Interpreting the SF-36 Health Survey
Barbara Gandek, M.S., Scientist, Health Assessment Lab, Boston, MA

One of the most frequently asked questions about health surveys such as the SF-36® is "what do the numbers mean"? Health status scores are abstractions, which obtain meaning through knowledge of how the scores are related to other variables. Although we may not think about it, this is also true for more commonly used measures. For example, a temperature of 0°C is an abstract number, which becomes less abstract when it is associated with a cold feeling outside and rain turning to snow. Similarly, health status scores obtain meaning when they can be interpreted in relation to clinically and socially relevant variables.

Interpretation Strategies

Interpretation strategies can be grouped into three main categories: (1) content-based interpretation; (2) construct-based interpretation; and (3) criterion-based interpretation [1]. These strategies will be illustrated with examples from the SF-36 Health Survey, a widely-used generic health status measure [2,3]. The SF-36 measures eight concepts: physical functioning (PF), role limitations due to physical health (RP), bodily pain (BP), general health perceptions (GH), vitality (VT), social functioning (SF), role limitations due to emotional problems (RE), and general mental health (MH). Two summary measures of physical (PCS) and mental (MCS) health are constructed from the eight scales. The SF-36 has been used in a number of cardiac rehabilitation studies [4-10].

Content-based interpretation

Content-based interpretation uses information about item content and patterns of response choices to assign meaning to scores. Scores for people at the top or bottom of a scale can be interpreted by looking at the items and response choices that must be chosen to earn those scores. For example, someone at the top score of the SF-36 Physical Functioning (PF) scale does not have limitations in any of the SF-36 activities due to health. A person scoring at the bottom of the PF scale is very limited in all activities, including bathing and dressing. Scale scores in between these extremes can be interpreted in relation to responses to a single item from the scale, as can the two SF-36 summary measures. An example for the SF-36 Physical Component Summary (PCS) is provided in Figure 1. The percentage who reported any limitation in walking a block ranged from 81% for people scoring less than 30 on the PCS, to 0.6% for those scoring 60 or greater (Figure 1). Similar interpretation guidelines, which present data for narrower score ranges, are provided in the SF-36 manuals [2,3].

Construct-based interpretation

Constructs are abstract properties, such as physical or mental health, which are measured with specific instruments, such as the SF-36. Construct-based interpretation answers questions about the underlying meaning of health concepts, such as "where does a scale fit into a general model of health"? One way to determine this is to examine the relationship among scales within a questionnaire. For example, the SF-36 Physical Functioning, Role Physical, and Bodily Pain scales have high correlations with each other in general U.S. and Canadian populations, as do the Mental Health, Role Emotional, and Social Functioning scales. This pattern of relationships is indicative of a relationship between the scales and the underlying constructs of physical and mental health. That is, Physical Functioning, Role Physical, and Bodily Pain are primarily measures of physical health, while the other three scales are primarily measures of mental health. Research has demonstrated that scales that are the strongest measures of the physical health construct tend to do best in detecting the impact of interventions primarily thought to affect physical health [2,3]. Similarly, scales that are the strongest measures of mental health tend to do best in detecting the impact of mental health interventions.

Criterion-based interpretation

Criterion-based interpretation uses information on the relationship of scores to external variables to determine their meaning. Scores can be interpreted in relation to clinically and socially meaningful variables, such as job loss, utilization of health care services, likelihood of a clinical diagnosis, or death. For example, data from the US general population showed that 48% of respondents scoring between 30-34 on the PCS visited a medical doctor in the past month, compared to 26% of
respondents scoring between 45-49 [3]. Data from clinical studies showed that hip replacement patients had a 9.5 point improvement in PCS after surgery, while heart valve replacement patients had a 7.6 point improvement [3]. Two studies of patients in cardiac rehabilitation programs demonstrated an 8-point improvement in PCS at 3-month follow-up [9,10]. As the SF-36 continues to be used in a variety of clinical areas, this type of evidence will accumulate, facilitating interpretation of results from specific applications.

Norm-based comparisons are a type of criterion-based interpretation, in which scores are understood in relation to typical scores (norms). Norms can be calculated at an individual level, to determine if scores are typical of an individual under stable conditions. They also can be calculated at a group level, either from a sample of the general population or from normative data for a specific population (e.g., norms for patients in cardiac rehabilitation). Norm-based comparisons facilitate judgment as to whether an observed score is typical of what would be expected for that individual or group. Because health status is generally related to age and gender, normative values are often presented separately for specific age and gender groups [2,3].

Recent advances have made it easier to interpret SF-36 scores in relation to normative data. The SF-36 now can be scored so that all scales and summary measures are on the same metric, where 50 is the mean for the US general population and 10 is the standard deviation. Details are provided in [3] and at http://www.qualitymetric.com/innohome/norm.shtml. This scoring advance makes it easier to compare scale scores, which have different average values and standard deviations on the traditional 0-100 metric. For example, general population norms (using 0-100 scoring) range from 84 for PF to 61 for VT, and the standard deviations vary as well. Norm-based scoring equates all scores, so scores above 50 are better than the general population average for all scales and summary measures, while scores below 50 are worse.

The advantages of norm-based scoring can be seen with re-analysis of a study by Morrin et al., which examined the impact of duration in a cardiac rehabilitation program on health-related quality of life and coronary risk [10]. A total of 126 patients completed the SF-36 at entry into the program, and at 3 and 6 months. In their article, the authors helpfully provide published US general population normative data, for comparative purposes. However, because the norms themselves vary (with 0-100 scoring), it is difficult to tell at a glance where cardiac rehabilitation programs have the greatest impact. By converting scores to (50,10) scoring, the initial burden of coronary disease, and the impact on cardiac rehabilitation on various aspects of health status, can be seen more quickly. Baseline and 3-month SF-36 data from Morrin et al. are presented in Figure 2. Cardiac rehabilitation had a substantial (0.8 to 1 standard deviation) impact on many measures, including physical functioning, role limitations due to physical health problems, and the physical summary score.

Interpreting results from specific SF-36 studies can be done through use of the content, construct and criterion-based techniques illustrated here. Different audiences are likely to be more receptive to different types of interpretation guidelines. For example, employers may be more interested in what scores mean in terms of absenteeism, while clinicians may be more interested in comparative data for different treatments. Interpretation guidelines are compiled over time, both within and across diagnostic and treatment groups. Essentially, evidence about the interpretation of the SF-36 will accumulate as long as the survey is in use. For updated information about the SF-36, please visit www.sf-36.com.

References


