Title
Health-Related Quality of Life Measurement in Public Health.

Permalink
https://escholarship.org/uc/item/8jx5c4dm

Journal
Annual review of public health, 43(1)

ISSN
0163-7525

Authors
Kaplan, Robert M
Hays, Ron D

Publication Date
2022-04-01

DOI
10.1146/annurev-publhealth-052120-012811

Peer reviewed
Abstract
Patient-reported outcomes are recognized as essential for the evaluation of medical and public health interventions. Over the last 50 years, health-related quality of life (HRQoL) research has grown exponentially from 0 to more than 17,000 papers published annually. We provide an overview of generic HRQoL measures used widely in epidemiological studies, health services research, population studies, and randomized clinical trials [e.g., Medical Outcomes Study SF-36 and the Patient-Reported Outcomes Measurement Information System (PROMIS®)-29]. In addition, we review methods used for economic analysis and calculation of the quality-adjusted life year (QALY). These include the EQ-5D, the Health Utilities Index (HUI), the self-administered Quality of Well-being Scale (QWB-SA), and the Health and Activities Limitation Index (HALex). Furthermore, we consider hybrid measures such as the SF-6D and the PROMIS-Preference (PROPr). The plethora of HRQoL measures has impeded cumulative science because incomparable measures have been used in different studies. Linking among different measures and consensus on standard HRQoL measurement should now be prioritized. In addition, enabling widespread access to common measures is necessary to accelerate future progress.
INTRODUCTION

The goal of health care and preventive medicine is to improve health. Over the past 50 years, there has been growing recognition among researchers and clinicians that comprehensive measurement of health outcomes includes a combination of life expectancy and health-related quality of life (HRQoL) during the years prior to death. HRQoL refers to patient reports of functioning and well-being in physical, mental, and social domains of life. Functioning includes physical functioning, such as self-care (e.g., bathing, dressing, walking); role functioning, such as work-related activities (whether paid or not) such as housework and career; and social functioning, the extent to which one is able to interact with family and friends. Self-reports of functioning can be compared with other sources of data such as observations or performance measures. Well-being is more subjective than functioning and includes happiness, sadness, depression or anxiety (emotional well-being), pain, and lethargy.

In theory, quality of life refers to aspects of life that extend beyond health status, such as access to nutritional food and water. But the terms quality of life and HRQoL have often been used interchangeably. Publications under the “quality of life” PubMed Medical Subject Headings (MeSH) search term grew dramatically between 1972 and 2019 (Figure 1). In 1972, there were 0 publications, but the number of articles that use the quality of life keyword grew to 17,011 in 2019.

Despite its impressive growth, the field is divided over fundamental theoretical and methodological issues. This review concentrates on the conceptualization of the HRQoL construct and on some of the most common measures used to measure it. We focus on generic HRQoL measures. Within the space allowed, we could not do justice to the hundreds of disease-targeted measures such as the Arthritis Impact Measurement Scale (AIMS) (77) or the Minnesota Living with Heart Failure Questionnaire (99). Targeted measures are designed to be relevant to subgroups (e.g., people with diabetes, hypertension, seniors, women). Systematic reviews of instruments for many of the major diseases are available, including heart disease (28), diabetes (14), and breast cancer.

Figure 1

HRQoL indexes
Preference-weighted aggregate scores summarizing overall health
Generic health status profiles
Vector of health status domain scales
Disease-specific scales
Do not necessarily cover all health domains
Multitude of health indicators

More aggregated = summarization, evaluation
More disaggregated = explanation, description

Figure 2
Data pyramid for population health. Figure adapted with permission from Michael Wolfson and Dennis Fryback.

(81). For vision, so many studies have been published that there are now systematic reviews of the systematic reviews (2).

The plethora of HRQoL measures is both a blessing and a curse. On the one hand, researchers and clinicians have many high-quality options from which to choose. On the other hand, inconsistent use of measures contributes to poor replicability and the noncumulative nature of public health science (73). In particular, investigators use the HRQoL term to refer to different constructs, and the many measures of HRQoL may not be interchangeable.

Health outcomes include mortality (death rates or life expectancy) and morbidity-based indicators that count disease prevalence (127). For example, we can examine incidence and prevalence of HIV or coronavirus disease 2019 (COVID-19). Morbidity counts are commonly displayed in comparisons such as America’s health ratings (94) or the World Health Organization (WHO) core indicators (125). Morbidity measures also include self-reported health status instruments such as the SF-36 or SF-12, which we describe below.

A typology developed by Michael Wolfson helps clarify the differences between measurement systems. Wolfson (127) described a data pyramid for health measures (Figure 2). At the bottom of the pyramid are multiple health indicators such as rates of heart attacks or strokes in a community. The next level of the pyramid includes quality of life or functioning relevant to a specific condition, such as diabetes (95) or vision (106). One level up are generic measures of health status such as the Medical Outcomes Study (MOS) health questionnaires and the Patient-Reported Outcomes Measurement Information System (PROMIS®) measures. At the top of the pyramid are HRQoL indexes that combine morbidity and mortality and place each person on a continuum from dead to optimum function.

PROFILE MEASURES

Profile measures assess multiple aspects of HRQoL. Generic profile measures are relevant to respondents in general. They are analogous to intelligence tests in the sense that different people can be compared to one another because they have taken the same exam. Among the most widely used profile measures are the SF-36 and the PROMIS measures.
SF-36

The MOS SF-36 is the most widely used generic profile HRQoL measure to date (122). It includes 36 items selected from a large pool of items in the MOS (40). Twenty of the items are administered using a reporting interval covering the past 4 weeks. The SF-36 assesses eight health concepts with multi-item scales (35 items): physical functioning (10 items), role limitations caused by physical health problems (4 items), role limitations caused by emotional problems (3 items), social functioning (2 items), emotional well-being (5 items), energy/fatigue (4 items), pain (2 items), and general health perceptions (5 items). An additional single item assesses change in perceived health during the last 12 months.

The standard physical (PCS) and mental (MCS) component summary scales were derived by using a principal components analysis that forces physical and mental health to be uncorrelated (120). PCS and MCS scores can yield counterintuitive results. For example, a study of 536 primary care patients who initiated antidepressant treatment showed that physical functioning, role limitations caused by physical health, pain, and general health perceptions scales improved significantly by 0.28–0.49 SD units, but the PCS did not change significantly (113). Similar anomalies have been reported in multiple other studies (88). As a result, summary scores that represent the true correlation between mental and physical health have been derived (18, 40).

Psychometric Approaches: Classical Test Theory and IRT/CAT

Classical test theory dominated the analyses of HRQoL measures prior to the PROMIS project. Item response theory (IRT) methods began to be used routinely during the first decade of the twenty-first century (100). An advantage of IRT is a focus on evaluating the assumptions underlying the scoring of unidimensional scales (38).

Unidimensionality means that scale items measure a single construct. Local independence means that the items are uncorrelated with each other when the latent trait has been controlled for. Monotonicity means that the response categories representing lower levels of the construct (e.g., “limited a lot” in physical function) should be more likely to be selected by those with lower levels of the construct (physical function), and those representing higher levels of the construct should be more likely to be selected by those with higher levels (118).

There is a plethora of IRT models, but a main distinction among them is the number of item parameters that are estimated. The simplest is the Rasch model (one-parameter model) that estimates only a difficulty or threshold parameter. Two-parameter models estimate item difficulty and item discrimination (slope). The discrimination parameter is like an item-scale correlation and suggests how well an item represents the underlying construct. Other models estimate lower (“guessing”) and/or upper asymptote parameters used in education but not in HRQoL measures.

One of the advantages of IRT is that it allows for an essentially unlimited number of items (item bank) as the basis for computerized adaptive tests (CATs). At the start of the CAT, if nothing is known about the respondent, then an item that taps into the middle range of difficulty can be administered. On the basis of the individual’s response to the first item, the person’s score and standard error of measurement (SEM) are estimated using prior calibrations of the item bank. This approach is used to select an item that is likely to provide additional information about where the individual is located on the scale and to reduce the SEM. For example, if someone reports on the first item that they are unable to walk 50 yards, then the next item would not ask them if they could run a mile but instead would ask something that represented a lower level of physical function, such as whether they could walk 10 yards.

A major benefit of IRT is the ability to assess differential item functioning or equivalence of measurement by subgroup (e.g., age, gender, race/ethnicity) or mode of administration (110, 116).
Another feature of IRT that is rarely exploited is the extent to which a person’s pattern of item responses is consistent with the underlying model (101). An example in PROMIS was an individual who reported on 13 physical function items that they were able to do them (including the ability to run 5 miles) without any difficulty but also reported having a little difficulty getting out of bed (35). This discrepancy could represent carelessness in responding or perhaps a condition such as back pain, which selectively impacts getting out of bed but does not affect other activities.

While proponents often assert that simple-summed scores are ordinal and Rasch scores are interval-level measurement, the IRT latent trait metric is simply a rescaling of the optimally weighted summed score metric (weights that maximize internal consistency). Simple-summed scores and IRT scores are extremely highly correlated, and associations with other variables tend to be robust (111, 118).

PROMIS

The National Institutes of Health Roadmap initiative funded a cooperative agreement to develop, evaluate, and standardize item banks to measure HRQoL across different medical conditions and in the general population (9). PROMIS developed item banks calibrated using IRT that allows for flexibility in administration in a variety of formats, including short forms and CATs (38). The compelling scientific basis for PROMIS measures is likely to lead to greater usages than competing measures.

The PROMIS® suite of measures (http://www.healthmeasures.net/explore-measurement-systems/promis) includes the PROMIS-29 v2.1 brief profile measure, which is analogous to the SF-36 (8). The PROMIS-29 v2.1 profile assesses pain intensity using a single 0–10 numeric rating item and 7 health domains (physical function, fatigue, pain interference, depressive symptoms, anxiety, ability to participate in social roles and activities, and sleep disturbance) using four items per domain. Physical and mental health summary scores are also available (41). In addition, there is a PROMIS preference-based measure, the PROPr (12), which is discussed later in this article.

The flexibility in using PROMIS measures is vast. In addition to the PROMIS-29 profile, additional options can increase the number of items in each domain from four to six (PROMIS-43) or eight (PROMIS-57). Users can also elect to administer entire item banks or a subset of items within the banks using computer-adaptive testing (see below). Because the items within each domain are calibrated on the same underlying metric, users can also pick and choose a subset of items (65).

METHODS OF ECONOMIC EVALUATION

HRQoL measures can be used in a wide variety of applications. One of the most important applications is for estimating the economic medical and health care interventions (107). The three most widely used methods for economic evaluation are cost-benefit analysis (CBA), cost-effectiveness analysis (CEA), and cost-utility analysis (CUA). The theoretical basis and conclusions that can be drawn from each can differ (87).

A CBA compares the cost of an intervention with the dollar value of the benefits that accrue from it. The analysis does not typically require the measurement of HRQoL. Both inputs and outputs are measured in dollar values. In CEA, the costs of a program contain elements like those in a CBA; however, the output is a measure of health benefit. For example, we might want to compare the value of assigning people to health insurance plans that include first dollar coverage with plans that require substantial copayments for services. The outcomes might be a measure of health status among groups assigned to each insurance alternative. Outcomes used to evaluate
the effect of the insurance alternatives could be as different as blood pressure and depression. With units of benefit that are not directly comparable to one another, direct comparison between programs is not possible. CUA is a special case of CEA in which the benefit is measured in terms of the quality-adjusted life year (QALY). A QALY represents 1 year of life, adjusted for its quality or value based on health. Quality is assessed across a patient’s physical, social, and psychological domains, with QALY weights empirically assigned to the various dimensions. A year in perfect health would be assigned a QALY value of 1.0, whereas a year of less-than-perfect health would be assigned a value less than 1.0. In theory, a wide range of programs and interventions can be compared on the basis of their ability to increase this common metric.

**Measuring Outcomes for Economic Analysis: Theory**

Utility weighting that assigns levels of wellness along the continuum between dead and optimum function is an essential part of CUA. Survival analysis, for example, ignores the impact of conditions that reduce HRQoL but do not shorten life expectancy. Arthritis and depression may have profound effects on HRQoL but little effect on mortality. Survival measures will miss the important benefits of treatments for these conditions. A comprehensive evaluation of health outcome must be able to distinguish between positive and negative effects of treatment and their side effects, prevention, or lifestyle. Overall, we want to know whether the patient benefits from the services they receive.

**First-Generation HRQoL Measures for Estimating QALYs**

Most approaches for obtaining QALYs are similar (30) and involve several steps (51). First, patients are classified according to levels of functioning and well-being. Human value studies are used to place the observable health states onto a preference continuum ranging from 0.0 (dead) to 1.0 ("perfect" health). Duration of stay in various health states may be noted. For example, having a cough or a headache for 1 day should not be scored the same as having the problem for 1 year.

**HRQoL Measures Used in Economic Analysis**

The most widely used preference measures include the EQ-5D, the Health Utilities Index (HUI), the Quality of Well-Being Scale (QWB), and the Health and Activities Limitation Index (HALex). In addition, hybrid methods such as the SF-6D can map utilities onto profile measures. We briefly review these methods next.

**EQ-5D.** The EQ-5D was developed by a collaborative group from Western Europe known as the EuroQol group (68). The concept of a common EuroQol was stimulated by the common European currency, the Euro. The original version of the EuroQol had 14 health states in 6 different domains. In addition, surveys in England, Sweden, and the Netherlands were used to place health states on a continuum ranging from dead (0.0) to perfect health (1.0). The next iteration was known as the EQ-5D (31, 45). Although the EQ-5D is comprehensive and easy to use, there were problems with ceiling effects. Substantial numbers of people obtain the highest possible score. The latest version of the measure, the EQ-5D-5L, changed the rating system to include five new levels for each of the domains: no problems, some problems, moderate problems, severe problems, and extreme problems. The older version with three response levels is labeled the EQ-5D-3L. To compare the EQ-5D-5L with the EQ-5D-3L, individuals with similar expected levels of health status (two or more chronic conditions) were compared in two separate years (117). Confirming better sensitivity, particularly among people with better health end of the continuum, the
EQ-5D-5L had fewer people with the highest possible score. Other studies confirmed its greater sensitivity in specific patient populations, such as hidradenitis suppurativa (3). Recent papers document the validity of the EQ-5D-5L in different countries, including Japan (112), Poland (80), and Russia (44).

Health Utilities Index. The HUI was developed in Canada by Torrance, Feeny, Furlong and associates (19, 20, 21). The HUI Mark I (HUI1) was developed for studies in the neonatal intensive care unit. The measure had 960 unique health states. In 1992, the HUI Mark II (HUI2) included 24,000 unique health states. The HUI Mark III (HUI3), released in 1995, had 972,000 health states. Eight components of the HUI3 include vision (six levels), hearing (six levels), speech (five levels), ambulation (six levels), dexterity (six levels), emotion (five levels), cognition (six levels), and pain (five levels). Multiplying the number of levels across the eight dimensions gives the 972,000 states. Using multi-attribute utility scaling methods, judges evaluate levels of wellness associated with each level of each domain. A multi-attribute model is used to map preference for the 972,000 possible states onto the 0.0–1.0 continuum.

The HUI has been used in many population and clinical studies to evaluate outcomes, chronic pruritus (124), total joint replacement, Duchenne muscular dystrophy (71), and multiple sclerosis (76). The HUI also continues to attract methodological evaluations. There is a crosswalk for mapping HUI utilities onto the SF-12 (76), and evaluations of the properties of the utility functions are ongoing, including cross-cultural assessments (89). Recently, data from the Canadian Community Health Survey were used to create Canadian utility scores for 17 chronic conditions. Among the conditions, utilities for asthma have the least detrimental weight (closest to 1.0), whereas those for Alzheimer’s disease had the most negative weight (32).

Self-Administered Quality of Well-Being Scale. The QWB-SA integrates several components into a single score. First, individuals are classified on scales of mobility, physical activity, social activity, and symptom/problem complexes. Weights for these levels of functioning were obtained from a community sample (54, 57, 60, 61, 63, 90, 98). The QWB-SA is unique among the measures in its inclusion of a comprehensive list of symptom/problem complexes. Health problems ranging from missing limbs to runny noses are captured, which allows for greater sensitivity at the top (healthy) end of the continuum.

The QWB has been used in numerous clinical trials and studies to evaluate medical and surgical therapies in conditions such as chronic obstructive pulmonary disease (COPD) (53), HIV (52, 59), cystic fibrosis (90, 91), diabetes mellitus (58), atrial fibrillation (27), lung transplantation (115), arthritis (50, 60), end-stage renal disease (104), cancer (47), depression (96, 97), and several other conditions (57). Furthermore, the method has been used for health resource allocation modeling and has served as the basis for an innovative experiment on rationing of health care by the state of Oregon (46, 48). The self-administered form of the QWB (QWB-SA) was developed more recently. It has been shown to be highly correlated with the interviewer-administered QWB and to have equivalent psychometric properties (61).

HALex. While European investigators invested in a standardized HRQoL instrument, the EQ-5D, and the Canadians have de facto adopted the HUI3 as a national survey instrument, the United States has no one standardized instrument used broadly in national data sets. However, the United States has several national surveys of health: the Longitudinal Study of Aging (LSOA), the Health and Retirement Study (HRS), the National Health and Nutrition Examination Study (NHANES), the National Health Interview Survey (NHIS), and the Medical Expenditure Panel Survey (MEPS). Gold and colleagues developed an ad hoc measure based on information
collected for the NHIS: the HALex (15), known as “years of healthy life” (17). The HALex has two dimensions: a seven-level classification of activities and function limitations ranging from “no limitations” to “limited in instrumental activities of daily living (IADLs)” to “limited in activities of daily living (ADLs),” and self-rated overall health using the five-level, “excellent, very good, good, fair, poor” classification. The resulting classification scheme has $7 \times 5 = 35$ health states. Building on prior attempts to develop a national composite index for health states (16), the 35 states were weighted to correspond with expected utilities from the HUI1 (15). The HALex has been shown to be correlated with other HRQoL measures (29).

**Utility Weighting Systems for Profile Measures**

A variety of weighting methods are now available for estimating utilities for profile measures. These systems facilitate the use of profile measures for CUA. The most widely recognized methods include the SF-6D and the PROPr.

**SF-6D.** As noted earlier, the SF-36 measures eight health concepts: physical functioning, physical health–related role limitations, bodily pain, general health perceptions, vitality, social functioning, and mental health–related role limitations. The SF-36 and the shorter SF-12 version were not designed for use in cost-utility studies. The first approach to put the SF-36 on the 0 to 1 preference continuum was described by Brazier et al. (6). He obtained independent utility ratings of 249 health states derived from combinations of SF-36 components. The ratings were used to estimate utilities for 18,000 different combinations of SF-36 subscales. The measure became known as the SF-6D. In addition to its use in the United Kingdom (82), the SF-6D has been evaluated in the United States (11), Portugal (22), Hong Kong (70), and Lebanon (66). In 2020, an updated SF-6D classification system was introduced to match the latest versions of the SF-36 version 2.

**PROPr.** One of the most important developments was the creation of utilities that allow PROMIS measures to be used in CUA. Hanmer et al. (34) developed PROMIS-Preference (PROPr), a generic preference-based scoring system for the PROMIS measures. They demonstrated that PROPr was more sensitive to minor variations in health in comparison to the HUI2 and the EQ-5D-3L. PROPr is sensitive to variations in kidney disease, with significant and substantial correlations with SF-6D, EQ-5D-5L and several different indicators of renal functioning (128). PROPr is also associated with the social determinants of health. In one evaluation using 4,142 participants, PROPr was significantly correlated with education, income, food and financial insecurity, and social interactions (33).

**Utility Measurement Methods**

Measures used for economic analysis require utility assessment. However, the methods used to obtain these weights are not uniform. The best-known method is the standard gamble (SG). Using this technique, a respondent is given a hypothetical choice between continued life in a current state of health or a gamble that would result in perfect health (with a probability of $p$) or death (with a probability of $1 - p$). An alternative method is the time trade-off (TTO), in which respondents are asked about the amount of time that they would be willing to give up to be in a better health state (74). Many researchers consider the TTO easier to implement in clinical studies than the standard gamble (74). A third approach involves the use of simple rating scales (RS) or a visual analog scale. Subjects are required to rate health conditions on a scale ranging from 0 to 10 or from 0 to 100. Ideally, the anchors are clearly defined with 0 equal to dead and 100 equal to perfect health. Unlike
with SG and TTO, subjects are not required to make a choice between alternatives. In addition, rating scales do not consider attitude toward risk nor do they incorporate time horizons (56).

Comparisons among SG, TTO, and RS show that the methods yield different preference weights (67). Preferences from SG are usually higher than those obtained using TTO. In turn, TTO preferences are higher than those measured using RS (23).

Some authors have argued that the SG is the best approach because of its linkage to theoretical concepts of utility (123). However, some evidence indicates that people believe that TTO better reflects their preferences (72). Others have argued that TTO is the most credible validity criterion (102). In terms of feasibility, TTO often fails to produce meaningful preferences, which has led some observers to prefer the RS method (7, 55). Furthermore, Kattan and colleagues have developed newer methods to adjust TTO to consider subjective fears of death and declining health (64). Investigators have expressed concerns about the cognitive burden that SG and TTO place on patients (49, 75).

The National Health Measurement Study
The National Health Measurement Study (NHMS) was designed to compare and cross-calibrate preference-based HRQoL indexes using a variety of methods. The study concentrated on the EQ-5D, HUI, QWB-SA, HALex, and SF-6D. While each index uses profiles of health states composed of similar dimensions (e.g., physical function, mental health, social function, pain, other symptoms), they are based on different survey items. Each index is scored so that perfect health is represented as 1.0 and dead is represented as 0.0; some preference-based HRQoL indexes allow health states to be valued worse than being dead, with scores less than 0.0. The indexes apply different utility weighting methods. One of the goals of the NHMS was to compare scores obtained using different measures that measure the same construct.

Although all the measures had been used in many studies, rarely had more than one measure been used in the same study. The NHMS provided the opportunity for head-to-head comparison of the instruments. To compare the methods, Fryback and his team administered the 5 measures to a national sample of 3,844 35–89-year-old US adults using random digit dialing telephone survey methods. People aged 65+ years and telephone exchanges with high proportions of African American households were oversampled. In addition to completing the five measures, respondents indicated whether they had been diagnosed with coronary heart disease, stroke, diabetes, arthritis, eye disease, sleep disorder, chronic respiratory disease, clinical depression or anxiety disorder, gastrointestinal ulcer, thyroid disorder, and/or severe chronic back pain.

The study demonstrated that the mean scores differed across the indexes, with males consistently obtaining higher (better health) scores than did females. Median scores were comparable across the indexes (range 0.79–0.88), except for the QWB-SA, which obtained lower values (median = 0.64). Estimates of the standard error of measurement were similar across the indexes. In addition, estimations of test–retest standard deviation, a separate index of reliability, were similar and varied between ~0.60–0.77 across the measures (93). The indexes were substantially correlated with one another (range r = 0.65 to r = 0.71, excluding HUI1 versus HUI2, r = 0.89). For all indexes, scores declined with increasing age with one exception. For the 65–74-years age group, there was a deviation from the declining pattern for both men and women (24). Other analyses suggested that all indexes performed as expected in relation to other health risk factors. For example, all indexes showed worse HRQoL with obesity. Yet, the pattern differed across indexes. African American respondents had scores that were influenced less by obesity, yet this pattern was not equally detected by the different indexes (5).

The study also identified methodological concerns. For example, a trial built into the NHMS compared responses of those randomized to complete their follow-up questionnaires by mail
versus those who were to complete their questionnaires by telephone interview. Those assigned to the telephone administration condition reported significantly better HRQoL scores for all measures except for the QWB-SA. Some of these differences were as large as one-half standard deviation (37).

In addition to the national survey, the NHMS evaluated responsiveness of the measures to change in two clinical populations: cataract surgery and heart failure. In the cataract study, which used a pre- versus postsurgery design, there were significant improvements for all indexes except the SF-6D. For patients being treated for heart failure, only the SF-6D demonstrated an improvement between baseline and one month following initiation of treatment, and only the QWB-SA detected a significant improvement between one month and six months (63). Another analysis considered agreement between the indexes for which patients would be classified as having improved, having remained stable, or having worsened following treatment. Overall, agreement between these classifications tended to be poor. These results suggest that the indexes, although substantially correlated with one another, may lead to different conclusions about individual patients who are improving, staying the same, or getting worse (21). One overall takeaway from the NHMS is the need for better harmonization across measures. Although each of the measures has been well studied and well evaluated, they produce similar but not directly comparable results. Thus, tables that list the cost per QALY for different investments in health can be misleading. The bottom line is that we need to develop better consensus for common metrics to be used in CUA.

USES OF HRQOL MEASURES

In this section, we review a selected set of applications for HRQoL measures. Several federal agencies use HRQoL measures to monitor populations. We offer two examples from Department of Health and Human Services agencies.

The Centers for Disease Control and Prevention

The most widely used single item (“In general, how would you rate your health?”) has been administered for decades in the United States on the NHIS and the Behavioral Risk Factor Surveillance System (42). PROMIS global health items were included in the NHIS, HealthStyles, and pilot data from the Division of Behavioral Surveillance (DBS) in the Population Health Surveillance and Informatics Program Office (PHSIPO) of the Centers for Disease Control and Prevention (CDC) (103). Similar global physical health and mental health scores were found in three of the four administrations, but the NHIS yielded more positive scores because it uses interviewer administration.

Center for Medicare and Medicaid Services

The Medicare Health Outcomes Survey (MHOS) is an annual survey administered to a random sample of 1,000 Medicare beneficiaries from each managed care plan under contract with the Centers for Medicare and Medicaid Services. The MHOS included the SF-36 survey when it commenced in 1998, but beginning in 2006, the Veterans RAND (VR-12) (109) was administered instead owing to proprietary issues (43) associated with the SF-36. The National Cancer Institute Surveillance, Epidemiology, and End Results (SEER) data set was linked to the MHOS to produce the SEER-MHOS data set (1).

Analyses of the SEER-MHOS data provided evidence of the negative impact of cancer diagnoses with HRQoL and the unique negative associations beyond that of older age, less education, and lower household income (10). In addition, one study found that depressive symptoms had the largest unique association with the SF-6D preference-based score, followed by arthritis of the
hip, COPD/asthma, stroke, and sciatica (39). In addition, most cancer types were significantly associated with the SF-6D score, with significant negative weights ranging from −0.01 to −0.02 on the 0–1 health utility scale. Distant stage of cancer was associated with large decrements in the SF-6D, ranging from −0.04 (prostate) to −0.08 (female breast). Because depressive symptoms were represented, to some extent, on both sides of the equation (36), the authors reran the model, dropping depressive symptoms, and found no impact on the interpretation of the associations for the other 20 comorbid conditions that had significant unique associations with the SF-6D score.

EVALUATING TREATMENT EFFECTIVENESS

In this section, we review selected applications of HRQoL measurement to evaluate health status and economic outcomes in clinical trials.

Economic Analysis in Clinical Trials

There has been increasing interest in estimating the cost utility of treatments evaluated in randomized clinical trials, but HRQoL is not usually measured. Instead, researchers attempt to impute health outcomes on the basis of other variables. This practice results in problematic estimates of benefit that are biased toward showing a treatment benefit. Ideally, an HRQoL utility measure should be included in the clinical trial. Unfortunately, only a few major trials have been prospectively designed to include HRQoL measures. Examples of such instances include the National Emphysema Treatment Trial (NETT), the diabetes prevention program, and the Look AHEAD trial.

National Emphysema Treatment Trial. The NETT evaluated lung volume reduction surgery (LVRS) in comparison with medical management of patients who have moderate to severe emphysema. Patients with moderate to severe COPD were assigned to the combination of LVRS along with maximal medical care or to maximal medical care alone. All participants were followed prospectively for vital status over 15 years. Data were available on 140 high-risk patients, with the QWB administered through 6 years of follow-up.

Early in the study, it appeared that some patients were at high risk for postsurgical death (83). Enrollment was discontinued for high-risk patients, but those already participating in the study were followed. Through the first 3 years of follow-up, surgical patients in the high-risk group had a significantly higher probability of dying. However, the curves crossed after 3 years. Thereafter, the probability of death was lower for those who had received surgery. HRQoL data suggested an advantage of surgery for the first 5 years of follow-up. However, QALYs favored medical management for the first few years of follow-up and favored surgery after 4 years. For high-risk patients who survived the first 30 days, deaths were lower, and eventually QALYs were superior compared with medical treatment (62). CUA suggested that lung volume reduction surgery produced a QALY at $190,000 over 3 years. If the modeling was extended to 10 years, the cost per QALY was $53,000. A subgroup that had predominantly upper lobe emphysema and low exercise capacity after pulmonary rehabilitation showed a cost per QALY of only $21,000 (84). These estimates suggest that the surgery produces benefit at costs comparable to, or lower than, several other well-established interventions.

Diabetes prevention program. Another example of a prospective CEA is the diabetes prevention program (129). In this randomized clinical trial, patients at risk for type 2 diabetes were randomly assigned to one of three conditions: intensive lifestyle modification, metformin, or placebo. The diabetes prevention program included 3,234 adults with impaired glucose tolerance. The patients
were evaluated using the QWB-SA prior to randomization and at annual intervals over 3 years. Over the course of the study, individuals randomly assigned to the lifestyle intervention accrued 0.05 more QALYs than did those assigned a regular dose of Metformin. Among the three interventions, the lifestyle approach was the most expensive (total cost $27,065 in 2000 US dollars). Metformin was less expensive ($25,937), whereas the placebo was the least expensive option ($23,525). Although both interventions offer significant benefits over placebo or doing nothing, the cost per QALY for the lifestyle invention was significantly lower than that for metformin. That is, the lifestyle intervention was more expensive, but it offered significantly better value for money.

Look AHEAD. One cautionary tale comes from the Action for Health in Diabetes (Look AHEAD) trial (129). The purpose of this investigation was to determine the impact and cost-effectiveness of an intensive lifestyle intervention compared with usual support and education for overweight or obese adults with type 2 diabetes; 4,827 participants were randomly assigned to one of these two conditions and then followed prospectively for 9 years. HRQoL was assessed using the HUI2/HUI3 and the SF-6D. In addition, the investigators collected outcome data using a feeling thermometer. The study is important because that the costs of long-term programs to manage weight can be very high. For example, intensive lifestyle intervention costs $6,666 more per person in comparison with education. Of interest, QALYs gained were not statistically significant using any of the measures, and there was no difference between the groups in mortality. But cost per QALY was relatively low (i.e., high value) when outcomes were measured using a feeling thermometer measure. From a practical perspective, there was no evidence that the intervention produced HRQoL or mortality benefits. But when using a highly subjective self-rating, it was possible for investigators to obtain an estimate that made the treatment look favorable. However, the substantial likelihood of bias necessitates cautious interpretation of the results. As with the NHMS, the finding argues in favor of developing international standards for the harmonization of outcome measurement for economic analysis.

Use in Clinical Practice

Wasson and colleagues’ use of the Dartmouth COOP Charts is the pioneering work for using HRQoL measures in clinical practice (85, 121). They demonstrated the necessity for providing guidance for interpreting HRQoL scores and support materials for interventions to promote the use and effectiveness of HRQoL measures. The studies to date indicate that use of HRQoL measures in clinical practice improves provider–patient communication and shared decision making, but the evidence about impact on change in HRQoL is mixed (114).

Many institutions, including the University of Utah, Northwestern University, Stanford University, Washington University, and Partners Healthcare (4), are now using PROMIS measures at the point of care. Some have suggested that use of HRQoL measures can improve the quality of health care and that these measures will grow in importance as policies and payment systems emphasize patient-centered care (92).

DIRECTIONS FOR FUTURE RESEARCH

Association for Psychological Science President Walter Mischel described some of the difficulty of achieving a cumulative science as the “toothbrush problem” (78). Theories and measures are like toothbrushes: “[N]o self-respecting person wants to use anyone else’s.” Career advancement, including achieving tenure, depends on originality. Creating a new measure is given more credit than using an established method. Building a cumulative science is difficult when investigators measure outcomes using noncomparable methods (79).
The case of economic analysis provides a useful illustration. In 1996, a distinguished panel created methodological guidelines for cost-effectiveness studies in medicine and health care (30). The standards were updated in 2016 (86, 87). Both publications offered detailed recommendations on the standardization of methods. Although both panels proclaimed that utility-based methods are needed for the analyses, they demurred on suggesting which HRQoL instruments should be used. As demonstrated in the NHMS, the methods are substantially correlated with one another, but they do not produce the same scores (26). The measures use different items, are built on different domains, and use different methods to obtain utility weights (26). Nevertheless, investigators commonly report “league tables” that compare the cost/utility for different investments in health care (126). Policy makers sometimes take these comparisons seriously (105), arguing, for example, that interventions yielding between $50,000 and $150,000 per QALY are of intermediate value (13). There is often little recognition that the comparison is built on the application of noncomparable measures. There have been several attempts to develop crosswalks between measures (25). These make it possible, for example, to predict HUI scores from the EQ-5D. Although these comparisons are attractive, the translations tend to be quite imperfect.

There are many generic HRQoL measures and approaches for estimating preference-based single summaries for use in evaluating health care outcomes. IRT has been used to solve the “Tower of Babel” problem of different profile measures by linking scores from one measure to others (108). Future efforts are needed to understand variations in results from different preference-based measures and to evaluate whether scores from one preference-based measure can be accurately predicted from another.

To move the field forward, we first need to make existing measures freely available so that all investigators can access them without high user fees (43). Next, it is essential to develop consensus around the optimal approach. Doing so may require additional analysis of existing measures. Most of the measures evaluate common constructs, although the actual questions differ. An alternative is to use the best currently available approach. Applications of the PROPr scoring system will enable PROMIS to be used for CEA and CUA (12, 69).

CONCLUSIONS
Over the last half-century, progress on HRQoL measurement has been remarkable. Profile and utility-based measures are now abundant, extensively evaluated, translated into multiple languages, and used in multiple studies. Further efforts to build and evaluate item banks that draw on the content of the existing measures may be needed (119). But the main challenge over the next decade may be consolidation rather than expansion. To develop a cumulative science of health outcomes, it may be necessary to achieve consensus around one or two standardized approaches that build on the lessons provided by the literature we have summarized.

DISCLOSURE STATEMENT
The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS
The authors gratefully acknowledge many colleagues who have shaped our thinking over the years. These include Jim Bush, David Cella, David Feeny, Dennis Fryback, Theodore Ganiats, Marthe Gold, and Paul Kind.
LITERATURE CITED


   uality of life: a systematic review of systematic reviews. JAMA Ophthalmol. 139:526–41


tions between obesity and nine health-related quality-of-life measures. Qual. Life Res. 20:665–74


10. Clau


14. El Achi


