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

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

Annual Review of Public Health

Health-Related Quality of Life Measurement in Public Health

Robert M. Kaplan[1] and Ron D. Hays[2]

[1]Clinical Excellence Research Center, Department of Medicine, Stanford University, Stanford, California, USA; email: Bob.kaplan@stanford.edu

[2]Division of General Internal Medicine, Department of Medicine, University of California, Los Angeles, California, USA

Annu. Rev. Public Health 2022. 43:9.1–9.19

The *Annual Review of Public Health* is online at publhealth.annualreviews.org

https://doi.org/10.1146/annurev-publhealth-052120-012811

Keywords

health-related quality of life, measurement, functional status, cost-effectiveness, quality-adjusted life year, QALY

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

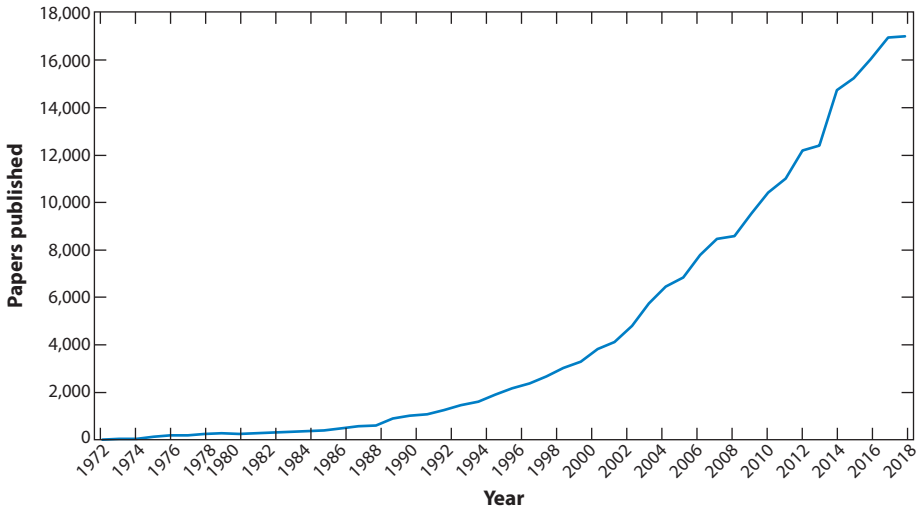

Figure 1

PubMed citations published in 1972–2019 under Medical Subject Headings (MeSH) "quality of life."

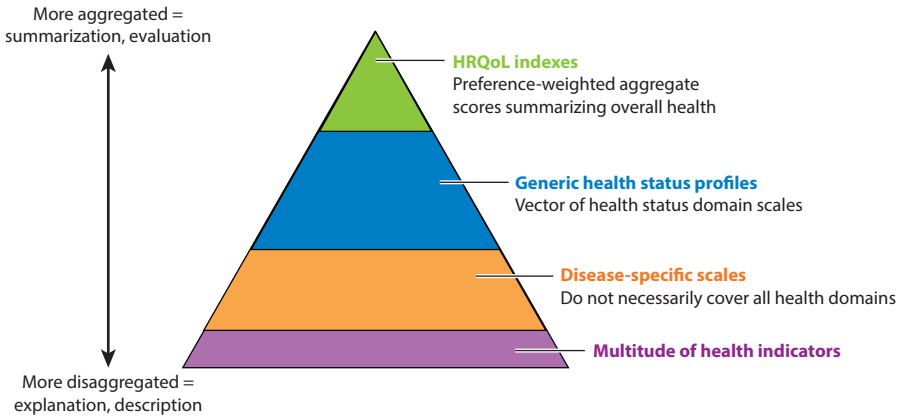

Figure 2

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 a 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-summated 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-summated 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

1. Ambs A, Warren JL, Bellizzi KM, Topor M, Haffer SCC, Clauser SB. 2008. Overview of the SEER—Medicare Health Outcomes Survey linked dataset. *Health Care Financ. Rev.* 29:5–21
2. Assi L, Chamseddine F, Ibrahim P, Sabbagh H, Rosman L, et al. 2021. A global assessment of eye health and quality of life: a systematic review of systematic reviews. *JAMA Ophthalmol.* 139:526–41
3. Bató A, Brodszky V, Gergely LH, Gáspár K, Wikonkál N, et al. 2021. The measurement performance of the EQ-5D-5L versus EQ-5D-3L in patients with hidradenitis suppurativa. *Qual. Life Res.* 30:1477–90
4. Baumhauer JF. 2017. Patient-reported outcomes—are they living up to their potential? *N. Engl. J. Med.* 377:6–9
5. Bentley TGK, Palta M, Paulsen AJ, Cherepanov D, Dunham NC, et al. 2011. Race and gender associations between obesity and nine health-related quality-of-life measures. *Qual. Life Res.* 20:665–74
6. Brazier J, Roberts J, Tsuchiya A, Busschbach J. 2004. A comparison of the EQ-5D and SF-6D across seven patient groups. *Health Econ.* 13:873–84
7. Broome L. 1993. The goal is quality improvement. *Nurs. Manag.* 24:51–52
8. Cella D, Choi SW, Condon DM, Schalet B, Hays RD, et al. 2019. PROMIS® adult health profiles: efficient short-form measures of seven health domains. *Value Health* 22:537–44
9. Cella D, Riley W, Stone A, Rothrock N, Reeve B, et al. 2010. Initial adult health item banks and first wave testing of the Patient-Reported Outcomes Measurement Information System (PROMIS™) network: 2005–2008. *J. Clin. Epidemiol.* 63:1179–94
10. Clauser SB, Arora NK, Bellizzi KM, Haffer SCC, Topor M, Hays RD. 2008. Disparities in HRQOL of cancer survivors and non-cancer managed care enrollees. *Health Care Financ. Rev.* 29:23–40
11. Craig BM, Pickard AS, Stolk E, Brazier JE. 2013. US valuation of the SF-6D. *Med. Decis. Mak.* 33:793–803
12. Dewitt B, Feeny D, Fischhoff B, Cella D, Hays RD, et al. 2018. Estimation of a preference-based summary score for the Patient-Reported Outcomes Measurement Information System: The PROMIS®-Preference (PROPr) scoring system. *Med. Decis. Mak.* 38:683–98
13. Dubois RW. 2016. Cost-effectiveness thresholds in the USA: Are they coming? Are they already here? *J. Comp. Eff. Res.* 5:9–12
14. El Achhab Y, Nejjari C, Chikri M, Lyoussi B. 2008. Disease-specific health-related quality of life instruments among adults diabetic: a systematic review. *Diabetes Res. Clin. Pract.* 80:171–84
15. Erickson P. 1998. Evaluation of a population-based measure of quality of life: the Health and Activity Limitation Index (HALex). *Qual. Life Res.* 7:101–14
16. Erickson P, Kendall EA, Anderson JP, Kaplan RM. 1989. Using composite health status measures to assess the nation's health. *Med. Care* 27:S66–76
17. Erickson P, Wilson R, Shannon I. 1995. Years of healthy life. *Healthy People 2000 Stat. Notes* 1995(7):1–15
18. Farivar SS, Cunningham WE, Hays RD. 2007. Correlated physical and mental health summary scores for the SF-36 and SF-12 Health Survey, V. 1. *Health Qual. Life Outcomes* 5:54
19. Feeny D, Furlong W, Mulhern RK, Barr RD, Hudson M. 1999. A framework for assessing health-related quality of life among children with cancer. *Int. J. Cancer Suppl.* 12:2–9
20. Feeny D, Furlong W, Torrance GW, Goldsmith CH, Zhu Z, et al. 2002. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. *Med. Care* 40:113–28
21. Feeny D, Spritzer K, Hays RD, Liu H, Ganiats TG, et al. 2012. Agreement about identifying patients who change over time: cautionary results in cataract and heart failure patients. *Med. Decis. Mak.* 32:273–86
22. Ferreira LN, Ferreira PL, Pereira LN, Brazier J, Rowen D. 2010. A Portuguese value set for the SF-6D. *Value Health* 13:624–30
23. Froberg DG, Kane RL. 1989. Methodology for measuring health-state preferences—II: scaling methods. *J. Clin. Epidemiol.* 42:459–71
24. Fryback DG, Dunham NC, Palta M, Hanmer J, Buechner J, et al. 2007. US norms for six generic health-related quality-of-life indexes from the National Health Measurement Study. *Med. Care* 45:1162

25. Fryback DG, Palta M, Cherepanov D, Bolt D, Kim J-S. 2007. *Cross-walks among five self-reported summary health utility indexes: progress and prospects*. Presented at Annual Meeting of the Society for Medical Making, Pittsburgh, PA, Oct. 20–24
26. Fryback DG, Palta M, Cherepanov D, Bolt D, Kim J-S. 2010. Comparison of 5 health-related quality-of-life indexes using item response theory analysis. *Med. Decis. Mak.* 30:5–15
27. Ganiats TG, Palinkas LA, Kaplan RM. 1992. Comparison of Quality of Well-Being scale and Functional Status Index in patients with atrial fibrillation. *Med. Care* 30:958–64
28. Garin O, Ferrer M, Pont À, Rué M, Kotzeva A, et al. 2009. Disease-specific health-related quality of life questionnaires for heart failure: a systematic review with meta-analyses. *Qual. Life Res.* 18:71–85
29. Gold M, Franks P, Erickson P. 1996. Assessing the health of the nation. The predictive validity of a preference-based measure and self-rated health. *Med. Care* 34:163–77
30. Gold MR, Siegel JE, Russell LB, Weinstein MC. 1996. *Cost-Effectiveness in Health and Medicine*. Oxford, UK: Oxford Univ. Press
31. Gudex C, Dolan P, Kind P, Williams A. 1996. Health state valuations from the general public using the visual analogue scale. *Qual. Life Res.* 5:521–31
32. Guertin JR, Humphries B, Feeny D, Tarride J-E. 2018. Health Utilities Index Mark 3 scores for major chronic conditions: population norms for Canada based on the 2013–2014 Canadian Community Health Survey. *Health Rep.* 29:12–19
33. Hanmer J. 2021. Cross-sectional validation of the PROMIS-Preference scoring system by its association with social determinants of health. *Qual. Life Res.* 30:881–89
34. Hanmer J, Dewitt B, Yu L, Tsevat J, Roberts M, et al. 2018. Cross-sectional validation of the PROMIS-Preference scoring system. *PLOS ONE* 13:e0201093
35. Hays RD. 2011. Applying item response theory for questionnaire evaluation. In *Question Evaluation Methods: Contributing to the Science of Data Quality*, ed. J Madans, K Miller, A Maitland, G Willis, pp. 125–35. Hoboken, NJ: Wiley
36. Hays RD, Fayers PM. 2021. Overlap of depressive symptoms with health-related quality-of-life measures. *Pharmacoeconomics* 39:627–30
37. Hays RD, Kim S, Spritzer KL, Kaplan RM, Tally S, et al. 2009. Effects of mode and order of administration on generic health-related quality of life scores. *Value Health* 12:1035–39
38. Hays RD, Morales LS, Reise SP. 2000. Item response theory and health outcomes measurement in the 21st century. *Med. Care* 38:II28–42
39. Hays RD, Reeve BB, Smith AW, Clauser SB. 2014. Associations of cancer and other chronic medical conditions with SF-6D preference-based scores in Medicare beneficiaries. *Qual. Life Res.* 23:385–91
40. Hays RD, Sherbourne CD, Mazel RM. 1993. The RAND 36-item health survey 1.0. *Health Econ.* 2:217–27
41. Hays RD, Spritzer KL, Schalet BD, Cella D. 2018. PROMIS®-29 v2. 0 profile physical and mental health summary scores. *Qual. Life Res.* 27:1885–91
42. Hays RD, Spritzer KL, Thompson WW, Cella D. 2015. U.S. general population estimate for "excellent" to "poor" self-rated health item. *J. Gen. Intern. Med.* 30:1511–16
43. Hays RD, Weech-Maldonado R, Teresi JA, Wallace SP, Stewart AL. 2018. Commentary: copyright restrictions versus open access to survey instruments. *Med. Care* 56:107–10
44. Hołownia-Voloskova M, Tarbastaev A, Golicki D. 2021. Population norms of health-related quality of life in Moscow, Russia: the EQ-5D-5L-based survey. *Qual. Life Res.* 30:831–40
45. Hurst NP, Kind P, Ruta D, Hunter M, Stubbings A. 1997. Measuring health-related quality of life in rheumatoid arthritis: validity, responsiveness and reliability of EuroQol (EQ-5D). *Br. J. Rheumatol.* 36:551–59
46. Kaplan RM. 1993. *Allocating health resources in California: learning from the Oregon experiment*. Calif. Policy Semin. Brief, Vol. 5, Calif. Policy Semin., Berkeley
47. Kaplan RM. 1993. Quality of life assessment for cost/utility studies in cancer. *Cancer Treat. Rev.* 19(Suppl. A):85–96
48. Kaplan RM. 1994. Value judgment in the Oregon Medicaid experiment. *Med. Care* 32:975–88

49. Kaplan RM. 1995. Utility assessment for estimating quality-adjusted life years. In *Valuing Health Care: Costs, Benefits, and Effectiveness of Pharmaceuticals and Other Medical Technologies*, ed. FA Sloan, pp. 31–60. Cambridge, UK: Cambridge Univ. Press
50. Kaplan RM, Alcaraz JE, Anderson JP, Weisman M. 1996. Quality-adjusted life years lost to arthritis: effects of gender, race, and social class. *Arthritis Care Res*. 9:473–82
51. Kaplan RM, Anderson JP. 1988. A general health policy model: update and applications. *Health Serv. Res.* 23:203–35
52. Kaplan RM, Anderson JP, Patterson TL, McCutchan JA, Weinrich JD, et al. 1995. Validity of the Quality of Well-Being Scale for persons with human immunodeficiency virus infection. HNRC Group. HIV Neurobehavioral Research Center. *Psychosom. Med.* 57:138–47
53. Kaplan RM, Atkins CJ, Timms R. 1984. Validity of a quality of well-being scale as an outcome measure in chronic obstructive pulmonary disease. *J. Chronic Dis.* 37:85–95
54. Kaplan RM, Bush JW, Berry CC. 1976. Health status: types of validity and the Index of Well-being. *Health Serv. Res.* 11:478–507
55. Kaplan RM, Ernst JA. 1983. Do category rating scales produce biased preference weights for a health index? *Med. Care* 21:193–207
56. Kaplan RM, Feeny D, Revicki DA. 1993. Methods for assessing relative importance in preference based outcome measures. *Qual. Life Res.* 2:467–75
57. Kaplan RM, Ganiats TG, Sieber WJ, Anderson JP. 1998. The Quality of Well-Being Scale: critical similarities and differences with SF-36. *Int. J. Qual. Health Care* 10:509–20
58. Kaplan RM, Hartwell SL, Wilson DK, Wallace JP. 1987. Effects of diet and exercise interventions on control and quality of life in non-insulin-dependent diabetes mellitus. *J. Gen. Intern. Med.* 2:220–28
59. Kaplan RM, Patterson TL, Kerner DN, Atkinson JH, Heaton RK, Grant I. 1997. The Quality of Well-Being scale in asymptomatic HIV-infected patients. HNRC Group. HIV Neural Behavioral Research Center. *Qual. Life Res.* 6:507–14
60. Kaplan RM, Schmidt SM, Cronan TA. 2000. Quality of well being in patients with fibromyalgia. *J. Rheumatol.* 27:785–89
61. Kaplan RM, Sieber WJ, Ganiats TG. 1997. The Quality of Well-being Scale: comparison of the interviewer-administered version with a self-administered questionnaire. *Psychol. Health* 12:783–91
62. Kaplan RM, Sun Q, Naunheim KS, Ries AL. 2014. Long-term follow-up of high-risk patients in the National Emphysema Treatment Trial. *Ann. Thorac. Surg.* 98:1782–89
63. Kaplan RM, Tally S, Hays RD, Feeny D, Ganiats TG, et al. 2011. Five preference-based indexes in cataract and heart failure patients were not equally responsive to change. *J. Clin. Epidemiol.* 64:497–506
64. Kattan MW, Fearn PA, Miles BJ. 2001. Time trade-off utility modified to accommodate degenerative and life-threatening conditions. *Proc. AMIA Symp*. 2001:304–8
65. Khanna D, Krishnan E, Dewitt EM, Khanna PP, Spiegel B, Hays RD. 2011. The future of measuring patient-reported outcomes in rheumatology: Patient-Reported Outcomes Measurement Information System (PROMIS). *Arthritis Care Res.* 63(Suppl. 11):S486–90
66. Kharroubi SA, Beyh Y, El Harake MD, Dawoud D, Rowen D, Brazier J. 2020. Examining the feasibility and acceptability of valuing the Arabic version of SF-6D in a Lebanese population. *Int. J. Environ. Res. Public Health* 17:1037
67. Kim S-H, Lee S-I, Jo M-W. 2017. Feasibility, comparability, and reliability of the standard gamble compared with the rating scale and time trade-off techniques in Korean population. *Qual. Life Res.* 26:3387–97
68. Kind P. 1997. The performance characteristics of EQ-5D, a measure of health related quality of life for use in technology assessment [abstract]. *Annu. Meet. Int. Soc. Technol. Assess. Health Care* 13:81
69. Klapproth CP, Fischer F, Merbach M, Matthias R, Obbarius A. 2021. Validity, reliability, and ceiling and floor effects of the PROMIS Preference score (PROPr) in patients with rheumatological and psychosomatic conditions. *Research Square*. **https://doi.org/10.21203/rs.3.rs-478767/v1**
70. Lam CLK, Brazier J, McGhee SM. 2008. Valuation of the SF-6D health states is feasible, acceptable, reliable, and valid in a Chinese population. *Value Health* 11:295–303
71. Landfeldt E, Lindberg C, Sejersen T. 2020. Improvements in health status and utility associated with ataluren for the treatment of nonsense mutation Duchenne muscular dystrophy. *Muscle Nerve* 61:363–68

72. Lipman SA, Brouwer WB, Attema AE. 2020. What is it going to be, TTO or SG? A direct test of the validity of health state valuation. *Health Econ*. 29:1475–81
73. Loken E, Gelman A. 2017. Measurement error and the replication crisis. *Science* 355:584–85
74. Lugnér AK, Krabbe PF. 2020. An overview of the time trade-off method: concept, foundation, and the evaluation of distorting factors in putting a value on health. *Expert Rev. Pharmacoecon. Outcomes Res.* 20:331–42
75. Makarov DV, Holmes-Rovner M, Rovner DR, Averch T, Barry MJ, et al. 2017. American Urological Association and Society for Medical Decision Making Quality Improvement Summit 2016: shared decision making and prostate cancer screening. *Urol. Pract.* **https://doi.org/10.1016/j.urpr.2017.11.005**
76. Marrie RA, Dufault B, Tyry T, Cutter GR, Fox RJ, Salter A. 2020. Developing a crosswalk between the RAND-12 and the health utilities index for multiple sclerosis. *Multiple Scler. J.* 26:1102–10
77. Meenan RF, Mason JH, Anderson JJ, Guccione AA, Kazis LE. 1992. AIMS2. The content and properties of a revised and expanded Arthritis Impact Measurement Scales Health Status Questionnaire. *Arthritis Rheum.* 35:1–10
78. Mischel W. 2008. The toothbrush problem. *APS Observer* 21. **https://www.psychologicalscience.org/observer/the-toothbrush-problem**
79. Mischel W. 2009. Becoming a cumulative science. *APS Observer* 22. **https://www.psychologicalscience.org/observer/becoming-a-cumulative-science**
80. Młyńczak K, Golicki D. 2021. Validity of the EQ-5D-5L questionnaire among the general population of Poland. *Qual. Life Res.* 30:817–29
81. Montazeri A. 2008. Health-related quality of life in breast cancer patients: a bibliographic review of the literature from 1974 to 2007. *J. Exp. Clin. Cancer Res.* 27:32
82. Mulhern BJ, Bansback N, Norman R, Brazier J. 2020. Valuing the SF-6Dv2 classification system in the United Kingdom using a discrete-choice experiment with duration. *Med. Care* 58:566–73
83. Natl. Emphysema Treat. Trial Res. Group. 2001. Patients at high risk of death after lung-volume–reduction surgery. *N. Engl. J. Med.* 345:1075–83
84. Natl. Emphysema Treat. Trial Res. Group. 2003. Cost effectiveness of lung-volume–reduction surgery for patients with severe emphysema. *N. Engl. J. Med.* 348:2092–102
85. Nelson EC, Eftimovska E, Lind C, Hager A, Wasson JH, Lindblad S. 2015. Patient reported outcome measures in practice. *BMJ* 350:g7818
86. Neumann PJ, Kim DD, Trikalinos TA, Sculpher MJ, Salomon JA, et al. 2018. Future directions for cost-effectiveness analyses in health and medicine. *Med. Decis. Mak.* 38:767–77
87. Neumann PJ, Sanders GD, Russell LB, Siegel JE, Ganiats TG. 2016. *Cost-Effectiveness in Health and Medicine*. Oxford, UK: Oxford Univ. Press
88. Nortvedt MW, Riise T, Myhr K-M, Nyland HI. 2000. Performance of the SF-36, SF-12, and RAND-36 summary scales in a multiple sclerosis population. *Med. Care* 38:1022–28
89. Noto S, Shiroiwa T, Kobayashi M, Murata T, Ikeda S, Fukuda T. 2020. Development of a multiplicative, multi-attribute utility function and eight single-attribute utility functions for the Health Utilities Index Mark 3 in Japan. *J. Patient-Rep. Outcomes* 4:1–8
90. Orenstein DM, Kaplan RM. 1991. Measuring the quality of well-being in cystic fibrosis and lung transplantation. The importance of the area under the curve. *Chest* 100:1016–18
91. Orenstein DM, Pattishall EN, Nixon PA, Ross EA, Kaplan RM. 1990. Quality of well-being before and after antibiotic treatment of pulmonary exacerbation in patients with cystic fibrosis. *Chest* 98:1081–84
92. Øvretveit J, Zubkoff L, Nelson EC, Frampton S, Knudsen JL, Zimlichman E. 2017. Using patient-reported outcome measurement to improve patient care. *Int. J. Qual. Health Care* 29:874–79
93. Palta M, Chen H-Y, Kaplan RM, Feeny D, Cherepanov D, Fryback DG. 2011. Standard error of measurement of 5 health utility indexes across the range of health for use in estimating reliability and responsiveness. *Med. Decis. Mak.* 31:260–69
94. Park H, Roubal AM, Jovaag A, Gennuso KP, Catlin BB. 2015. Relative contributions of a set of health factors to selected health outcomes. *Am. J. Prev. Med.* 49:961–69
95. Pereira EV, Tonin FS, Carneiro J, Pontarolo R, Wiens A. 2020. Evaluation of the application of the Diabetes Quality of Life Questionnaire in patients with diabetes mellitus. *Arch. Endocrinol. Metab.* 64:59–65

96. Pyne JM, Patterson TL, Kaplan RM, Gillin JC, Koch WL, Grant I. 1997. Assessment of the quality of life of patients with major depression. *Psychiatr. Serv.* 48:224–30
97. Pyne JM, Patterson TL, Kaplan RM, Ho S, Gillin JC, et al. 1997. Preliminary longitudinal assessment of quality of life in patients with major depression. *Psychopharmacol. Bull.* 33:23–29
98. Pyne JM, Sieber WJ, David K, Kaplan RM, Rapaport MH, Williams DK. 2003. Use of the quality of well-being self-administered version (QWB-SA) in assessing health-related quality of life in depressed patients. *J. Affect. Disord.* 76:237–47
99. Rector TS, Cohn JN. 1992. Assessment of patient outcome with the Minnesota Living with Heart Failure questionnaire: reliability and validity during a randomized, double-blind, placebo-controlled trial of pimobendan. *Am. Heart J.* 124:1017–25
100. Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, et al. 2007. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Med. Care* 45(5 Suppl. 1):S22–31
101. Reise SP. 1990. A comparison of item- and person-fit methods of assessing model-data fit in IRT. *Appl. Psychol. Meas.* 14:127–37
102. Richardson J. 1994. Cost utility analysis: What should be measured? *Soc. Sci. Med.* 39:7–22
103. Riley W, Hays RD, Kaplan RM, Cella D. 2013. *Sources of comparability between probability sample estimates and nonprobability web sample estimates*. Presented at Proceedings of the 2013 Federal Committee on Statistical Methodology (FCSM) Research Conference, Washington, DC, Nov. 4–6. **https://nces.ed.gov/FCSM/pdf/B4_Riley_2013FCSM.pdf**
104. Rocco MV, Gassman JJ, Wang SR, Kaplan RM. 1997. Cross-sectional study of quality of life and symptoms in chronic renal disease patients: the Modification of Diet in Renal Disease Study. *Am. J. Kidney Dis.* 29:888–96
105. Romanens M, Sudano I, Szucs T, Adams A. 2017. Medical costs per QALY of statins based on Swiss Medical Board assumptions. *Cardiovasc. Med.* 20:96–100
106. Rosen PN, Kaplan RM, David K. 2005. Measuring outcomes of cataract surgery using the Quality of Well-Being Scale and VF-14 Visual Function Index. *J. Cataract Refract. Surg.* 31:369–78
107. Sanders GD, Neumann PJ, Basu A, Brock DW, Feeny D, et al. 2016. Recommendations for conduct, methodological practices, and reporting of cost-effectiveness analyses: second panel on cost-effectiveness in health and medicine. *JAMA* 316:1093–103
108. Schalet BD, Lim S, Cella D, Choi SW. 2021. Linking scores with patient-reported health outcome instruments: a validation study and comparison of three linking methods. *Psychometrika* 86(3):717–46
109. Selim A, Rogers W, Qian S, Rothendler JA, Kent EE, Kazis LE. 2018. A new algorithm to build bridges between two patient-reported health outcome instruments: the MOS SF-36® and the VR-12 Health Survey. *Qual. Life Res.* 27:2195–206
110. Setodji CM, Reise SP, Morales LS, Fongwa MN, Hays RD. 2011. Differential item functioning by survey language among older Hispanics enrolled in Medicare managed care: a new method for anchor item selection. *Med. Care* 49:461–68
111. Shallcross AJ, Lu NY, Hays RD. 2020. Evaluation of the psychometric properties of the Five Facet of Mindfulness Questionnaire. *J. Psychopathol. Behav. Assess.* 42:271–80
112. Shiroiwa T, Ikeda S, Noto S, Fukuda T, Stolk E. 2021. Valuation Survey of EQ-5D-Y based on the international common protocol: development of a value set in Japan. *Med. Decis. Mak.* 41:597–606
113. Simon GE, Revicki DA, Grothaus L, Vonkorff M. 1998. SF-36 summary scores: Are physical and mental health truly distinct? *Med. Care* 36:567–72
114. Snyder CF, Aaronson NK, Choucair AK, Elliott TE, Greenhalgh J, et al. 2012. Implementing patient-reported outcomes assessment in clinical practice: a review of the options and considerations. *Qual. Life Res.* 21:1305–14
115. Squier HC, Ries AL, Kaplan RM, Prewitt LM, Smith CM, et al. 1995. Quality of well-being predicts survival in lung transplantation candidates. *Am. J. Respir. Crit. Care Med.* 152:2032–36
116. Teresi JA, Wang C, Kleinman M, Jones RN, Weiss DJ. 2021. Differential item functioning analyses of the Patient-Reported Outcomes Measurement Information System (PROMIS®) measures: methods, challenges, advances, and future directions. *Psychometrika* 86(3):674–711

117. Thompson AJ, Turner AJ. 2020. A comparison of the EQ-5D-3L and EQ-5D-5L. *Pharmacoeconomics* 38:575–91
118. Uy V, Hays RD, Xu JJ, Fayers PM, Auerbach AD, et al. 2020. Do the unlabeled response categories of the Minnesota Living with Heart Failure Questionnaire satisfy the monotonicity assumption of simple-summated scoring? *Qual. Life Res.* 29:1349–60
119. Voshaar MO, Vonkeman H, Courvoisier D, Finckh A, Gossec L, et al. 2019. Towards standardized patient reported physical function outcome reporting: linking ten commonly used questionnaires to a common metric. *Qual. Life Res.* 28:187–97
120. Ware J, Kosinski M, Keller S. 1994. *SF-36 Physical and Mental Health Summary Scales: A User's Manual*. Boston: Health Inst.
121. Wasson J, Hays R, Rubenstein L, Nelson E, Leaning J, et al. 1992. The short-term effect of patient health status assessment in a health maintenance organization. *Qual. Life Res.* 1:99–106
122. Weinberger M, Samsa GP, Hanlon JT, Schmader K, Doyle ME, et al. 1991. An evaluation of a brief health status measure in elderly veterans. *J. Am. Geriatr. Soc.* 39:691–94
123. Weinstein MC, Stason WB. 1985. Cost-effectiveness of interventions to prevent or treat coronary heart disease. *Annu. Rev. Public Health* 6:41–63
124. Whang KA, Khanna R, Williams KA, Mahadevan V, Semenov Y, Kwatra SG. 2021. Health-related QOL and economic burden of chronic pruritus. *J. Investig. Dermatol.* 141:754–60.e1
125. WHO (World Health Organ.). 2015. *Global reference list of 100 core health indicators*. Rep., WHO, Geneva
126. Wilson N, Davies A, Brewer N, Nghiem N, Cobiac L, Blakely T. 2019. Can cost-effectiveness results be combined into a coherent league table? Case study from one high-income country. *Popul. Health Metrics* 17:10
127. Wolfson M. 2014. Measuring the health component of quality of life. *Stat. J. IAOS* 30:193–207
128. Zhang J, Dewitt B, Tang E, Breitner D, Saqib M, et al. 2021. Evaluation of PROMIS Preference Scoring System (PROPr) in patients undergoing hemodialysis or kidney transplant. *Clin. J. Am. Soc. Nephrol.* 16:1328–36
129. Zhang P, Atkinson KM, Bray GA, Chen H, Clark JM, et al. 2021. Within-trial cost-effectiveness of a structured lifestyle intervention in adults with overweight/obesity and type 2 diabetes: results from the Action for Health in Diabetes (Look AHEAD) Study. *Diabetes Care* 44:67–74