high quality of measurement based on patients' reports (6). The committee compared and contrasted four different sets of questions about symptoms of BPH (7). Samples of patients and nonpatients answered questions

4 A reliability of 1.0 would mean that the instrument yielded identical answers every time. A score of 0.85 is generally considered acceptably high (58).

twice, one week apart. The test-retest reliabilities of all four scores exceeded 0.75. The internal-consistency measure, Cronbach's alpha, for all four indices exceeded 0.80. The intercorrelations among the four indices, which partly reflected the overlap of items, were all above 0.75. These statistics show that BPH has meaningful symptoms that sets of questions can measure in a consistent and apparently valid way. Furthermore, when the answers of patients diagnosed as having BPH were compared with the sample of healthy individuals, 85 percent of the people would have been correctly classified as BPH patients or nonpatients based on their answers to those questions.

Comparison with Other Tools

Although the results of diagnostic tests and results based on patients' reports are difficult to compare directly, studies of diagnostic tests and of measurements taken in clinical settings almost always reveal considerable error. Blood pressure readings, for example, are often inaccurate: using the wrong cuff size is common and produces serious overestimates of blood pressure (30,53); and in a phenomenon known as "white coat" response, 20 to 40 percent of people who have elevated blood pressure readings in doctors' offices have normal blood-pressure readings in other settings (49,65). Thus, even though the measurement of blood pressure is considered an important procedure upon which important diagnoses and treatment decisions are based, the measurement process is fraught with potential error.

The lack of correspondence between the results of image studies and the symptoms of people with lower back pain provides another example of a traditional medical test that is not a consistently reliable or valid indicator of a health condition (12,67,89). Similar problems have arisen with the use of image studies to diagnose arthritis (23,81).

To help evaluate BPH, urologists have traditionally used a measure of the residual urine left in the bladder after voiding and have also begun using a measure of the rate of urine flow to assess obstruction. These measures correlate poorly with patients' symptoms, however (5,14). Although factors that have nothing to do with BPH status (such as recent fluid intake and patients' anxiety) apparently affect the measures, they continue to be a common part of the urologic diagnostic process.

There are at least three reasons why clinical tests may not be valid measures.

- First, the variable state being measured may not be a reliable indicator of the condition of a patient. (Because blood pressures go up and down in response to circumstances, the reading at any point in time may not be a good indicator of the usual state of a person's blood pressure.)
- Second, the measurement may be performed inconsistently or incorrectly, affecting the results. (Using the wrong cuff to measure blood pressure yields an erroneous reading.)
- Third, what can be measured may not be informative about the condition of interest. (In the case of back pain, some of the things that affect the nerves coming out of the spine are apparently not visible in image studies.)

Measuring clinical or medical states by asking people questions is subject to the same kinds of problems. Whether people can answer questions that provide valid measurements of a clinical state is an empirical question to be tested, and the validity of patients' reports can vary from condition to condition. (No matter how well they can describe their pain or functioning, for instance, patients cannot say what their blood pressures are based on feelings alone.) Aspects of the data collection procedures, such as the quality of interviewing in those cases where interviewers are used, also can affect the results (33).

Patients' reports cannot substitute for other strategies of clinical observation and diagnosis, nor are medical tests inherently unreliable. The reliability and validity of clinical and laboratory tests vary, as do those of measures based on patients' reports. For measuring what patients observe and experience, however, properly designed questions can produce measures that compare favorably in reliability and validity with traditional clinical measures of health.

ISSUES

Broad-based agreement on how to conduct good studies of the outcomes of medical care is emerging, but consensus on the details is lacking. Many studies are now designed to collect data about the treated condition and complications of treatment using patients' questionnaires or interviews. Ideally, for the sake of simplicity and comparability, all studies of a particular condition would use the same measures. There are very few conditions, however, for which a specific set of questions is widely accepted. In general, researchers are still developing and revising questions to meet their perceptions of what is best.

The lack of consensus on specific measures of health states does not mean that studies cannot be compared. Questions that validly measure the same underlying conditions will produce similar results, even if the wording is different. The scores of the resulting indices of four recently evaluated series of questions that have been used in published studies to measure the severity of BPH, for example, intercorrelated very highly (6). Consequently, studies using any of the four measures are likely to produce similar conclusions about the effect of treatment.

Medical outcomes studies now routinely include general, as well as condition-specific, measures of functioning and perceived well-being. The domains (i.e., the aspects of health and well-being) covered in the general measures are similar, drawing on those covered in the SIP and SF-36, but the particular indices and questions vary.

Some of the diversity in the choice of measures reflects the characteristics of the condition or the populations being studied. The range of functioning to be measured in studies of stroke victims, for example, is very different from that in studies of women who have had Cesarean sections, both because of the patients' ages and because of the way the conditions affect people.

Most of the outcomes studies being done by the PORTS, funded by the Agency for Health Care Policy and Research, are using some of the indices from the SF-36. In addition, various PORTS are using all or part of the SIP, asking specific questions about activities of daily living and instrumental activities of daily living,[5] and inquiring about disability days or restricted-activity days, as part of their protocols to assess the effects of treatment comprehensively. One argument for using the entire SIP or SF-36 is that these measures produce comprehensive profiles of the patients. Researchers differ, however, in how much they value measures of domains that are not likely to be affected by a particular treatment.

Thus, although there is virtually complete agreement on the need for measures of patients' self-reported health, there is diversity in the questions chosen by different researchers. The differences reflect the conditions being studied, the populations of patients, the burdens deemed appropriate for respondents in particular projects, and the personal convictions of the researchers about which specific measures are best suited for studying particular treatments.

Some convergence will probably occur as researchers gain experience. More systematic evaluation of questions is needed, however. Questions need to be tested with cultural minorities, for example, to ensure the questions really mean the same thing to everyone. Optimal questions about role limitations-questions that apply equally well to all age groups, including children and retired persons—have yet to be found.

Although these issues are important and need to be addressed, they are relatively minor problems that should not detract from the agreement about the need for general measures of health-related quality of life and about the advances in developing good measures of the major aspects of quality of life. Nonetheless, at least three major methodological challenges remain. If medical

[5] See box **1-1**.

outcomes studies are to live up to their promise, each of the following issues must be resolved:

- How should prospective and retrospective designs be modified to ensure accurate measurements of the effects of treatment?
- How should researchers collect information about the results that would have been expected had a particular treatment not been given?
- How should the effects of treatment be calibrated to facilitate comparisons across conditions?

Prospective vs. Retrospective Designs

Both prospective and retrospective designs are used to assess treatment effects with patient surveys. In a prospective study, patients are asked the question *"How are you doing?"* before treatment and again at a later point in time. The effect of the treatment is assessed by comparing the two answers. In a retrospective approach, people who have already been treated are asked to compare their present state with how they felt prior to the treatment: *"DO you think you are doing better now, worse now, or about the same?"*

The two methods do not always yield the same results. Some people report that they feel better after treatment even though comparisons of their reported symptoms before and after treatment indicate no changes in their conditions (39,54). Some studies of medical outcomes are most easily done retrospectively: one cannot easily identify (or collect data from) the individuals who will later have heart attacks or suffer accidental injuries, whereas surgical patients are relatively easy to identify after they have had surgery. Therefore, it is important to develop an understanding of how best to conduct both prospective and retrospective studies. The measurement implications of the two kinds of designs should also be considered.

1 Better-Than-Expected Results

Assessing whether results are better than expected is another methodological problem that needs work. Although ascertaining whether patients change for the better by virtue of being treated may seem a good way to assess treatment results, considerable medical care is intended merely to keep patients from getting worse. The management of a patient who has had an acute myocardial infarction (AMI),[6] for example, is designed to make the recovery process as good as possible. Because people who have suffered AMIs cannot reasonably be expected to be better off than they were before the AMIs, their health status must be compared with what it would have been had they been treated differently. By the same token, although measuring the reduction in symptoms may be a good way to assess the value of the treatment (where symptom relief was the primary goal of the treatment), some people improve without treatment.

These examples underscore the fact that all treatment or outcomes studies require controls or comparisons and for those, studies of untreated people (or some other "control" group) are necessary. The traditional standard in clinical research is a randomized controlled trial (RCT). The design is good when it is feasible, but such studies often have not been done and sometimes cannot be done.

In the absence of good RCTS, researchers have been trying to do better descriptive studies of patients who undergo particular treatments, but valid conclusions about the treatments' effects are difficult to reach without good data about what would have happened to the patients had they received no treatment or alternative treatments. Such data are scarce. One critical gap is the relative lack of natural history studies. Patients who present themselves to physicians and meet criteria for surgical treatment are 1ikel y to get the surgical treatment, particularly in the United States. There is a dearth of studies that systematically follow candidates for surgery or hospitalization who do not actually receive the surgery or hospitalization.

6 An acute myocardial infarction is a type of heart attack.

On a related issue, cohorts who are not given an extreme treatment, such as surgery, tend to be different from the aggressively treated group. As a result, appropriate data about the symptom status, comorbidities, and general health of both cohorts must be collected so that appropriate controls can be used in analytic comparisons of the outcomes. The Medical Outcomes Study used this kind of design. Its approach was a major advance over having no comparison group at ali, but researchers often have difficulty making adjustments to ensure that the comparisons are appropriate (43). Agreed-upon methods for cohort studies of people who receive different treatments or no treatments must be developed to provide data that will enable researchers to reach valid conclusions about treatment effects.

Calibrating Measurements of Treatment Effect

Measuring patients' views of the significance of particular clinical states, including complications of treatment, is central to the problem of how to measure the value of a treatment. In the past, there were few good studies of the overall health status of patients before and after treatments. As more such studies are conducted, however, the question of how to calibrate benefits will become much more salient.

A single summary measure of the net significance or value of medical treatment would be useful for decision analysis,[7] for ranking the value of performing various medical procedures, or for deciding whether a particular treatment is one for which we are willing to pay.

In clinical practice at the individual patient level, a single summary measure can be obtained simply by describing the various possible results to the patient, who then makes his or her own choice based on personal preferences and values. Some problems, however, raise social questions of cost, ethics, or best medical practice (box 1-3). Producing good statistical descriptions of the results of treatments might improve judgments and social choices in these cases.

The QWB Scale and the Health Utilities Index seek to address this need by asking groups of people to rate quantitatively how they value vari-

BOX 1-3: Questions That Might Benefit from a Summary Measure of Treatment Effect

- Suppose two treatments are available: one has an 80-percent chance of relieving the symptoms and a 20-percent chance of producing certain side effects; the alternative is less effective but has fewer side effects. Which is the best treatment?
- Suppose the costs of two treatments are significantly different. Is the more expensive treatment justified?
- Suppose a treatment is found that can make a measurable improvement in the cognitive functioning of mentally impaired elderly patients, but the treated patients remain substantially impaired after treatment How should the value, if any, of such a treatment be calculated?
- Suppose a treatment will prevent 30 premature deaths for every 1,000 people treated, but most of the treated people will have significant short-term side effects, a few will have long-term quality-of-life loss, and the treatment is expensive. Should the treatment be used?

SOURCE F.J. Fowler, 1995

[7] In a decision analysis, the analyst considers the variety of possible treatment options, associates each treatment option with a set of probabilities for good and bad outcomes, and tries to put them together to illuminate the implications of each treatment (45). Critical components of any decision analysis are the values assigned to the various health states in which patients may find themselves, with or without treatment. These measures of significance-numbers assigned to describe how good or bad patients' states **are--can be derived from each individual patient, from the average ratings of a group of patients, or from independent ratings (59).**

ous states of health. In doing so, instruments like these raise issues about the method by which the ratings of health status are derived—issues that do not arise with instruments such as the SIP and SF-36, which do not attempt to come up with a summary measure of health-related quality of life. Two issues are especially central:

1. What questions should be asked to rate health status?
2. Who should be the raters?

One way to measure what significance a condition holds for people is to ask them to rate it numerically: on a scale from O to 100, where O is death and 100 is perfect health, what number would you give to, for example, lower back pain? This is the approach used by Kaplan and his associates (42). Other researchers, however, believe that the valid measurement of the significance or importance of a condition requires asking people how much they are willing to pay, to risk, or to lose in order to get rid of a condition—an approach that leads to an entirely different set of questions (26,35,36,37,38,73). These researchers prefer the standard reference gamble, which goes something like this:

> Option A is to have no treatment at all and stay in your current state of health. Option B involves accepting a treatment. If the treatment is successful, it will cure your condition and return you to perfect health. If it is unsuccessful, you will die. With what chance of success would you choose Option B over Option A?

A variation is called the time tradeoff. It also trades off life against health, quantity versus quality of life, but in a different way:

> Consider the possibility that you will live 10 years with your health just the way it is now. Suppose I could offer you a treatment that would return you to perfect health, without the condition, but you would live fewer years. How many years of perfect health would you consider to be the same as 10 years in your current health state?

These approaches presume that the greater the risks people are willing to take or the more of their lives they are willing to give up to improve their current health, the worse the states of health in which they find themselves.

Studies assessing the significance of health conditions have used all of these approaches: asking patients to rate how they think they are affected by various health conditions, asking people to rate how they think they would feel if they were in various health states, and asking expert raters (such as physicians) to say how they think patients would feel if they got into various states. The QWB and Health Utilities Index use ratings by samples of people to assign weights and produce a summaries of well-being. Both have been used in clinical studies of medical outcomes. In addition, a variation of QWB was used in Oregon to set priorities for proposed revisions in the Medicaid payments system (34,77), and the Health Utilities Index was used by Statistics Canada to assess well-being in a general population survey in Ontario.

Researchers disagree about whether scale- or risk-based approaches are better. Many researchers believe that questions based on the standard gamble or time tradeoff approach are by far the best way to measure how significant particular health states are to people (72). Others point out that these are very hard questions to answer, and that the answers may not have the meaning the researchers hoped for. Moreover, questions based on gambles and tradeoffs reflect not only the value of health states but also the individuals' attitudes about trading quality and quantity of life and toward taking risks, and thus they have been criticized as producing confounded—rather than better—measures of the value of health states. Research using both approaches continues.

As for the issue of whose values should be reflected in the ratings, the answer depends in part on the purposes for which data are being collected. If a physician is treating an individual patient, the patient's preferences should have priority. For managed health care, however, the values of the average patient might be the most relevant **(59). A** different set of priorities might be appropriate for an insurance company. In that context, the perspectives of the people who are paying the pre-

miums might be most appropriate. Applying that logic to government-funded health care might entail using the values of a cross-section of the general public to determine the ratings (64). But in the context of government, where the question of whose values matter is a political as well as an academic one, there is no unambiguous answer.

Thus, although the SIP, the SF-36, and similar survey instruments can yield summary measures, their strength lies in producing profiles of the various ways a health condition affects people's lives. The QWB and Health Utilities Index researchers address the problem more directly, but they have not resolved the perplexing issues of which questions to ask, whose values to measure, and how to create an overall summary of the quality of life. Describing patients' post-treatment status on various indices may actually be the best form in which to convey information to patients and physicians, but those who want a simple summary number—for decision analysis, for ranking the value of hysterectomies and fixing broken legs, or for deciding whether to pay for a particular treatment-do not yet agree about how to proceed.

CONCLUSION

It is not accidental that researchers' recent interest in developing measures of health-related quality of life has coincided with widespread interest in better assessing the value of current medical treatments. Patients' reports about their perceptions of their symptoms, about the significance of their conditions, and about their general functioning and quality of life are essential to documenting what benefits, if any, patients derive from treatments.

One of the contributions of the PORT concept has been to emphasize the patients' perspective in the evaluation of medical treatments. Although some very good work was done in the 1970s and became the foundation of current work, the focus on how patients fare after treatment is mainly a recent phenomenon. That patients' reports can provide valid and reliable measures of their health status has been clearly demonstrated. Indeed, measures from patients' reports often prove better than those from commonly used clinical and laboratory tests, and studies that include patients' perspectives have produced sound results that sometimes raise questions about standard medical practice.

Work remains to be done in developing and improving measures. Researchers need to increase their understanding of how best to conduct these studies to reach valid conclusions and how best to assess the significance of the results. Nonetheless, in a comparatively short time, an appreciation for patient-oriented outcomes studies and how to do them has developed a great deal. They can be done, and they produce considerable knowledge that neither patients, clinicians, nor researchers have had before.

REFERENCES

1. Abrams, P. H., and Greffeths, D.J., "The Assessment of Prostatic Obstruction from Urodynamic Measurements and from Residue Urine," *British Journal of Urology 51:129, 1979.*
2. American College of Cardiology/American Heart Association, "Special Report: A Report of the American College of Cardiology/ American Heart Association Task Force on Assessment of Diagnostic and Therapeutic Cardiovascular Procedures (Subcommittee on Coronary Artery Bypass Graft Surgery)," *Circulation 83(3) :1125-1172, 1991.*
3. Andersen, J.T., Nordling, J., and Walter, S., "Prostatism-I: The Correlation Between Symptoms, Cytometric, and Urodynamic Findings," *Scandinavian Journal of Urolology and Nephrology 13;229, 1979.*
4. Andrews, F. M., and Withey, S. B., *Social Indicators of Well-Being: Americans' Perceptions of Life Quality (New* York, NY: Plenum, 1976).
5. Barry, M.J., Cockett, A. T. K., Holtgrewe, H.L., et al., "Relationship of Symptoms of Prostatism to Commonly-Used Physiological and Anatomical Measures of the Severity of Benign Prostatic Hyperplasia," *Journal of Urology,* in press.

6. Barry, M.J., Fowler, F.J., O'Leary, M. P., et al., "The American Urological Association Symptom Index for Benign Prostatic Hyperplasia," *Journal of Urology 148: 1549-1557, 1992.*
7. Barry, M.J., Fowler, F.J., O'Leary, M. P., et al., "Correlation of the American Urological Association Symptom Index with Self-Administered Versions of the Madsen-Iversen, Boyarsky, and Maine Medical Assessment Program Symptom Indexes," *Journal of Urology 148:1558-1563, 1992.*
8. Bergner, M., Bobbitt, R. A., Carter, W. B., et al., "The Sickness Impact Profile: Development and Final Revision of a Health Status Measure," *Medical Care 19(8):787-805, 1981.*
9. Berwick, D. M., Murphy, J. M., Goldman, P. A., et al., "Performance of a Five-Item Mental Health Screening Test," *Medical Care 29(2): 169-176, 1991.*
10. Bigos, S.J., Battie, M. C., Fisher, L. D., et al., "A Prospective Evaluation of Preemploy - ment Screening Methods for Acute Industrial Back Pain," *Spine 17(8):922-926, 1992.*
11. Black, N., Petticrew, M., Ginzler, M., et al., "Do Doctors and Patients Disagree? Views of the Outcome of Transurethral Resection of the Prostate," *International Journal of Technology Assessment in Health Care 7(4):533-54-4,*
12. Boden, S. D., Davis, D. O., Dina, T. S., et al., "Abnormal Magnetic-Resonance Scans of the Lumbar Spine in Asymptomatic Subjects," *Journal of Bone and Joint Surgery* 72(3): 403-408, 1990.
13. Brook, R. H., Ware, Jr., J. E., Davies-Avery, A., et al., *Conceptualization and Measurement of Health for Adults in the Health Insurance Study: Vol. VIII—Overview,* Publication No. R-1987/8-HEW (Santa Monica, CA: RAND Corp., 1987).
14. Bruskewitz, R. C., Iversen, P., and Madsen, P. O., "Value of Postvoid Residual Urine Determination in Evaluation of Prostatism," *Urology* 20:260, 1982.
15. Carlson, K.J., Miller, B.A., and Fowler, Jr., F.J., "The Maine Women's Health Study: I-Outcomes of Hysterectomy," *American Journal of Obstetrics and Gynecology 83:556-565,* 1994a.
16. Carlson, K. J., Miller, B. A., and Fowler, Jr., F.J., "The Maine Women's Health Study: II—Outcomes of Nonsurgical Treatment for Fibroids, Abnormal Bleeding, and Chronic Pelvic Pain," *American Journal of Obstetrics and Gynecology 83:566-572,* 1994b.
17. Cleary, P. D., Edgman-Levitan, S., Roberts, M., et al., "Patients Evaluate Their Hospital Care: A National Study," *Health Affairs 10(4):254-267, 1991.*
18. Cleary, P. D., and McNeil, B.J., "Patient Satisfaction as an Indicator of Quality of Care," *Inquiry 25:25-36, 1988.*
19. Collins, R., Pete, R., MacMahon, S., et al., "Blood Pressure, Stroke, and Coronary Heart Disease," *Lancet 335:827-838, 1990.*
20. Cronbach, L.J., "Coefficient Alpha and the Internal Structure of Tests," *Psychometrika 16:297, 1951.*
21. Croog, S. H., Levine, S., Testa, M. A., et al., "The Effects of Antihypertensive Therapy on the Quality of Life," New *England Journal of Medicine 314:1657-1664, 1986.*
22. Deyo, R. A., "The Quality of Life, Research and Care" (editorial), *Annals of Internal Medicine* 114:695-697, 1991.
23. Deyo, R. A., "Comparative Validity of the Sickness Impact Profile and Shorter Scales for Functional Assessment in Low Back Pain," *Spine* 11:951-54, 1986.
24. Epstein, A. M., Hall, J. A., Tognetti, J., et al., "Using Proxies To Evaluate Quality of Life: Can They Provide Information About Patients' Health Status and Satisfaction with Medical Care?" *Medical Care 27(3):S91-98, 1989.*
25. EuroQol Group, "EuroQol-A New Facility for the Measurement of Health-Related Quality of Life," *Health Policy 16: 199-208, 1990.*
26. Feeny, D. H., and Torrance, G. W., "Incorporating Utility-Based Quality-of-Life Assess-

ment Measures in Clinical Trials: Two Examples," *Medical Care* 27(3)(suppl.):S *190-S204, 1989.*

27. Fillenbaum, G.G., and Smyer, M. A., "The Development, Validity, and Reliability of the OARS Multidimensional Functional Assessment Questionnaire," *Journal of Gerontology* 36:428-434, *1981.*

28. Fisher, B., Redmond, C., Poisson, R., et al., "Eight-Year Results of a Randomized Clinical Trial Comparing Total Mastectomy and Lumpectomy With or Without Irradiation in the Treatment of Breast Cancer," New *England Journal of Medicine* 320:822-828, *1989.*

29. Fleming, C., Wasson, J. H., Albertsen, P. C., et al., "A Decision Analysis of Alternative Treatment Strategies for Clinically Localized Prostate Cancer," *Journal of the American Medical Association* 269:2650-2658, *1993.*

30. Floras, J. S., Hassan, M. O., Osikowska, B., et al., "Cuff and Ambulatory Blood Pressure in Subjects with Essential Hypertension," *Lancet* 2(8238): 107-109, 1981.

31. Fowler, F.J., "Patient Reports of Symptoms and Quality of Life Following Prostate Surgery," *European Journal of Urology 20* (suppl. 2):44-49, 1991.

32. Fowler, F. J., and Mangione, T. W., *Standardized Survey Interviewing, Minimizing Interviewer-Related Error* (Newbury Park, CA: Sage Publications, 1990).

33. Fowler, F. J., Wennberg, J. E., Timothy, R. P., et al., "Symptom Status and Quality of Life Following Prostatectomy," *Journal of the American Medical Association 259(20): 3018-3022, 1988.*

34. Fox, D. M., and Leichter, H. M., "Rationing Care in Oregon: The New Accountability," *Health Affairs* 10(2):728, *1991.*

35. Froberg, D.G., and Kane, R. L., "Methodology for Measuring Health-State Preferences—1: Measurement *Strategies, ''Journal of Clinical Epidemiology* 42(4):345-354,1989.

36. Froberg, D. G., and Kane, R. L., "Methodology for Measuring Health-State Preferences—II: Scaling *Methods, ''Journal of Clinical Epidemiology 42(5):459-471, 1989.*

37. Froberg, D. G., and Kane, R. L., "Methodology for Measuring Health-State Preferences—III: Population and Context Effects," *Journal of Clinical Epidemiology* 42(6):585-592, *1989.*

38. Froberg, D. G., and Kane, R. L., "Methodology for Measuring Health-State Preferences—IV: Progress and a Research Agenda," *Journal of Clinical Epidemiology* 42(7):675-685, *1989.*

39. Guyatt, G. H., Townsend, M., Keller, J. L., et al., "Should Study Subjects See Their Previous Responses? Data from a Randomized Control Trial," *Journal of Clinical Epidemiology* 42(9):913-920, *1989.*

40. Herron, L. D., and Turner, J., "Patient Selection for Lumbar Laminectomy and Discectomy with a Revised Objective Rating S ystem," *Clinical Orthopedics and Related Research* 199:145-152, *1985.*

41. Hunt, S. M., McEwen, J., and McKenna, S. P., *Measuring Health Status* (Dover, NH: Croom Helm, 1986).

42. Kaplan, R. M., Anderson, J. P., Wu, A. W., et al., "The Quality of Well-Being Scale: Applications in AIDS, Cystic Fibrosis, and Arthritis," *Medical Care* 27(3)(suppl.):S27-S43, *1989.*

43. Kaplan, R. M., and Berry, C. C., "Adjusting for Confounding Variables," *Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data--Conference Proceedings,* (Rockville, MD: Agency for Health Care Policy and Research, May 1990)

44. Kaplan, R. M., and Bush, J. W., "Health-Related Quality of Life Measurement for Evaluation and Research and Policy Analysis," *Health Psychology* 1:61, *1982.*

45. Kassirer, J.P., Moscowitz, A. J., Law, J., et al., "Decision Analysis: A Progress Report," *Annals of Internal Medicine 106:275-291,* 1987.

46. Katz, S., Ford A. B., Moskowitz, R. W., et al., "Studies of Illness in the Aged. The Index of ADL: A Standardized Measure of Biological

and Psychosocial Function," *Journal of the American Medical Association* 185:914-919, *1963.*

47. Kirshner, B., and Guyatt, G., "A Methodologic Framework for Assessing Health Indices," *Journal of Chronic Disease* 38:27, *1985.*

48. Lee, N. C., Dicker, R.C., Rubin, G. L., et al., "Confirmation of the Preoperative Diagnoses for Hysterectomy," *American Journal of Obstetrics and Gynecology,* 150:283-287, *1984.*

49. Lerman, C. E., Brody, D. S., Hui, T., et al., "The White-Coat Hypertension Response: Prevalence and Predictors," *Journal of General Internal Medicine* 4:226-231, 1989.

50. Magaziner, J., Simonsick, E. M., Kashner, T. M., et al., "Patient-Proxy Response Comparability on Measures of Patient Health and Functional Status," *Journal of Clinical Epidemiology 41(1* 1):1065-1074, *1988.*

51. Mahoney, F. I., and Barthel, D. W., "Functional Evaluation: The Barthel Index," *Maryland State Medical Journal* 14:61-65, 1965.

52. Manning, W.G., Newhouse, J. P., and Ware, Jr., J. E., "The Status of Health in Demand Estimation, or Beyond Excellent, Good, Fair, and Poor," *Economic Aspects of Health,* V.R. Fuchs (cd.) (Chicago, IL: University of Chicago Press, 1982).

53. Manning, D. M., Kuchirka, C., and Kaminski, J., "Miscuffing: Inappropriate Blood Pressure Cuff Application," *Circulation* 68(4):763-766, *1983.*

54. MacKenzie, R. C., Charlson, M. E., DiGioia, D., et al., "Can the Sickness Impact Profile Measure Change? An Example of Scale Assessment," *Journal of Chronic Disease* 39(6):429-438, *1986.*

55. Mebust, W. K., Holtgrewe, H. L., Cockett, A.T., et al., "Transurethral Prostatectomy: Immediate and Postoperative Complications. A Cooperative Study of 13 Participating Institutions Evaluating 3,885 Patients, ''Journal *of Urology* 141:243-247, 1989.

56. Medical Research Council Working Party, "Stroke and Coronary Heart Disease in Mild Hypertension: Risk Factors and the Value of Treatment," *British Medical Journal 296: 1565-1570, 1988.*

57. Meenan, R.F., "The AIMS Approach to Health Status Measurement: Conceptual Background and Measurement Properties," *Journal of Rheumatology* 9:785-788, 1982.

58. McDowell, I., and Newell, C., *Measuring Health, A Guide to Rating Scales and Questionnaires (New* York, NY: Oxford University Press, 1987).

59. Mulley, A.G., "Assessing Patients' Utilities: Can the Ends Justify the Means?" *Medical Care 27(suppl.):S269-S281, 1989.*

60. Neal, D.E., Styles, R. A., Ng, T., et al., "Relationship Between Voiding Pressures, Symptoms, and Urodynamic Findings in 253 Men Undergoing Prostatectomy," *British Journal of Urology* 60:554, 1987.

61. Paajamen, H., Erkintalo, M., Dahlstrom, S., et al., "Disc Degeneration and Lumbar Instability. Magnetic-Resonance Examination of 16 Patients," *Acta Orthopaedica Scandinavia* 60(4):375-378, *1989.*

62. Parliament, M. B., Danjoux, C. E., and Clayton, T., "Is Cancer Treatment Toxicity Accurately Reported?" *International Journal of Radiation Oncology, Biology, Physics* 11:603-608, 1985.

63. Patrick, D. L., Darby, S.C., Green, S., et al., "Screening for Disability in the Inner City," *Journal of Epidemiology and Community Health* 35:65-70, *1981.*

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

65. Pickering, T. G., James, G. D., Boddie, C., et al, "How Common Is White Coat Hypertension?" *Journal of the American Medical Association* 259:225-228, *1988.*

66. Rothman, M. L., Hedrich, S. C., Bulcrot, K. A., et al., "The Validity of Proxy-Generated Scores as Measures of Patient Health Status," *Medical Care 29(2):* 115-124, 1991.

67. Spangfort, E. V., "Lumbar Disc Herniation: A Computer-Aided Analysis of 2,504 Opera-

tions," *Acta Orthopaedica Scandinavia, 142(suppl.):l-95, 1972.*

68. Spengler, D. M., Ouellette, E. A., Battie, M., et al., "Elective Discectomy for Herniation of Lumbar Disc. Additional Experience with an Objective Method," *Journal of Bone and Joint Surgery* 72(2):230-237, *1990.*

69. Stewart, A. L., Greenfield, S., Hays, R. D., et al., "Functional Status and Well-Being of Patients with Chronic Conditions: Results from the Medical Outcomes Study," *Journal of the American Medical Association 262(7):907-913, 1989.*

70. Stewart, A. L., Hays, R. D., and Ware, J. E., "The MOS Short-Form General Health Survey, Reliability and Validity in a Patient Population," *Medical Care* 26(7):724-735, *1988.*

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

72. Torrance, G.W., "Measurement of Health State Utilities for Economic Appraisal," *Journal of Health Economics* 5:1-30, *1986.*

73. Torrance, G. W., "Utility Approach to Measuring Health-Related Quality of Life," *Journal of Chronic Disease* 40(6):593-600, *1987.*

74. Turner J. A., Ersek, M., Herron, L., et al., "Surgery for Lumbar Spinal Stenosis: Attempted Meta-Analysis of the Literature," *Spine* 17(1):1-8, *1992.*

75. Turner, C., and Martin, E., *Surveying Subjective Phenomena (New* York, NY: Russell Sage, 1984).

76. Uhlmann, R. F., Pearlman, R. A., and Cain, K. C., "Physicians' and Spouses' Predictions of Elderly Patients' Resuscitation Preferences," *Journal of Gerontology* 43(5):M115-M121, 1988.

77. U.S. Congress, Office of Technology Assessment, *Evaluation of the Oregon Medicaid Proposal,* OTA-H-531 (Washington, DC: U.S. Government Printing Office, May 1992),

78. U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control and Prevention, National Center for Health Statistics, "Hysterectomies in the United States, 1965 -1984," *Vital and Health Statistics,* Series 13, No. 155, DHHS Pub. No. (PHS)87-1753 (Hyattsville, MD: 1987).

79. U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control, National Center for Health Statistics, "The National Ambulatory Medical Care Survey: United States, 1975-1981 and 1985 Trends," *Vital and Health Statistics, Se*ries 13, No. 93, DHHS Pub. No. (PHS)88-1754 (Hyattsville, MD: 1988).

80. U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control, National Center for Health Statistics, "Data Systems of the National Center for Health Statistics," *Vital and Health Statistics, Series* 1, No. 23, (Hyattsville, MD: 1989).

81. Verbrugge, L. M., "Scientific and Professional Allies in Validity Studies," *Health Survey Research Methods-Conference Proceedings* (Rockville, MD: National Center for Health Services Research and Health Care Technology Assessment, September, 1989).

82. Verbrugge, L. M., and Balaban, D.J., "Patterns of Change in Disability and Well-Being," *Medical Care 27(3)(* suppl.):S128-S147, *1989.*

83. Veronesi, U., Banfi, A., Salvadori, B., et al., "Breast Conservation Is the Treatment of Choice in Small Breast Cancer: Long-Term Results of a Randomized Trial," *European Journal of Cancer* 26:668-670,1990.

84. Ware, Jr., J. E., Brook, R. H., Davies, A. R., et al., *Conceptualization and Measurement of Health for Adults in the Health Insurance Study: Volume I—Model of Health and Methodology,* Publication No. R-1987/ 1-HEW. (Santa Monica, CA: RAND Corp., 1987)

85. Ware, J., and Davies, A., *Scoring the Short-Form Mental Health Inventory* (MHI-5) (Santa Monica, CA: RAND Corp., 1983).

86. Ware, Jr., J.E., and Sherboume, C. D., "The MOS 36-Item Short-Form Health Survey (SF-36)—I: Conceptual Framework and Item Selection, ''Medical *Care* 30(6):473-483, 1992.

87. Ware, J. E., Snyder, M. R., Wright, R., et al., "Defining and Measuring Patient Satisfaction with Medical Care," *Evaluation and Program Planning* 6:247-263,1983.

88. Wasson, J. H., Cushman, C. C., Bruskewitz, R. C., et al., "A Structured Literature Review of Treatment for Localized Prostate Cancer," *Journal of the American Medical Association, 1993.*

89. Weisel, S. W., Tsourmas, N., Feffer, H. L., et al., "A Study of Computer-Assisted Tomography—1: The Incidence of Positive CAT Scans in an Asymptomatic Group of Patients," *Spine* 9(6):549-551, *1984.*

90. Wilson, R.G., Hart, A., and Dawes, P. J. D. K., "Mastectomy or Conservation: The Patient's Choice," *British Medical Journal 297(6657):* 1167-1169, 1988.