Research White Paper

A Primer for Systematic Reviewers on the Measurement of Functional Status and Health-Related Quality of Life in Older Adults

AHRQ
Agency for Healthcare Research and Quality
Advancing Excellence in Health Care • www.ahrq.gov

Research White Paper

A Primer for Systematic Reviewers on the Measurement of Functional Status and Health-Related Quality of Life in Older Adults

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

Contract No. 290-2007-10057-I

Prepared by:
Kaiser Permanente Research Affiliates Evidence-based Practice Center
Center for Health Research
Portland, OR

Investigators:
David H. Feeny, Ph.D.
Elizabeth Eckstrom, M.D., M.P.H.
Evelyn P. Whitlock, M.D., M.P.H.
Leslie A. Perdue, M.P.H.

AHRQ Publication No. 13-EHC128-EF
September 2013

This report is based on research conducted by the Kaiser Permanente Research Affiliates Evidence-based Practice Center (EPC) under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. 290-2007-10057-I). The findings and conclusions in this document are those of the authors, who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

The information in this report is intended to help health care decisionmakers—patients and clinicians, health system leaders, and policymakers, among others—make well informed decisions and thereby improve the quality of health care services. This report is not intended to be a substitute for the application of clinical judgment. Anyone who makes decisions concerning the provision of clinical care should consider this report in the same way as any medical reference and in conjunction with all other pertinent information, i.e., in the context of available resources and circumstances presented by individual patients.

This report may be used, in whole or in part, as the basis for development of clinical practice guidelines and other quality enhancement tools, or as a basis for reimbursement and coverage policies. AHRQ or U.S. Department of Health and Human Services endorsement of such derivative products may not be stated or implied.

Persons using assistive technology may not be able to fully access information in this report. For assistance contact EffectiveHealthCare@ahrq.hhs.gov.

David H. Feeny has a proprietary interest in Health Utilities Incorporated, Dundas, Ontario, Canada. HUInc. distributes copyrighted Health Utilities Index (HUI) materials and provides methodological advice on the use of HUI. None of the other investigators have any affiliations or financial involvement that conflicts with the material presented in this report.

Suggested citation: Feeny DH, Eckstrom E, Whitlock EP, Perdue LA. A Primer for Systematic Reviewers on the Measurement of Functional Status and Health-Related Quality of Life in Older Adults. (Prepared by the Kaiser Permanente Research Affiliates Evidence-based Practice Center under Contract No. 290-2007-10057-I.) AHRQ Publication No. 13-EHC128-EF. Rockville, MD: Agency for Healthcare Research and Quality. September 2013. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

Preface

The Agency for Healthcare Research and Quality (AHRQ), through its Evidence-based Practice Centers (EPCs), sponsors the development of evidence reports and technology assessments to assist public- and private-sector organizations in their efforts to improve the quality of health care in the United States. The National Cancer Institute provided funding for this report through an inter-agency agreement.

The reports and assessments provide organizations with comprehensive, science-based information on common, costly medical conditions and new health care technologies and strategies. The EPCs systematically review the relevant scientific literature on topics assigned to them by AHRQ and conduct additional analyses when appropriate prior to developing their reports and assessments.

To bring the broadest range of experts into the development of evidence reports and health technology assessments, AHRQ encourages the EPCs to form partnerships and enter into collaborations with other medical and research organizations. The EPCs work with these partner organizations to ensure that the evidence reports and technology assessments they produce will become building blocks for health care quality improvement projects throughout the Nation. The reports undergo peer review and public comment prior to their release as a final report.

AHRQ expects that the EPC evidence reports and technology assessments will inform individual health plans, providers, and purchasers as well as the health care system as a whole by providing important information to help improve health care quality.

We welcome comments on this Research White Paper. They may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, or by email to epc@ahrq.hhs.gov.

Richard G. Kronick, Ph.D.
Director
Agency for Healthcare Research and Quality

Jean Slutsky, P.A., M.S.P.H.
Director, Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Stephanie Chang, M.D., M.P.H.
Director
Evidence-based Practice Program
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Suchitra Iyer, Ph.D.
Task Order Officer
Evidence-based Practice Program
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Acknowledgments

The authors thank Dr. Gordon H. Guyatt, Dr. David B. Reuben, Dr. Jennifer Lin, Dr. Ron Klein, Dr. Tracy Wolff, Dr. Erica S. Breslau, Dr. Elliot J. Roth, Dr. Robert Kane, and participants in the Research In Progress Seminar Series, Division of General Internal Medicine and Geriatrics, Oregon Health & Science University, June 1, 2010, for their constructive comments and suggestions on the work reported here. Further, the authors thank Tracy Beil, Jill Pope, Heather Baird, and Debra Burch for their assistance with the preparation of the manuscript.

Peer Reviewers

Mary Butler, Ph.D., M.B.A.
Division of Health Policy & Management
School of Public Health
University of Minnesota
Minneapolis, MN

Kirstie L. Haywood, D.Phil.
Senior Research Fellow (Patient Reported Outcomes)
Royal College of Nursing Research Institute
School of Health and Social Studies
University of Warwick
Coventry, United Kingdom

Christine McDonough, Ph.D., P.T.
Research Assistant Professor
Health & Disability Research Institute
Department of Health Policy and Management
Boston University School of Public Health, Boston University
Boston, MA

Mari Palta, Ph.D.
Professor of Population Health Sciences and of Biostatistics and Medical Informatics
Department of Population Health Sciences
School of Medicine and Public Health
University of Wisconsin–Madison
Madison, WI

Kathleen W. Wyrwich, Ph.D.
Senior Research Leader
Evidera
Bethesda, MD

A Primer for Systematic Reviewers on the Measurement of Functional Status and Health-Related Quality of Life in Older Adults

Structured Abstract

Objectives. Provide a primer for systematic reviewers, clinicians, and researchers on assessing functional status and health-related quality of life (HRQL) in older adults. Systematic reviewers are increasingly focusing on interventions that address the problems of older people, who often have functional impairments and multiple morbidities. Key outcomes are function and HRQL. The paper provides an overview of the methods for assessing function and HRQL, and evidence on the measurement properties of prominent measures.

Methods. The paper provides an overview of the methods for assessing function and HRQL, and evidence on the measurement properties of prominent instruments.

Results. Key measurement properties include construct validity (does the instrument measure what it is supposed to measure?), responsiveness (the ability to detect meaningful change) and interpretation (is the magnitude of change trivial or important?). Special challenges in older adult populations include sometimes sparse evidence on the measurement properties; using proxy respondents; a paucity of evidence on the magnitude of change that is patient-important; and threats to detecting patient-important changes due to floor and ceiling effects.

Discussion. While further study of the measurement properties of measures in older populations is needed, studies of older adults should include measures of HRQL and function. Further, to generate rigorous evidence on effectiveness, older adults should be included in randomized controlled clinical trials. HRQL evidence from natural-history cohorts is important in interpreting results from intervention studies.

Contents

Introduction

A key to improving the health of our aging population is developing evidence-based guidelines that can inform best practices at the patient, health system, and policy levels. The field of systematic review has evolved to include sophisticated meta-analytic techniques and highly structured evidence reviews. However, evidence-based guidelines have focused largely on single diseases and general populations, and have left gaps in recommendations for older, functionally impaired populations with multiple morbidities.[1] Systematic reviewers have often focused on objective outcomes such as mortality, with less consideration to health-related quality of life (HRQL) and functional outcomes. Yet, these patient reported outcomes could be very useful—both to evaluate the effectiveness of interventions to improve health in older adults, and as a means of defining risk status and identifying important subgroups for analyses.

This paper will identify important issues in using evidence from these measures in systematic reviews, and interpret these issues for clinicians, researchers, and systematic reviewers. In Section 2, we briefly define important relevant measurement properties and summarize evidence on the measurement properties of common measures used to assess patient-reported outcomes, including the classification of measures, the populations studied, reliability, validity, special considerations for older adults, floor and ceiling effects, and using proxy respondents. Section 3 focuses on how to interpret functional status and HRQL evidence. Sections 4 and 5 discuss implications for researchers and systematic reviewers; Section 6 provides a summary and conclusions.

The paper highlights several challenges for systematic reviewers in synthesizing evidence to improve HRQL and reduce functional decline in older adults. First, the evidence on how well various measures perform in studies of older adults is often sparse. Second, because some older adults are not able to respond on their own behalf, investigators must often rely on family members and friends acting as proxy respondents. Third, evidence on the magnitude of change that is patient-important to older adults is often lacking. Finally, floor and ceiling effects often attenuate the usefulness of many measures.

Patient-Reported Outcomes, Health-Related Quality of Life, and Function: An Overview of Measurement Properties

In general, we rely on patient reports to assess HRQL and function. This section provides a discussion of the most important considerations when using evidence derived from the application of such measures. But first it is necessary to clarify terminology and provide definitions for the important relevant concepts and measurement properties.

Terminology

There is considerable heterogeneity in the terms used to describe HRQL and functional status. Recently the United States Food and Drug Administration introduced the term patient-reported outcomes, PROs.[2] A key component of the FDA definition is that the measure conveys information reported by the patient that is not filtered by an observer or clinician. In the United Kingdom the term patient-reported outcome measures, PROMs, is widely used. Some authors use the terms HRQL, health status, PROs, and PROMs interchangeably; we and many others do not. Rather we provide the following definitions.

Health Status: A person's current state of health. Typically that includes functional status, morbidity, physiologic outcomes, and some notion of well-being.[3]

Functional Status: Starfield: "The capacity to engage in activities of daily living and social activities".[4]

Frailty: Fried's definition is the presence of at least three of five factors: (1) unintentional weight loss (10 pounds or more in a year), (2) general feeling of exhaustion, (3) weakness (as measured by grip strength), (4) slow walking speed, and (5) low levels of physical activity.[5] Frailty is a risk factor for further decline in functional status and mortality, and can be associated with a wide variety of chronic conditions.

The concepts health status, functional status, and frailty, an important type of functional status for older populations, focus on a description of the current state of health of the subject. As noted below in the definition of HRQL used in this paper, the concept of HRQL includes health status but goes further by including the value attached to that health status.

Health-Related Quality of Life: There are a wide variety of definitions of HRQL. Some focus on the domains of health status that comprise HRQL, usually including physical health, mental health, social and role function, and pain and discomfort. Patrick and Erickson provide a useful definition (1993, p 22).[3] "Health-related quality of life is the value assigned to duration of life as modified by the impairments, functional states, perceptions, and social opportunities that are influenced by disease, injury, treatment, or policy."

Classification of Health-Related Quality of Life Measures

One taxonomy focuses on the types of persons for whom the measure is applicable.[6] Generic measures typically include both physical and mental health, are applicable to virtually any adult population, and can be used to make comparisons across diseases and conditions. There are two major categories of generic measures:[7] health profiles such as the Short-Form 36[8] and preference-based measures such as the Health Utilities Index.[9] Each of these will be discussed in more detail below. Specific measures are applicable to people with a particular disease (breast cancer), condition (frailty), or symptom (pain). Specific measures are often more responsive to

change than generic measures[10,11] but may not capture the effects of comorbidities, do not allow for comparisons across conditions, and thus have limited usefulness for cost-effectiveness analyses and other broader analyses. Some generic measures have condition-specific adaptations.[12]

Measures can also be classified by their intended purpose.[6,13,14] Evaluative measures capture "within person change" over time. Discriminative measures detect differences among groups (or individuals) at a point in time. Many measures of functional status were designed for this purpose. In practice, however, most measures are used for both purposes. Because systematic reviewers are interested in assessing the effectiveness of interventions, our focus is on evaluative applications of measures and the measurement properties that are important for assessing change over time.

It is also useful to note that there are three major intellectual paradigms upon which most measures of functional status and HRQL are based: psychometric, clinimetric, and economics/decision science.[15] The psychometric paradigm draws from psychology and typically relies on a latent-variable model.[16] In this paradigm, the underlying construct is not measured directly but rather the items in a measure reflect that construct. The clinimetric tradition builds a measure by selecting items that are important to patients with that condition or problem; this approach is often used to develop specific measures. Finally the economics and decision science paradigm, like the clinimetric tradition, selects domains and items on the basis of their importance to patient or members of the general population. The economics/decision science paradigm also focuses on the value attached to the health state, typically on a scale in which dead = 0.00 and perfect health = 1.00, thus enabling the integration of morbidity and mortality. There is considerable cross-fertilization among the three paradigms. Examples of measures based on each of these intellectual traditions are described below.

Quality-Adjusted Survival

The goal of interventions is to improve functional status or HRQL outcomes for older adults or reduce the rate of decline in their functional status or HRQL. That is, the goal is to improve quality-adjusted survival.[17,18] A unique feature of preference-based measures is their ability to integrate mortality and morbidity and provide estimates of quality-adjusted survival or quality-adjusted life years gained. Preference-based (utility) measures are on a scale in which 0.00 = dead and perfect health = 1.00. Preference or utility scores are derived directly using choice-based techniques such as the standard gamble and time-tradeoff or through the use of multi-attribute utility measures.[19] In the standard gamble, the subject is given a choice between remaining in an impaired state of health for sure or taking a lottery with a probability p of achieving perfect health and probability 1-p of dead. The probability at which the subject is indifferent between the lottery and the sure thing provides an estimate of the value attached to the sure-thing health state. Similarly, in the time-tradeoff, the subject places value on a health state by determining the number of years in that state she/he would be willing to give up to enjoy a shorter period in perfect health.[20] In the multi-attribute approach, the subject completes a questionnaire based on the measure; examples of multi-attribute measures include the EQ-5D[21] and Health Utilities Index (HUI).[9] The health status of the subject obtained by completing the questionnaire is then valued using a scoring function for that measure based on community preferences. Given their ability to provide estimates of quality-adjusted survival, preference-based measures have a special role in evaluating interventions in older populations. Further detail on preference-based measures can be found in Torrance 1986.[19,22,23]

Definitions

There are three key categories of measurement properties: reliability, validity, and responsiveness (see Table 1 at the end of the paper for brief definitions).

Reliability. A reliable measure is consistent and reproducible. *Internal consistency* is the extent to which items intended to assess health or functional status in a particular domain are correlated with each other and not correlated with items intended to measure other domains. Internal consistency is often measured with Cronbach's alpha. Scores > 0.70 are usually considered to have acceptable internal consistency for group comparisons.[24]

Intra- and Inter-Observer Reliability. This form of reliability examines the agreement between two raters—for instance, self-assessment at two points in time (intra-rater) or self- and proxy-assessment (inter-rater). The intra-class correlation coefficient (ICC [continuous response scale]) or kappa statistic (categorical responses) is used to assess the extent of agreement; kappas and ICCs > 0.70 are generally regarded as acceptable.[24]

Test-Retest Reliability. Test-retest reliability examines the agreement among scores in stable persons at two points in time. The interval between testing is generally one to two weeks—long enough that the person is unlikely to recall their previous response and short enough that it is unlikely the condition of the person has changed. Again, ICCs > 0.70 are regarded as acceptable for group comparisons. A good measure provides stable scores for stable persons.

Content Validity. Content validity is the "extent to which the items are sensible and reflect the intended domain of interest."[13] Does the content of the measure make sense? Are the items included relevant to the domain of interest? Do the items cover the full range relevant to that domain? Are the items comprehensible to respondents? There is no formal statistical test to evaluate content validity. In practice, content validity is evaluated using a structured set of criteria, including those listed above.[25-27] Face validity, "the degree to which the items indeed look as though they are an adequate reflection of the construct to be measured," is a sub-category of content validity.[26]

Criterion Validity. Criterion validity is the extent to which a measure agrees with a gold standard measure (the criterion). Predictive validity relies on criterion validity. For instance, in the question, "Does baseline self-rated health predict admission to a nursing home or mortality?"; mortality or nursing home admission is regarded as the criterion. In applications other than the assessment of predictive validity, the field of HRQL lacks gold standards and thus relies on the evaluation of construct validity.

Construct Validity. Construct validity is a measure's ability to perform as expected. It involves specifying *a priori* hypotheses about how the measure should perform based on an underlying model or conceptual framework, testing those hypotheses, and accumulating evidence over time and across settings. *Cross-sectional construct validity* involves making comparisons at a point in time. In *convergent validity* we expect a high correlation between two different measures of the same concept or measures of highly related domains such as mobility and self-care, or anxiety and depression. In *discriminant validity* we expect little or no correlation between measures of domains that are unrelated, such as vision and pain. Another strategy for assessing construct

validity is *known-groups comparisons*. For example, we would expect the scores for a measure of mobility to be systematically related to known groups based on the categories in the New York Heart Association functional classification system.[28]

Responsiveness (Longitudinal Construct Validity). Longitudinal construct validity measures within-person change over time. Does the measure capture meaningful change when it occurs? Change scores for those known to have changed (by some other criterion) should exceed change scores for those known not to have changed. For those who have changed, change scores should be systematically related to the degree of change. Measures for which there is substantial evidence of responsiveness in the relevant area enhance the confidence of the reviewer in the validity of the estimates of change.

Responsiveness is often assessed using effect size (ES, the magnitude of the change divided by the standard deviation of baseline scores), the standardized response mean (SRM, the magnitude of change divided by the standard deviation of change scores) or other related measures that are ratios of signal to noise.[29] Cohen provides a scheme to interpret the magnitude of ES: small (0.20); moderate (0.50); or large ($\geq$ 0.80) change.[30] A related measure, the standard error of measurement (SEM), is also frequently used. SEM is computed as the standard deviation at baseline times the square root of one minus test-retest reliability.[31]

The Distinction Between Predictive Validity and Responsiveness

Predictive validity refers to the ability of a baseline score to predict subsequent events. For instance, in both population health survey and clinical studies, self-rated health (SRH) (excellent, very good, good, fair, or poor), has been shown to predict mortality, admission to nursing homes, and other major health outcomes.[32-41] However, as there are only five options, the responsiveness of SRH is limited. Predictive validity does not necessarily imply that a measure will be able to detect within-person change over time. Further, predictive validity is based on the association between a baseline value and a subsequent outcome. In contrast, responsiveness instead focuses on the degree of change between the baseline and followup assessments.

How Should Reviewers Approach These Measurement Properties?

In assessing the measurement properties of functional status and HRQL measures there are a number of key questions.[42] How extensive is the evidence on the relevant measurement properties, especially responsiveness and interpretability, of the measures? How rigorous is that evidence? Is the evidence directly applicable to the issues at hand? Evidence on cross-sectional and longitudinal construct validity and interpretation is central to evaluating the effects of interventions. Construct validity involves the accumulation of evidence. The interpretation of that evidence also involves subjective judgments. If a systematic reviewer is confident that the measure is valid and responsive in the setting being reviewed, the reviewer can be more confident in the evidence on the effectiveness of an intervention. If the evidence on validity and responsiveness of the measure in that context is equivocal, interpreting results based on that measure will be challenging. The focus in this paper is on using evidence for making group-level comparisons rather than using evidence for making individual-patient-level decisions. The same methodological issues are relevant both for measures of functional status and HRQL.

Reviewing Measures of Health-Related Quality of Life: Special Considerations for Older Adults Populations

Many measures of functional status and HRQL were not designed specifically for use in older adult populations. A useful review both of generic measures that have been applied to older populations and older-population specific measures is provided in Haywood and colleagues.[43] Further, evidence on the construct validity and responsiveness of many measures is based on studies in populations whose mean age was 64-86, but age ranges vary by measure.[43,44] Extensive evidence on the reliability and validity of a measure does not necessarily imply that there is abundant evidence supporting its use among older adults, especially those at the upper extremes of the age ranges. Potential ceiling and floor effects, discussed below, are also very important in the context of studies of older adults.

To illustrate this we briefly review measurement properties for several widely used generic measures of HRQL: the Short-Form 36 (SF-36) and its preference-based version, the Short-Form 6D (SF- 6D or Six Dimensions), EuroQol-5D (EQ-5D), the Health Utilities Index Mark 3 (HUI3), and the Quality of Well-Being Scale (QWB). The SF-36 includes eight domains: physical functioning (PF), role-physical, bodily pain, general health, vitality, social functioning, role-emotional, and mental health.[8,45] The EQ-5D includes a five attribute health-status classification system: mobility, self-care, usual activity, pain/discomfort, and anxiety/depression, with three levels per attribute: no problem, some problem, or extreme problem.[21] The HUI3 system includes eight attributes: vision, hearing, speech, ambulation, dexterity, emotion, cognition, and pain and discomfort, with five or six levels per attribute, from severely impaired ("so unhappy that life is not worthwhile") to no problem or normal ("happy and interested in life").[46] The original version of the Quality of Well Being Scale (QWB) included three attributes (mobility, physical activity, and social activity) and a problem/symptom complex.[47] The more recent QWB-SA (self-administered) retains the same structure but includes fewer levels within each attribute and fewer problems/symptoms.[48]

How well do these measures work in older adults? In a prospective cohort study of patients 75+, Brazier and colleagues examined test-retest reliability in patients who self-identified as stable: patients who indicated that their health had not changed. Correlations for domains of the SF-36 ranged from 0.28 to 0.70; the correlation for EQ-5D scores was 0.67.[49] In a paper based on one of the original Medical Outcome Study (MOS) surveys (n = 3,445), one of the major studies upon which the SF-36 is based, McHorney and colleagues reported lower completion rates by item for those ≥ 75 than for the 65-74 group, who in turn had lower completion rates than persons <65. However, estimates of internal consistency reliability (Cronbach's alpha) did not vary by age, education, poverty status, diagnosis, or disease severity.[50]

A study of patients 65+ who identified themselves as stable reported intraclass correlation coefficients (ICCs) for SF-36 domains ranging from 0.65 to 0.87. Andresen and colleagues also showed evidence of cross-sectional construct validity for the SF-36 in that domain scores were lower for those who were older and for those with more severe comorbidities.[51]

Naglie and colleagues reported test-retest reliability estimates for patients with mild (mini-mental state examination [MMSE] scores 19-26) or moderate (MMSE 10-18) cognitive impairment, and proxy family caregivers for three generic preference-based measures, EQ-5D, HUI3, and the QWB[52]. Follow-up assessments were done approximately 2 weeks after the initial assessment. Examining consistency between initial and re-test responses by patients, the ICCs for the entire cohort were 0.79 (EQ-5D), 0.47 (HUI3), and 0.70 (QWB), respectively; for those with mild cognitive impairment the ICCs were 0.70, 0.75, and 0.81, respectively; for moderate

impairment 0.83, 0.25, and 0.59, respectively. Examining consistency between initial and re-test proxy responses, the ICCs were 0.71, 0.81, and 0.70, respectively. The results for HUI3 and the QWB were sensible; test-retest reliability for those with mild cognitive impairment was reasonable but persons with moderate cognitive impairment were not reliable respondents.[52] But for HUI3 and the QWB test-retest reliability was much lower for subjects with moderate cognitive impairment. This result has implications for the use of proxy respondents for subjects with moderate and severe cognitive impairment, a topic which is discussed below.

Two generalizations emerge from the studies reporting results for SF-36 and the Naglie and colleagues paper. First, the severely cognitively impaired are, in general, not capable of providing reliable and valid responses. Second, if the highly cognitively impaired are excluded, reliability in samples of older adults appear to be of the same order of magnitude as in general adult samples.

Floor and Ceiling Effects

If the range of function covered by a measure is less than the range experienced by patients, especially frail older adults, the measure may lack responsiveness. The potential for floor and ceiling effects is often assessed by examining response patterns. If there are spikes at the highest or lowest response option this is often interpreted as evidence of ceiling or floor effects, respectively. However, when using measures to assess the effectiveness of interventions prospective evidence of the performance of a measure is more important than whether or not there are spikes. Results from longitudinal studies indicate that the SF-36 (and therefore SF-6D) has well known floor effects that have been recognized in a wide variety of clinical settings and samples.[53-67] In a prospective cohort study comparing utility scores before and after elective total hip arthroplasty a gain of 0.10 was registered by SF-6D and a gain of 0.23 by HUI3.[55] (It should be noted that Version 2 of SF-36 is less prone to floor effects than Version 1. However, floor effects have been observed in studies using both versions.) In a natural history cohort of 124 patients recruited shortly after a stroke and followed for 6 months, the gain in overall HRQL observed in the 98 survivors (18 lost to followup) was 0.24 according to the EQ-5D[21] and 0.25 according to the HUI3,[46] but only 0.13 according to SF-6D.[68] Floor effects attenuated the ability of SF-36 and SF-6D to capture gains when many patients had moderate or severe burdens at baseline. The magnitude of improvement experienced by patients was underestimated because some patients were "worse off" than the measure could capture before the intervention; this underestimation could seriously bias estimates of the magnitude of change associated with interventions and cost-effectiveness estimates of those interventions.

Similarly, ceiling effects can threaten responsiveness. The absence of levels for mild problems in the EQ-5D probably accounts for the ceiling effects associated with the measure in population health survey and clinical applications. In a review of generic preference-based measures used in studies of patients with rheumatoid arthritis, ceiling effects associated with EQ-5D attenuated its responsiveness.[52,69-71] Similarly, a lack of responsiveness of EQ-5D has been reported in clinical studies of urinary incontinence in females [72] and treatments for leg ulcers.[73] The recently developed five-level EQ-5D may reduce ceiling effects.[74] Ceiling effects in population health surveys have also been observed for HUI2 and HUI3.[75,76]

Proxy Respondents

Cognitive impairment or physical disability may attenuate older adults' ability to respond, and this situation may be temporary or chronic. One approach to this problem is to rely on a

proxy respondent—a family member or caregiver who is familiar with the subject's current status. Agreement between self and proxy report then becomes an important issue; the "proxy as an agent" (if "X" could respond, what would she/he say) must be distinguished from the "proxy as an informed observer" (which of the following best describes the current condition of "X"). Most investigations of agreement have adopted the informed-observer approach.

Magaziner and colleagues examined agreement between self- and proxy-report in a prospective cohort of patients ≥ 65 (n = 361) being followed after hip fracture.[77] Both sets of respondents independently completed questionnaires on activities of daily living, instrumental activities of daily living, mental status, and depressive symptoms. Proxies tended to rate patients as more disabled than the patients rated themselves. Agreement was higher when the proxy and subject lived together; agreement was also higher when the proxy was a sibling or spouse as compared with offspring and nonrelative. Even mild cognitive impairment in the patient was associated with less agreement. Agreement was often lower on less observable aspects of physical and mental health.

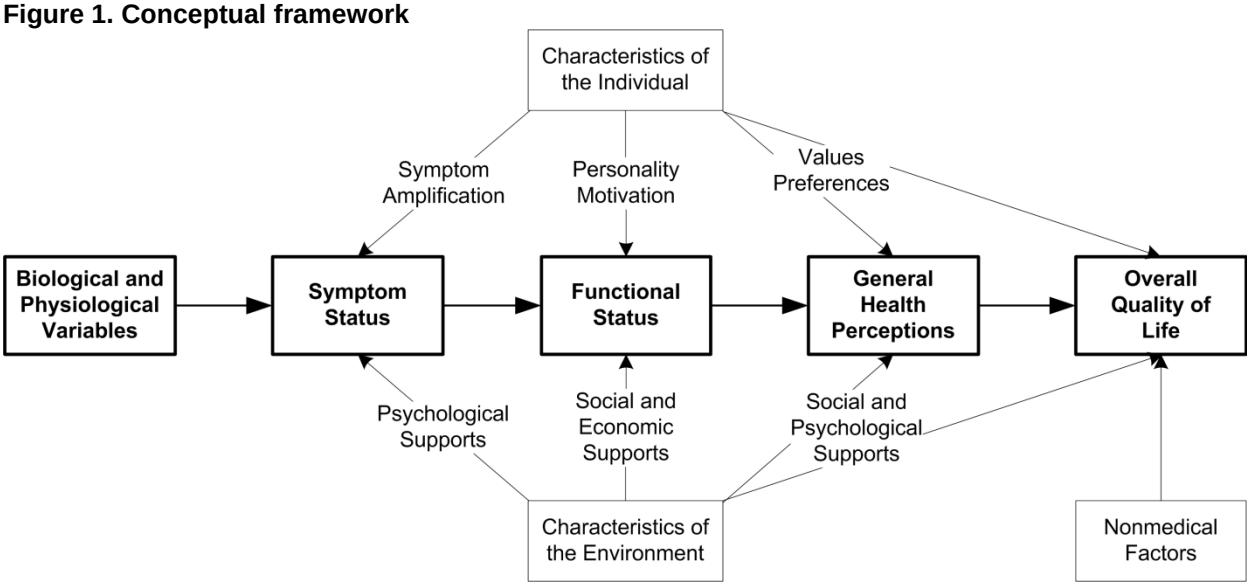
Clearly, in studies that gathered responses from both patients and their proxies, the responses were not interchangeable. The degree of agreement was affected by the observability of that aspect of health status, the degree of familiarity of the proxy with the current condition of the patient, and in some cases, the burden being experienced by the proxy caregiver.[78] The extent to which agreement varies with respect to these factors varies across studies. However, in general, these factors are associated with quantitatively important differences in the degree of agreement. Nonetheless, the results indicate a reasonable amount of agreement. Furthermore, evidence suggests that more reliable and valid information is available from proxy respondents who have frequent contact with patients who are becoming incapable of responding than is available from the patient directly. Whether differences in source of measure, patient versus proxy, impacts results in a systematic review could be evaluated through meta-regression or other techniques. Failure to obtain data from proxy respondents entails a substantial risk of overestimating the health of a cohort because the most severely affected will often not be able to respond.[79] Thus, in general, it is wise to collect both self and proxy assessments and to analyze them separately.

What To Assess and How To Interpret Measures of Functional Status and Health-Related Quality of Life

Conceptual Framework

An intervention is grounded in some conceptual framework about how the intervention would work and what dimensions of health status it would impact. The systematic reviewer must examine the conceptual framework of the original work to determine if the study included measures capable of capturing the intended effects. Wilson and Cleary[80] provide a useful framework that can guide the choice of measures and the interpretation and presentation of results (Figure 1).

Figure 1. Conceptual framework



Source: Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. JAMA. 1995;273:59-65. Reprinted with permission.

Another prominent framework is provided by the World Health Organization International Classification of Functioning, Disability, and Health (ICF).[81] The ICF focuses on the concepts of impairment, disability, participation, and the physical and social environment surrounding the subject. Like the Wilson and Cleary framework, the ICF can provide guidance to the systematic reviewer about what effects of an intervention and characteristics of the context in which the intervention took place to assess.

Minimum Important Difference

How to interpret the results of an intervention to prevent functional decline or improve HRQL is a key issue. Is the magnitude of change important? A statistically significant effect may not always translate to an important change from baseline. Thus, one must also consider the clinically important difference (CID), or minimum important difference (MID), defined as: "The smallest difference in score in the outcome of interest that informed patients or informed proxies perceive as important, either beneficial or harmful, and that would lead the patient or clinician to

consider a change in management".[82] Guyatt and colleagues suggest the use of the term "patient-important" rather than "clinically important" to focus on the "preeminence of the patient's values and preferences."[83,84]

MIDs are estimated using anchor-based or distribution-based approaches.[24,85-90] In the anchor-based approach, the change in HRQL score is related to a well-established meaningful measure. The anchor itself must be an independent measure and be readily interpretable—for example, the categories of the New York Heart Association functional classification system or ability to climb a flight of stairs. There must be an appreciable correlation when measured at the same time in the same person between the anchor and the target measure.[88] In contrast, the distribution-based approach is based on statistical criteria. It compares the magnitude of change compared to some measure of the variability of scores such as effect size (ES). Similar to evidence on construct validity, evidence on the usefulness of MIDs accumulates and evolves over time. If a guideline on the MID for a measure generates results that are congruent with clinical evidence and evidence from other measures, confidence in the usefulness of that MID increases over time. There is mixed evidence on the extent to which MIDs are context free.

Ware and Keller, using SF-36 data from the Medical Outcomes Study, provided examples of the usefulness of anchor-based approaches.[91] The PF scale ranges from 0 to 100. Thirty-two percent of respondents with a score of 40 can walk one block without limitations, a readily interpretable anchor; at a score of 50, 49.7 percent are able to walk a block. A change in PF score of 10 is clearly important. Yet a change of 10 in PF from 80 to 90 implies that 98.8 percent of respondents instead of 98.4 percent of respondents will be able to walk a block. Given that the standard deviation (SD) for the PF scale is 23.3,[45] a change of 10 is equivalent to an ES of 0.43, moderate in the scheme proposed by Cohen.[30] In this example, the anchor-based interpretation is meaningful while the distribution-based interpretation (ES) has the potential to be misleading.

A number of studies report results in which anchor-based and distribution-based approaches provide similar estimates of the threshold for a patient-important difference.[89] If a systematic reviewer cannot find evidence in support of an anchor-based criterion for a measure, one default option is to use 1.0 standard error of measurement or 0.5 SD. Nonetheless, anchor-based MIDs that focus on the importance of the magnitude of the change provide a conceptually preferable guide to interpreting results than do distribution-based criteria.[92,93]

Absolute or Relative Change?

An advantage of ES and SRM is that they can be used to make comparisons across studies and among measures. However, within a study the stimulus (intervention) is the same so one can compare the absolute magnitude of change among measures that use a common scale such as the conventional scale for preference-based (utility) measures in which 0.00 = dead and perfect health = 1.00. In general, clinicians and systematic reviewers are more interested in the absolute magnitude of change than in the relative magnitude of change expressed in SD units (ES or SRM).

One level of interpretation is to consider if the mean magnitude of change observed in a RCT is patient- or clinically important. More relevant is the proportion of patients achieving no, small, moderate, or large change.[88] As Guyatt and colleagues note, mean change can be misleading if, for instance, there is heterogeneity in treatment effect such that an important minority of patients improved moderately while the majority experienced no change.[94] The proportion that benefit can then be used to calculate the number needed to treat, the inverse of the absolute risk reduction, which has intuitive appeal to clinicians. Johnston and colleagues discuss an algorithm

that allows systematic reviewers to make comparisons across studies using MIDs.[95] A related approach focuses on individual-level data and the classification of patients as responders if the change that individual experienced is greater than or equal to a threshold—in a sense, a criterion for a patient-important difference defined at the individual level.[2,96]

Observational Data

Although the advantages of using evidence derived large randomized controlled clinical trials are well known, relevant trial data is often sparse or not available to systematic reviewers. Sometimes a reviewer must rely on observational data on the effects of an intervention. It may therefore be useful to compare the trajectory observed in the study to trajectories observed in natural history cohort studies of older adults. Evidence from such cohort studies might also be useful for interpreting results from controlled clinical trials.[97] When serving a high-risk population, slowing the rate of decline in functional status or HRQL may be a realistic goal and the maintenance of stability (as opposed to improvement) may be a marker of success.[98] Evidence from a natural history cohort can serve as a useful comparator for the results of an intervention tested in an observational study. Such evidence provides an answer to a counterfactual question: what would have happened in the absence of an intervention? Two examples are the longitudinal National Population Survey (NPHS) and Beaver Dam study.

The Statistics Canada NPHS displayed 10-year trends in overall HRQL for a cohort of respondents 40 years and older living in the community at baseline in 1994/95. The rate of decline in HRQL (measured by HUI3) accelerated in respondents aged in their mid-70s. The rate of decline is higher when those who were institutionalized and those who died during the followup period were included in the analyses.[99] Data from the NPHS could serve as a benchmark for comparisons.

Another example of a longitudinal natural history cohort is the Beaver Dam study. Begun in 1987-1988 in Beaver Dam, Wisconsin, a cohort of 4,926 respondents 43-84 years old was enrolled.[100-102] Respondents have been followed since, most recently surveyed at 15 years of followup (from 2003 to 2005).[103] HRQL measures used in the Beaver Dam study included the SF-36, the QWB,[104] and the time-trade off.[20] The Beaver Dam study provides a rich source of natural history data, and although participants reported a wide range of income levels, its mainly white population (99 percent) may not generalize to the entire U.S.

Implications for Researchers

Assessing Health-Related Quality of Life

As Feinstein suggested, "assessments of health status are important because improvements in symptoms, other clinical problems, and functional capacity are usually the main goals of patients in seeking clinical care".[105] Similarly, Osoba and King argue that "the ultimate goal of health care is to restore or preserve functioning and well-being related to health, that is health-related quality of life."[106] These ideas are underscored in the Public Comment Draft Report of the Patient-Centered Outcomes Research Institute Methodology Committee presented on July 23, 2012 (www.pcori.org/2012/methodology-report/).

Studies investigating ways to improve or maintain functional status in older adults need to include the assessment of HRQL. But using which measures? Generic measures provide the basis for broad comparisons, the ability to reflect comorbidities, and the ability to detect side effects and other consequences. More targeted measures often focus on the most salient domains and are often more responsive than generic measures.

One criterion in guiding the choice of and mix of type of measures should be the availability of evidence on the reliability, construct validity, and responsiveness of the measure in the context in which it will be applied. Sometimes there is a tension between choosing measures with well documented measurement properties in that application and choosing widely-used measures that permit comparison to other studies.

There is, however, a risk of "premature" standardization.[107,108] For instance, if a widely-used generic measure is chosen to enhance the ability to make comparisons to other studies, but that generic measure has inferior measurement properties relative to some other generic measures in the relevant area of application, then neither internal validity nor external generalizability are well served. Measures need to be chosen on the basis of relevance and their track record in the context of the study at hand. Further, as the examples presented in the paper illustrate, in general scores and change scores among generic measures are not interchangeable.

Studies of older adults must attend to these multiple challenges. Inclusion and exclusions criteria need to match the level of vulnerability of study participants appropriate for an intervention. Interventions need to be systematized and reproducible. Control and intervention groups must reflect the variability of health trajectories in older people (no easy task, as this is infrequently known at the start of a trial). Multiple inter-related outcomes need to be considered. Measures need to be appropriate for the baseline population to avoid floor and ceiling effects. Comparability across studies would be enhanced if researchers were to agree on a small number of measures that are appropriate for older populations. Examples of such efforts include Core Outcome Measures for Effectiveness Trials initiative (www.comet-initiative.org) and Outcome Measures in Rheumatology Trials (www.omeract.org). MIDs should be determined beforehand in a study.

Systematically adding measures of HRQL and functional status to studies of older adults and the routine use of these measures in chronic care management would importantly add to the evidence available.[84,109,110] In particular, routine collection would provide evidence on persons seldom included in clinical trials, such as patients with multiple chronic conditions, concomitant medications, and older adults.[111,112] The use of health profile measures and measures of symptom or function from the Patient Reported Outcomes Measurement Information System (PROMIS) has the potential to enhance the ability to make focused comparisons across populations and studies (www.nihpromis.org).[113-115] Of course, substantial evidence on the measurement

properties of PROMIS measures in older adult population will be required before those measures can be recommended. The usefulness of adding HRQL assessments to studies and registries would be enhanced by adherence to reporting standards for HRQL evidence.[116-120]

An additional very important criterion in selecting measures for use in older adult populations is evidence of their acceptability to respondents, family members, and clinicians.[15] In particular Haywood and colleagues report that measures acceptable to general adult populations are sometimes not acceptable to older respondents.[43]

Implications for Systematic Reviewers

Systematic reviewers should review the methods of the empirical studies assessing HRQL and functional status in light of the normative study design criteria outlined above in Section 4. But what if not all the criteria are met? For instance, if the existing evidence is based solely or almost exclusively on the basis of condition-specific measures of HRQL, the risk of a false negative result on the effectiveness of the intervention may be lower (specific measures are often more responsive) but the risk of a false positive result (concluding that the net benefits of the intervention are positive) may be higher because of the attenuated ability to detect side effects of treatment, effects which might offset some or even all of the treatment effects. Further, the ability to make broad comparisons will be attenuated because no generic measure was used in the underlying studies. Alternatively, if the underlying research is based mainly on results from generic measures, the risk of a false negative may be higher (generic measures are often less responsive than specific ones) while the risk of a false positive may be lower due to the ability to detect important side effects and the ability to make broad comparisons will be enhanced. Of course, if the generic measure was not carefully selected, floor and/or ceiling effects may attenuate the advantages of generic measures.

If a natural history cohort study that matches the characteristics (or which has a subset of participants who match) of the one being studied in the systematic review is available and that study included suitable measures of HRQL and functional status, then evidence from the observational study can help interpret the results of the intervention study being reviewed. If the match is less than perfect, the systematic reviewer will have to compromise.

Appendix Table 1 provides a brief summary of the relevant measurement properties for a number of widely-used generic measures and a few of the disease-specific measures chosen to illustrate the issues covered in this paper. This is intended to be illustrative, rather than a comprehensive review. Measures based on each of the three major paradigms of HRQL are included in Appendix A, Table 1, the psychometric paradigm (SF-36), the clinimetric paradigm (Chronic Respiratory Questionnaire), and the preference-based/economics/decision science paradigm (HUI3), along with two widely used measures of functional status (activities of daily living; instrumental activities of daily living).

Summary and Conclusions

Several conclusions emerge. First, randomized trials must include the right patients—those who have enough impairment to make intervention worthwhile, but who are not so ill that an intervention would be unlikely to improve their situation or slow their rate of decline. Second, current HRQL and functional status measures are not always responsive to subtle but important changes in older populations. Third, the older population has substantial heterogeneity in disease progression. Fourth, the natural history of disease in older adults is highly variable. An intervention might slow functional decline, but that can be difficult to demonstrate but nonetheless be important.

As the field of geriatrics embraces these and other recommendations to strengthen the evidence base for evaluating interventions that can prevent functional decline in older adults, systematic reviewers will be able to apply a more rigorous set of criteria that will allow for stronger evidence to guide patient care. Systematic reviewers can employ our framework to ensure that all the challenges inherent in interpreting the literature for this growing population are considered.

Table 1. Brief definitions of important measurement properties

Term	Definition
Reliability	A reliable measure is consistent and reproducible[2]
Internal Consistency	The extent to which items are measuring the same concept.[121]
Intra- and Inter-Observer Reliability	The extent of agreement across assessments or among individuals.[122]
Validity	The measure accurately reflects the concept it is intended to measure.[121]
Content Validity	The extent to which the measure covers the full range of meanings included in the concept.[121]
Criterion Validity	The extent of agreement between the measure and a gold standard measure of the same concept.[2]
Construct Validity	Evidence that the relationships among items and domains conform to a priori hypotheses and that logical relationships exist between the measure and characteristics of patients and patient groups.[2]
Convergent Validity	Convergent validity refers to evidence of a moderate or strong relationship between measures of the same concept or construct.[122]
Discriminant Validity	Discriminant validity refers to evidence of the lack of relationship between measures of a different concept or construct.[13,121]
Cross-Sectional Construct Validity	Evidence of construct validity based on comparisons at a point in time.
Responsiveness (Longitudinal Construct Validity).	The ability of a measure to capture meaningful change when it occurs.[121]
Interpretation	The ability to attach meaning to the scores provided by a measure.[123]

Note: The brief definitions provided are not meant to be definitive; each of the concepts is expanded upon in the text. The sources from which the definitions are paraphrased are cited.

References

1. Tinetti ME FTBC. Designing health care for the most common chronic condition–multimorbidity. JAMA 2012 Jun 20;307(23):2493-4. PMID: doi: 10.1001/jama.2012.5265.

2. U.S.Department of Health and Human Services, FDA, CDER, et al. Guidance for Industry: Patient-Reported Outcome Measures: Use in Medical Product Development to Support Label Claims. Rockville, MD: 2009.

3. Patrick DL, Erickson P l.s. Health status and health policy: Quality of life in health care evaluation and resource allocation. New York: Oxford University Press; 1993.

4. Starfield B. Basic concepts in population health and health care. J Epidemiol Community Health 2001 Jul;55(7):452-4.

5. Fried LP, Tangen CM, Walston J, et al. Frailty in older adults: evidence for a phenotype. J Gerontol A Biol Sci Med Sci 2001 Mar;56(3):M146-M156.

6. Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. Ann Intern Med 1993 Apr 15;118(8):622-9.

7. Garratt A, Schmidt L, Mackintosh A, et al. Quality of life measurement: bibliographic study of patient assessed health outcome measures. BMJ 2002 Jun 15;324(7351):1417.

8. Ware JE, Jr., Sherbourne CD. The MOS 36-item short-form health survey (SF-36). 1. Conceptual framework and item selection. Med Care 1992 Jun;30(6):473-83.

9. Horsman J, Furlong W, Feeny D, et al. The Health Utilities Index (HUI®): concepts, measurement properties and applications. Health Qual Life Outcomes 2003 Oct 16;1(1):54.

10. Guyatt GH, King DR, Feeny DH, et al. Generic and specific measurement of health-related quality of life in a clinical trial of respiratory rehabilitation. J Clin Epidemiol 1999 Mar;52(3):187-92.

11. Wiebe S, Guyatt G, Weaver B, et al. Comparative responsiveness of generic and specific quality-of-life instruments. J Clin Epidemiol 2003 Jan;56(1):52-60.

12. Longworth L, Yang Y, Brazier JE, et al. Valuing a vision 'bolt-on' item for the EQ-5D. Quality of Life Research 2013 Oct 1;21(Supplement 1):53-4.

13. Fayers PM, Machin D, , s. Quality of life. The assessment, analysis and interpretation of patient-reported outcomes. West Sussex, England: John Wiley & Sons Ltd; 2007.

14. Beaton DE, Bombardier C, Katz JN, et al. A taxonomy for responsiveness. J Clin Epidemiol 2001 Dec;54(12):1204-17.

15. Feeny DH, Furlong W, Mulhern RK, et al. A framework for assessing health-related quality of life among children with cancer. International Journal of Cancer Supplement 1999;12:2-9.

16. Fayers PM, Hand DJ. Factor analysis, causal indicators and quality of life. Quality of Life Research 1997 Mar;6(2):139-50.

17. Eckstrom E, Feeny DH, Walter LC, et al. Individualizing cancer screening in older adults: a narrative review and framework for future research. J Gen Intern Med 2013 Feb;28(2):292-8.

18. Feeny D. Health-related quality-of-life data should be regarded as a vital sign. J Clin Epidemiol 2013 Jul;66(7):706-9.

19. Torrance GW. Measurement of health state utilities for economic appraisal. J Health Econ 1986 Mar;5(1):1-30.

20. Torrance GW, Thomas WH, Sackett DL. A utility maximization model for evaluation of health care programs. Health Serv Res 1972;7(2):118-33.

21. Rabin R, de Charro F. EQ-5D: A measure of health status from the EuroQol Group. Ann Med 2001 Jul;33(5):337-43.

22. Feeny DH. Preference-based measures: Utility and quality-adjusted life years. In: Fayers P, Hays R, eds. Assessing Quality of Life in Clinical Trials. Second ed. Oxford: Oxford University Press; 2005. p. 405-29.

23. Feeny DH. The roles for preference-based measures in support of cancer research and policy. In: Lipscomb J, Gotay CC, Snyder C, eds. Assessment in Cancer: Measures, Methods, and Applications. New York: Cambridge University Press; 2005. p. 69-92.

24. Revicki DA, Osoba D, Fairclough D, et al. Recommendations on health-related quality of life research to support labeling and promotional claims in the United States. Qual Life Res 2000;9(8):887-900.

25. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res 2010 May;19(4):539-49.

26. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol 2010 Jul;63(7):737-45.

27. Mokkink LB, Terwee CB, Knol DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Med Res Methodol 2010 Mar 18;10:22.

28. The Criteria Committee of the New York Heart Association. Diseases of the Heart and Blood Vessels: Nomenclature and Criteria for Diagnosis. Boston, Mass: Little Brown; 1964.

29. Terwee CB, Dekker FW, Wiersinga WM, et al. On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. Qual Life Res 2003 Jun;12(4):349-62.

30. Cohen J. Statistical power analysis for the behavioral sciences. Hillsdale, New Jersey: Lawrence Erlbaum Associates; 1988.

31. Norman GR, Wyrwich KW, Patrick DL. The mathematical relationship among different forms of responsiveness coefficients. Qual Life Res 2007 Jun;16(5):815-22.

32. Idler EL, Kasl SV, Lemke JH. Self-evaluated health and mortality among the elderly in New Haven, Connecticut, and Iowa and Washington counties, Iowa, 1982-1986. Am J Epidemiol 1990 Jan;131(1):91-103.

33. Idler EL, Kasl SV. Self-ratings of health: do they also predict change in functional ability? J Gerontol B Psychol Sci Soc Sci 1995 Nov;50(6):S344-53.

34. Idler EL, Russell LB, Davis D. Survival, functional limitations, and self-rated health in the NHANES I Epidemiologic Follow-up Study, 1992. First National Health and Nutrition Examination Survey. Am J Epidemiol 2000 Nov 1;152(9):874-83.

35. Idler EL, Benyamini Y. Self-rated health and mortality: A review of twenty-seven community studies. J Health Soc Behav 1997;38(1):21-37.

36. Benjamins MR, Hummer RA, Eberstein IW, et al. Self-reported health and adult mortality risk: an analysis of cause-specific mortality. Soc Sci Med 2004 Sep;59(6):1297-306.

37. Kaplan MS, Berthelot JM, Feeny DH, et al. The predictive validity of two measures of health-related quality of life: mortality in a longitudinal population-based study. Qual Life Res 2007 Nov;16(9):1539-46.

38. Diehr P, Williamson J, Burke GL, et al. The aging and dying processes and the health of older adults. J Clin Epidemiol 2002 Mar;55(3):269-78.

39. Diehr P, Williamson J, Patrick DL, et al. Patterns of self-rated health in older adults before and after sentinel health events. J Am Geriatr Soc 2001 Jan;49(1):36-44.

40. Bjorner JB, Fayers P, Idler EL. Self-rated health. In: Fayers P, Hays R, eds. Assessing quality of life in clinical trials: Methods and Practice. 2nd ed. New York: Oxford University Press; 2005. p. 309-23.

41. DeSalvo KB, Bloser N, Reynolds K, et al. Mortality prediction with a single general self-rated health question. a meta-analysis. J Gen Intern Med 2006 Mar;21(3):267-75.

42. Terwee CB, Mokkink LB, Knol DL, et al. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res 2012 May;20(4):651-7. PMID: 10.1007/s11136-011-9960-1 [doi].

43. Haywood K, Garratt AM, Schmidt L, et al. Health status and quality of life in older people. A structured review of patient-reported health instruments. Report to the Department of Health April 2004. National Centre for Health Outcomes Development; 2004.

44. Haywood KL, Garratt AM, Fitzpatrick R. Older people specific health status and quality of life: a structured review of self-assessed instruments. J Eval Clin Pract 2005 Aug;11(4):315-27.

45. Ware JE. The SF-36 Health Survey. Quality of Life and Pharmacoeconomics in Clinical Trials. Second ed. Philadelphia: Lippincott-Raven; 1996. p. 337-45.

46. Feeny DH, Furlong W, Torrance GW, et al. Multi-attribute and single-attribute utility functions for the health utilities index mark 3 system. Med Care 2002 Feb;40(2):113-28.

47. Kaplan RM, Bush JW, Berry CC. Health status: types of validity and the index of well-being. Health Serv Res 1976;11(4):478-507.

48. Kaplan RM, Anderson JP. The general health policy model: an integrated approach. In: Spilker B, ed. Quality of Life and Pharmacoeconomics in Clinical Trials. Second ed. Philadelphia: Lippincott-Raven Press; 1996. p. 309-22.

49. Brazier JE, Walters SJ, Nicholl JP, et al. Using the SF-36 and Euroqol on an elderly population. Qual Life Res 1996 Apr;5(2):195-204.

50. McHorney CA, Ware JE, Jr., Lu JF, et al. The MOS 36-item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. Med Care 1994 Jan;32(1):40-66.

51. Andresen EM, Bowley N, Rothenberg BM, et al. Test-retest performance of a mailed version of the Medical Outcomes Study 36-Item Short-Form Health Survey among older adults. Med Care 1996 Dec;34(12):1165-70.

52. Naglie G, Tomlinson G, Tansey C, et al. Utility-based quality of life measures in Alzheimer's disease. Qual Life Res 2006 May 15;15(4):631-43.

53. Bosch JL, Hunink MG. Comparison of the Health Utilities Index Mark 3 (HUI3) and the EuroQol EQ-5D in patients treated for intermittent claudication. Qual Life Res 2000;9(6):591-601.

54. Brazier J, Roberts J, Tsuchiya A, et al. A comparison of the EQ-5D and SF-6D across seven patient groups. Health Econ 2004 Sep;13(9):873-84. PMID: 10.1002/hec.866 [doi].

55. Feeny DH, Wu L, Eng K. Comparing short form 6D, standard gamble, and Health Utilities Index Mark 2 and Mark 3 utility scores: results from total hip arthroplasty patients. Qual Life Res 2004 Dec;13(10):1659-70.

56. Grieve R, Grishchenko M, Cairns J. SF-6D versus EQ-5D: reasons for differences in utility scores and impact on reported cost-utility. Eur J Health Econ 2008 Mar 9;10(1):15-23.

57. Guo N, Marra CA, Marra F, et al. Health state utilities in latent and active tuberculosis. Value in Health 2008 Dec;11(7):1154-61.

58. Harrison MJ, Davies LM, Bansback NJ, et al. The comparative responsiveness of the EQ-5D and SF-6D to change in patients with inflammatory arthritis. Qual Life Res 2009 Nov;18(9):1195-205.

59. Hatoum HT, Brazier JE, Akhras KS. Comparison of the HUI3 with the SF-36 preference based SF-6D in a clinical trial setting. Value in Health 2004 Sep;7(5):602-9.

60. Kopec JA, Willison KD. A comparative review of four preference-weighted measures of health-related quality of life. J Clin Epidemiol 2003 Apr;56(4):317-25. PMID: S0895435602006091 [pii].

61. Longworth L, Bryan S. An empirical comparison of EQ-5D and SF-6D in liver transplant patients. Health Econ 2003 Dec;12(12):1061-7.

62. O'Brien BJ, Spath M, Blackhouse G, et al. A view from the bridge: agreement between the SF-6D utility algorithm and the Health Utilities Index. Health Econ 2003 Nov;12(11):975-81.

63. Pickard AS, Johnson JA, Feeny DH. Responsiveness of generic health-related quality of life measures in stroke. Qual Life Res 2005 Feb;14(1):207-19.

64. Suarez-Almazor ME, Kendall C, Johnson JA, et al. Use of health status measures in patients with low back pain in clinical settings. Comparison of specific, generic and preference-based instruments. Rheumatology (Oxford) 2000 Jul;39(7):783-90.

65. Szende A, Svensson K, Stahl E, et al. Psychometric and utility-based measures of health status of asthmatic patients with different disease control level. Pharmacoeconomics 2004;22(8):537-47. PMID: 2285 [pii].

66. Szende A, Leidy NK, Stahl E, et al. Estimating health utilities in patients with asthma and COPD: evidence on the performance of EQ-5D and SF-6D. Qual Life Res 2009 Mar;18(2):267-72. PMID: 10.1007/s11136-008-9429-z [doi].

67. Taylor SJ, Taylor AE, Foy MA, et al. Responsiveness of common outcome measures for patients with low back pain. Spine 1999 Sep 1;24(17):1805-12.

68. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ 2002 Mar;21(2):271-92.

69. Luo N, Chew LH, Fong KY, et al. A comparison of the EuroQol-5D and the Health Utilities Index mark 3 in patients with rheumatic disease. J Rheumatol 2003 Oct;30(10):2268-74.

70. Klassen A, Fitzpatrick R, Jenkinson C, et al. Contrasting evidence of the effectiveness of cosmetic surgery from two health related quality of life measures. J Epidemiol Community Health 1999 Jul;53(7):440-1.

71. Dobscha SK, Corson K, Perrin NA, et al. Collaborative care for chronic pain in primary care: a cluster randomized trial. JAMA 2009 Mar 25;301(12):1242-52.

72. Haywood KL, Garratt AM, Lall R, et al. EuroQol EQ-5D and condition-specific measures of health outcome in women with urinary incontinence: reliability, validity and responsiveness. Qual Life Res 2008 Apr;17(3):475-83. PMID: 10.1007/s11136-008-9311-z [doi].

73. Walters SJ, Morrell CJ, Dixon S. Measuring health-related quality of life in patients with venous leg ulcers. Qual Life Res 1999 Jun;8(4):327-36.

74. Pickard AS, De Leon MC, Kohlmann T, et al. Psychometric comparison of the standard EQ-5D to a 5 level version in cancer patients. Med Care 2007 Mar;45(3):259-63.

75. Fryback DG, Dunham NC, Palta M, et al. US norms for six generic health-related quality-of-life indexes from the National Health Measurement study. Med Care 2007 Dec;45(12):1162-70.

76. Cherepanov D, Palta M, Fryback DG. Underlying Dimensions of the Five Health-Related Quality-of-Life Measures Used in Utility Assessment: Evidence From the National Health Measurement Study. Med Care 2010 Aug;48(8):718-25.

77. Magaziner J, Simonsick EM, Kashner TM, et al. Patient-proxy response comparability on measures of patient health and functional status. J Clin Epidemiol 1988;41(11):1065-74.

78. Sneeuw KC, Aaronson NK, Sprangers MA, et al. Evaluating the quality of life of cancer patients: assessments by patients, significant others, physicians and nurses. Br J Cancer 1999 Sep;81(1):87-94.

79. Jones CA, Feeny DH. Agreement between patient and proxy responses of health-related quality of life after hip fracture. J Am Geriatr Soc 2005 Jul;53(7):1227-33.

80. Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. JAMA 1995 Jan 4;273(1):59-65.

81. World Health Organization. International Classification of Functioning, Disability and Health. Geneva: World Health Organization; 2001. Accessed June 21, 2013.

82. Schunemann HJ, Puhan M, Goldstein R, et al. Measurement properties and interpretability of the Chronic respiratory disease questionnaire (CRQ). COPD 2005 Mar;2(1):81-9.

83. Guyatt G, Montori V, Devereaux PJ, et al. Patients at the center: in our practice, and in our use of language. ACP J Club 2004 Jan;140(1):A11-A12. PMID: ACPJC-2004-140-1-A11 [pii].

84. Schunemann HJ, Guyatt GH. Commentary--goodbye M(C)ID! Hello MID, where do you come from? Health Serv Res 2005 Apr;40(2):593-7. PMID: HESR374 [pii];10.1111/j.1475-6773.2005.00374.x [doi].

85. Revicki D, Hays RD, Cella D, et al. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol 2008 Feb;61(2):102-9.

86. Hays RD, Farivar SS, Liu H. Approaches and recommendations for estimating minimally important differences for health-related quality of life measures. COPD 2005 Mar;2(1):63-7.

87. Leidy NK, Wyrwich KW. Bridging the gap: using triangulation methodology to estimate minimal clinically important differences (MCIDs). COPD 2005 Mar;2(1):157-65.

88. Guyatt GH, Osoba D, Wu AW, et al. Methods to explain the clinical significance of health status measures. Mayo Clin Proc 2002 Apr;77(4):371-83.

89. Wyrwich KW, Bullinger M, Aaronson N, et al. Estimating clinically significant differences in quality of life outcomes. Qual Life Res 2005 Mar;14(2):285-95.

90. Wyrwich KW, Norquist JM, Lenderking WR, et al. Methods for interpreting change over time in patient-reported outcome measures. Qual Life Res 2013 Apr;22(3):475-83.

91. Ware JE, Keller SD. Interpreting general health measures. In: Spilker BF (eds). Quality of Life and Pharmacoeconomics in Clinical Trials. Second ed. Philadelphia: Lippincott-Raven; 1996. p. 445-60.

92. de Vet HC, Terwee CB, Ostelo RW, et al. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. Health Qual Life Outcomes 2006 Aug 22;4:54. PMID: 1477-7525-4-54 [pii];10.1186/1477-7525-4-54 [doi].

93. Turner D, Schunemann HJ, Griffith LE, et al. The minimal detectable change cannot reliably replace the minimal important difference. J Clin Epidemiol 2010 Jul;63(1):28-36.

94. Guyatt GH, Juniper EF, Walter SD, et al. Interpreting treatment effects in randomised trials. BMJ 1998 Feb 28;316(7132):690-3.

95. Johnston BC, Thorlund K, Schunemann HJ, et al. Improving the interpretation of quality of life evidence in meta-analyses: the application of minimal important difference units. Health Qual Life Outcomes 2010;8:116.

96. McLeod LD, Coon CD, Martin SA, et al. Interpreting patient-reported outcome results: US FDA guidance and emerging methods. Expert Rev Pharmacoecon Outcomes Res 2011 Apr;11(2):163-9.

97. Norris SL, High K, Gill TM, et al. Health care for older Americans with multiple chronic conditions: a research agenda. J Am Geriatr Soc 2008 Jan;56(1):149-59.

98. Klein S, McCarthy D. CareOregon: Transforming the role of a Medicaid Health Plan from payer to partner. 50. 2010.

99. Orpana HM, Ross N, Feeny D, et al. The natural history of health-related quality of life: a 10-year cohort study. Health Reports 2009 Mar;20(1):1-7.

100. Klein R, Klein BE, Linton KL, et al. The Beaver Dam Eye Study: visual acuity. Ophthalmology 1991 Aug;98(8):1310-5.

101. Klein R, Klein BE, Moss SE. Age-related eye disease and survival. The Beaver Dam Eye Study. Arch Ophthalmol 1995 Mar;113(3):333-9.

102. Fryback DG, Dasbach EJ, Klein R, et al. The Beaver Dam Health Outcomes Study: initial catalog of health-state quality factors. Med Decis Making 1993 Apr;13(2):89-102.

103. Klein R, Klein BE, Lee KE, et al. Changes in visual acuity in a population over a 15-year period: The Beaver Dam Eye Study. Am J Ophthalmol 2006 Oct;142(4):539-49.

104. Patrick DL, Bush JW, Chen MM. Methods for measuring levels of well-being for a health status index. Health Serv Res 1973;8(3):228-45.

105. Feinstein AR. Benefits and obstacles for development of health status assessment measures in clinical settings. Med Care 1992 May;30(5 Suppl):MS50-MS56.

106. Osoba D, King M. Meaningful differences. In: Fayers P, Hays R, eds. Assessing Quality of Life in Clinical Trials. Second ed. Oxford: Oxford University Press; 2005. p. 243-57.

107. Feeny D, Spritzer K, Hays RD, et al. Agreement about Identifying Patients Who Change over Time: Cautionary Results in Cataract and Heart Failure Patients. Med Decis Making 2012;32(2):273-86. PMID: 0272989X11418671 [pii];10.1177/0272989X11418671 [doi].

108. Feeny D. Standardization and regulatory guidelines may inhibit science and reduce the usefulness of analyses based on the application of preference-based measures for policy decisions. Med Decis Making 2013 Apr;33(3):316-9.

109. Velikova G, Booth L, Smith AB, et al. Measuring quality of life in routine oncology practice improves communication and patient well-being: a randomized controlled trial. J Clin Oncol 2004 Feb 15;22(4):714-24. PMID: 10.1200/JCO.2004.06.078 [doi];JCO.2004.06.078 [pii].

110. Santana MJ, Feeny D, Johnson JA, et al. Assessing the use of health-related quality of life measures in the routine clinical care of lung-transplant patients. Qual Life Res 2010 Apr;19(3):371-9.

111. Abernethy AP, Etheredge LM, Ganz PA, et al. Rapid-learning system for cancer care. J Clin Oncol 2010 Sep 20;28(27):4268-74.

112. Tinetti ME, Studenski SA. Comparative Effectiveness Research and Patients with Multiple Chronic Conditions. N Engl J Med 2011 Jun 22;364(26):2478-81. PMID: 10.1056/NEJMp1100535 [doi].

113. Cella D, Gershon R, Lai JS, et al. The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment. Qual Life Res 2007;16(Suppl 1):133-41.

114. Cella D, Yount S, Rothrock N, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. Med Care 2007 May;45(5 Suppl 1):S3-S11.

115. Fries JF, Bruce B, Cella D. The promise of PROMIS: using item response theory to improve assessment of patient-reported outcomes. Clin Exp Rheumatol 2005 Sep;23(5 Suppl 39):S53-S57.

116. Staquet M, Berzon R, Osoba D, et al. Guidelines for reporting results of quality of life assessments in clinical trials. Qual Life Res 1996 Oct;5(5):496-502.

117. Lee CW, Chi KN. The standard of reporting of health-related quality of life in clinical cancer trials. J Clin Epidemiol 2000 May;53(5):451-8.

118. Brundage M, Bass B, Davidson J, et al. Patterns of reporting health-related quality of life outcomes in randomized clinical trials: implications for clinicians and quality of life researchers. Qual Life Res 2010 Nov 26;20(5):653-64.

119. Brundage M, Bass B, Jolie R, et al. A knowledge translation challenge: clinical use of quality of life data from cancer clinical trials. Qual Life Res 2011 Jan 29;20(7):979-85.

120. Calvert M, Blazeby J, Altman DG, et al. Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension. JAMA 2013 Feb 27;309(8):814-22.

121. Guyatt G, Feeny D, Patrick D. Glossary. Control Clin Trials 1991 Aug;12(4 Suppl 1):274S-80S.

122. Osoba D, ed. The effect of cancer on quality of life. Boca Raton, FL: CRC Press Inc.; 1991.

123. Glossary: Health Outcomes Methodology. Medical Care 2000;38(9 Suppl):II7-II13.

124. McDowell I, . Measuring health: a guide to rating scales and questionnaires. New York, NY: Oxford University Press; 2006.

125. Sikkes SA, de Lange-de Klerk ES, Pijnenburg YA, et al. A systematic review of Instrumental Activities of Daily Living scales in dementia: room for improvement. J Neurol Neurosurg Psychiatry 2009 Jan;80(1):7-12.

126. Guyatt GH, Berman LB, Townsend M, et al. A measure of quality of life for clinical trials in chronic lung disease. Thorax 1987 Oct;42(10):773-8.

127. Haywood, K, Garratt, AM, Schmidt, L, et al. Health status and quality of life in older people. A structured review of patient-reported health instruments. National Centre for Health Outcomes Development; 2004.

128. Laake K, Laake P, Ranhoff AH, et al. The Barthel ADL index: factor structure depends upon the category of patient. Age Ageing 1995 Sep;24(5):393-7.

129. de Morton NA, Keating JL, Davidson M. Rasch analysis of the barthel index in the assessment of hospitalized older patients after admission for an acute medical condition. Arch Phys Med Rehabil 2008 Apr;89(4):641-7.

130. Hsueh IP, Lin JH, Jeng JS, et al. Comparison of the psychometric characteristics of the functional independence measure, 5 item Barthel index, and 10 item Barthel index in patients with stroke. J Neurol Neurosurg Psychiatry 2002 Aug;73(2):188-90.

131. Hajiro T, Nishimura K, Tsukino M, et al. Comparison of discriminative properties among disease-specific questionnaires for measuring health-related quality of life in patients with chronic obstructive pulmonary disease. Am J Respir Crit Care Med 1998 Mar;157(3 Pt 1):785-90.

132. Lacasse Y, Wong E, Guyatt G. A systematic overview of the measurement properties of the Chronic Respiratory Questionnaire. Canadian Respiratory Journal 1997;4(3):131-9.

133. Sainsbury A, Seebass G, Bansal A, et al. Reliability of the Barthel Index when used with older people. Age Ageing 2005 May;34(3):228-32.

134. Wyller TB, Sveen U, Bautz-Holter E. The Barthel ADL index one year after stroke: comparison between relatives' and occupational therapist's scores. Age Ageing 1995 Sep;24(5):398-401.

135. Ranhoff AH, Laake K. The Barthel ADL index: scoring by the physician from patient interview is not reliable. Age Ageing 1993 May;22(3):171-4.

136. Hsueh IP, Lee MM, Hsieh CL. Psychometric characteristics of the Barthel activities of daily living index in stroke patients. J Formos Med Assoc 2001 Aug;100(8):526-32.

137. Hung SY, Pickard AS, Witt WP, et al. Pain and depression in caregivers affected their perception of pain in stroke patients. J Clin Epidemiol 2007 Sep;60(9):963-70.

138. Pickard AS, Johnson JA, Feeny DH, et al. Agreement between patient and proxy assessments of health-related quality of life after stroke using the EQ-5D and Health Utilities Index. Stroke 2004 Feb;35(2):607-12.

139. Dorman P, Slattery J, Farrell B, et al. Qualitative comparison of the reliability of health status assessments with the EuroQol and SF-36 questionnaires after stroke. United Kingdom Collaborators in the International Stroke Trial. Stroke 1998 Jan;29(1):63-8.

140. Coons SJ, Rao S, Keininger DL, et al. A comparative review of generic quality-of-life instruments. Pharmacoeconomics 2000 Jan;17(1):13-35.

141. Fisk JD, Brown MG, Sketris IS, et al. A comparison of health utility measures for the evaluation of multiple sclerosis treatments. J Neurol Neurosurg Psychiatry 2005 Jan;76(1):58-63.

142. Marra CA, Rashidi AA, Guh D, et al. Are indirect utility measures reliable and responsive in rheumatoid arthritis patients? Qual Life Res 2005 Jun;14(5):1333-44.

143. Thoma A, Sprague S, Veltri K, et al. Methodology and measurement properties of health-related quality of life instruments: a prospective study of patients undergoing breast reduction surgery. Health Qual Life Outcomes 2005 Jul 22;3:44.

144. Wiebe S, Eliasziw M, Matijevic S. Changes in quality of life in epilepsy: how large must they be to be real? Epilepsia 2001 Jan;42(1):113-8.

145. Boyle MH, Furlong W, Feeny D, et al. Reliability of the Health Utilities Index--Mark III used in the 1991 cycle 6 Canadian General Social Survey Health Questionnaire. Quality of Life Research 1995 Jun;4(3):249-57.

146. Wijkstra PJ, TenVergert EM, Van AR, et al. Reliability and validity of the chronic respiratory questionnaire (CRQ). Thorax 1994 May;49(5):465-7.

147. Martin LL. Validity and reliability of a quality-of-life instrument: the chronic respiratory disease questionnaire. Clin Nurs Res 1994 May;3(2):146-56.

148. Green J, Forster A, Young J. A test-retest reliability study of the Barthel Index, the Rivermead Mobility Index, the Nottingham Extended Activities of Daily Living Scale and the Frenchay Activities Index in stroke patients. Disabil Rehabil 2001 Oct 15;23(15):670-6.

149. Jones CA, Feeny D, Eng K. Test-retest reliability of health utilities index scores: evidence from hip fracture. Int J Technol Assess Health Care 2005;21(3):393-8.

150. Naglie G, Tomlinson G, Tansey C, et al. Utility-based quality of life measures in Alzheimer's disease. Qual Life Res 2006 May 15;15(4):631-43.

151. Jones CA, Pohar SL, Warren S, et al. The burden of multiple sclerosis: a community health survey. Health Qual Life Outcomes 2008;6:1.

152. Wilkinson PR, Wolfe CD, Warburton FG, et al. Longer term quality of life and outcome in stroke patients: is the Barthel index alone an adequate measure of outcome? Qual Health Care 1997 Sep;6(3):125-30.

153. Barberger-Gateau P, Commenges D, Gagnon M, et al. Instrumental activities of daily living as a screening tool for cognitive impairment and dementia in elderly community dwellers. J Am Geriatr Soc 1992 Nov;40(11):1129-34.

154. Hancock P, Larner AJ. The diagnosis of dementia: diagnostic accuracy of an instrument measuring activities of daily living in a clinic-based population. Dement Geriatr Cogn Disord 2007;23(3):133-9.

155. Grutters JP, Joore MA, van der HF, et al. Choosing between measures: comparison of EQ-5D, HUI2 and HUI3 in persons with hearing complaints. Quality of Life Research 2007 Oct;16(8):1439-49.

156. Neumann PJ, Kuntz KM, Leon J, et al. Health utilities in Alzheimer's disease: a cross-sectional study of patients and caregivers. Med Care 1999 Jan;37(1):27-32.

157. Sawatzky R, Liu-Ambrose T, Miller WC, et al. Physical activity as a mediator of the impact of chronic conditions on quality of life in older adults. Health Qual Life Outcomes 2007;5:68.

158. Robert SA, Cherepanov D, Palta M, et al. Socioeconomic status and age variations in health-related quality of life: results from the national health measurement study. J Gerontol B Psychol Sci Soc Sci 2009 May;64(3):378-89.

159. Huguet N, Kaplan MS, Feeny D. Socioeconomic status and health-related quality of life among elderly people: results from the Joint Canada/United States Survey of Health. Soc Sci Med 2008 Feb;66(4):803-10.

160. Wexler DJ, Grant RW, Wittenberg E, et al. Correlates of health-related quality of life in type 2 diabetes. Diabetologia 2006 Jul;49(7):1489-97.

161. Garster NC, Palta M, Sweitzer NK, et al. Measuring health-related quality of life in population-based studies of coronary heart disease: comparing six generic indexes and a disease-specific proxy score. Qual Life Res 2009 Sep 16;18(9):1239-47.

162. Pohar SL, Jones CA. The burden of Parkinson disease (PD) and concomitant comorbidities. Arch Gerontol Geriatr 2009 Sep;49(2):317-21. PMID: S0167-4943(08)00229-X [pii];10.1016/j.archger.2008.11.006 [doi].

163. Harper R, Brazier JE, Waterhouse JC, et al. Comparison of outcome measures for patients with chronic obstructive pulmonary disease (COPD) in an outpatient setting. Thorax 1997 Oct;52(10):879-87.

164. Guyatt GH, Townsend M, Keller J, et al. Measuring functional status in chronic lung disease: conclusions from a randomized control trial. Respir Med 1989 Jul;83(4):293-7.

165. Wallace D, Duncan PW, Lai SM. Comparison of the responsiveness of the Barthel Index and the motor component of the Functional Independence Measure in stroke: the impact of using different methods for measuring responsiveness. J Clin Epidemiol 2002 Sep;55(9):922-8.

166. van der Putten JJ, Hobart JC, Freeman JA, et al. Measuring change in disability after inpatient rehabilitation: comparison of the responsiveness of the Barthel index and the Functional Independence Measure. J Neurol Neurosurg Psychiatry 1999 Apr;66(4):480-4.

167. Ware JE. SF-36 health survey update. 2011. http://www.sf-35.org/tools/sf35.shtml#MODEL. Accessed November 2, 2010.

168. O'Mahony PG, Rodgers H, Thomson RG, et al. Is the SF-36 suitable for assessing health status of older stroke patients? Age Ageing 1998 Jan;27(1):19-22.

169. O'Connor RJ, Cano SJ, Thompson AJ, et al. Exploring rating scale responsiveness: does the total score reflect the sum of its parts? Neurology 2004 May 25;62(10):1842-4.

170. Wellwood I, Dennis MS, Warlow CP. A comparison of the Barthel Index and the OPCS disability instrument used to measure outcome after acute stroke. Age Ageing 1995 Jan;24(1):54-7.

171. Duncan PW, Samsa GP, Weinberger M, et al. Health status of individuals with mild stroke. Stroke 1997 Apr;28(4):740-5.

172. Balu S. Differences in psychometric properties, cut-off scores, and outcomes between the Barthel Index and Modified Rankin Scale in pharmacotherapy-based stroke trials: systematic literature review. Curr Med Res Opin 2009 Jun;25(6):1329-41.

173. Wyrwich KW, Tierney WM, Babu AN, et al. A comparison of clinically important differences in health-related quality of life for patients with chronic lung disease, asthma, or heart disease. Health Serv Res 2005 Apr;40(2):577-91.

174. Ware JE, Keller SD. Interpreting general health measures. In: Spilker BF (eds). Quality of Life and Pharmacoeconomics in Clinical Trials. Second ed. Philadelphia: Lippincott-Raven; 2010. p. 445-60.

175. Hsieh YW, Wang CH, Wu SC, et al. Establishing the minimal clinically important difference of the Barthel Index in stroke patients. Neurorehabil Neural Repair 2007 May;21(3):233-8.

176. Bowling A. Measuring Health: A Guide to Rating Scales and Questionnaires. Buckingham, UK: Open University Press; 2001.

177. Redelmeier DA, Guyatt GH, Goldstein RS. Assessing the minimal important difference in symptoms: a comparison of two techniques. J Clin Epidemiol 1996 Nov;49(11):1215-9.

178. Hobart JC, Cano SJ, Thompson AJ. Effect sizes can be misleading: is it time to change the way we measure change? J Neurol Neurosurg Psychiatry 2010 Sep;81(9):1044-8.

Appendix A. Measurement Properties

Table 1. Brief review of evidence on measurement properties for generic measures and selected disease-specific measures frequently used to assess health status and health-related quality of life in the older adults*

	SF-36	Barthel ADL	Lawton IADL	HUI3	CRQ
Content Validity	Built upon previous measures. Represents health concepts most frequently included in health surveys and additional concepts strongly supported by empirical evidence[8]	Developed to determine the amount of nursing care hospital patients undergoing rehabilitation would need[124]	No information on how items were selected[125]	Theoretical and empirical evidence have guided the creation of HUI3; selection of attributes was guided by the importance the general population placed on each attribute[9]	The items were generated from a literature review, consultations with health professionals, and interviews with patients about concerns/problems most important to them[126]
Internal Consistency	Moderate to high levels for all domains, Cronbach's alpha ranging from 0.49 (Social Functioning) to 0.96 (Physical Functioning)[127] Eight factor solutions supporting each of the domains were supported by factor analysis as well as two factor solutions for the 2 component scores (physical and mental health)[127]	Factor analysis indicated that the instrument was unidimensional for stroke patients, but multidimensional for geriatric and hip-fracture patients[128] Rasch analysis indicated that using a total score was not appropriate for older adults in the acute care setting[129] Cronbach's alpha of 0.84 upon admission and 0.85 upon discharge for stroke inpatients receiving rehabilitation[130] Internal consistency coefficients of 0.87 (admission) to 0.92 (discharge)[124]	Acceptable internal consistency (Cronbach's alpha > 0.7)[125] Two subscales found with factor analysis with Cronbach alpha of 0.91 and 0.78[125]	Little overlap among attributes, ranging from 0.02 (vision and speech correlations) to 0.35 (emotion and cognition correlations)[22]	Cronbach's alpha ranges from 0.76 (Mastery) to 0.90 (Emotional function) and 0.93 total[131] Cronbach's alpha ranges from 0.51 (Dyspnea) to 0.88 (Mastery)[132]
Inter-Observer Reliability *In general*					

	SF-36	Barthel ADL	Lawton IADL	HUI3	CRQ
Inter-Observer Reliability *In older adults*		Fair-moderate agreement for individual items, high percentage of agreement for the total score[133] Acceptable agreement for the total score between a doctor interview of a close relative and occupational therapist; kappa values ranged from 0.42 to 0.92[134] Self-report was the least reliable compared to physiotherapist testing and nurse assessment or testing and; agreement was also lowest for items on transfers, feeding, dressing, grooming, and toileting[124] Physician's score from interview tended to be higher than nurses' scores from observations among short-stay patients; only 4 individual items had a kappa coefficient above 0.40[135] For stroke patients, weighted kappa statistics ranged from 0.53 to 0.94 for individual items and the ICC was 0.94 for the total score[136]		Among caregivers and stroke patients, caregivers with pain overestimated patient pain and depressed caregivers underestimated patient pain[137] ICC>0.7 for patient-proxy responses at 1, 3, and 6 months post-stroke. ICC=0.59 at baseline[138] Overall score had ICC=0.70 for patient-proxy responses at baseline and ICC=0.86 at 6 months[79]	

	SF-36	Barthel ADL	Lawton IADL	HUI3	CRQ
Test-Retest Reliability *In general*	Generally good test-retest (ICC ranging from 0.57 to 0.80, except Mental health at 0.28) among stroke patients[139]	For patients retested after 3 weeks, scores for 35 of 41 patients were within 10 points of the original score[124]		Kappa values for attributes ranged from 0.14 to 0.73. ICC for overall scores was 0.73[23] 8 of 10 individual questions and 6 of the 8 attributes had moderate or better kappa coefficients[140] ICC of 0.87 for MS subjects[141] Test-retest for rheumatoid arthritis patients, ICC: 0.81 (0.66-0.90)[142] Good test-retest for breast hypertrophy patients, ICC=0.84[143] Test-retest among epilepsy patients, 0.87 +/- 0.3 (95%)[144] ICC of 0.77 from a population survey[145]	High test-retest reliability for Fatiguq, Emotion and Master (Spearman-Brown reliability coefficient ≥0.9), lower test-retest for Dyspnea (0.73)[146] No trends towards improvement or deterioration in stable COPD patients who were administered the test 6 times at 2-week intervals[126] High degree of test-retest reliability among Dyspnea, Emotional function, and Mastery[147]
Test-Retest Reliability *In older adults*	Low to high levels for all domains, ranging from 0.24 (Social Functioning) to 0.87 (General Health Perceptions). Most domains have high levels of reliability, except Social Functioning and Role Limitations – Emotional[127]	No studies of test-retest in general older adult population[133] Among stroke patients, agreement was >75% for individual items[148]		Acceptable test-retest reliability for hip fracture patients[149] Test-retest reliability intra-class correlation coefficient for mild cognitive impairment 0.75 (0.32-0.92); moderate impairment, 0.25 (0.00-0.74)[150]	

	SF-36	Barthel ADL	Lawton IADL	HUI3	CRQ
Cross-sectional Construct Validity *In general*	Discriminating between individuals with chronic medical illness and psychiatric, varying severity of medical conditions, osteoarthritis, epilepsy, depressive symptoms, panic disorder, total hip replacement, migraine, missing work due to illness, and varicose vein surgery[140]			Demonstrated validity with childhood cancer, adult oncology, population health survey[22], colorectal cancer, stroke, arthritis[23], neurological disability[141], and MS[151]	
Cross-sectional Construct Validity *In older adults*	Items correlate more highly with the proposed domain than with other domains[127]	Rank correlation coefficient with SF-36 in stroke patients ranged from 0.22 for Role Limitations – Emotional and 0.81 for Physical Functioning subscales[152] Rank correlation coefficient with the Nottingham health profile for stroke patients ranged from -0.19 for Sleep and -0.84 for Physical Mobility subscales[152] Scores correlated with Berg balance scale and Fugl-Meyer motor assessment at stroke recovery stages[136]	4 IADL items (telephone, medications, finances, and transportation) were associated with cognitive impairment in older community-dwelling adults[153] Was not helpful in identifying dementia in a clinic-based population[154] Indeterminate construct validity[125]	Demonstrated validity among groups with hearing loss[155], Alzheimer's disease[156], chronic conditions[157], socioeconomic status[158,159], type 2 diabetes[160], coronary heart disease[161], Parkinson disease[162], socioeconomic status	

	SF-36	Barthel ADL	Lawton IADL	HUI3	CRQ
Longitudinal Construct Validity (Responsiveness) *In general*	SF-36 scales and summary scores have been linked to utilization of health care services, progression of depression, loss of job within 1 year, and 5-year survival. Physical functioning, Role-physical, and Bodily pain are responsive to knee and hip replacement and heart valve surgery. Mental health, Role-emotional, and Social functioning are responsive to recovery from depression[45]			Responsive to treatments of osteoarthritis of the knee and elective total hip arthroplasty for osteoarthritis[22]	Responsive to changes after respiratory rehabilitation (Guyatt 1987, de Torres 2002), changes 10 days post acute COPD exacerbation (Aaron 2002), improvements and deteriorations in how patients felt[163] Dyspnea section was responsive to changes after treatment for patients with chronic airflow limitation[164]

	SF-36	Barthel ADL	Lawton IADL	HUI3	CRQ
Longitudinal Construct Validity (Responsiveness) *In older adults*	Hypothetical improvement in health states was associated with small to large effect sizes in community-dwelling older women[127]	Admission scores predicted mortality, length of hospital stay and subsequent progress among stroke patients[124] It was difficult for the index to obtain a change score for those at the upper or lower score ranges for older adults in the acute care setting; the index does not have appropriate scale width to monitor changes[129] Responsive to change from 1 to 3 months in recovering stroke patients[165] Responsive to change in patients undergoing inpatient neurorehabilitation[166] Responsive to changes over time in stroke patients[136]	Indeterminate responsiveness[125]	Measured significant improvement after hearing aid fitting[155]	

	SF-36	Barthel ADL	Lawton IADL	HUI3	CRQ
Evidence of Floor Effects *In general*	No floor effects for the Physical and Mental component summary scores observed in the general U.S. population. About 10% were observed to have the lowest scores in role-emotional and role-physical[167] Floor effects for people aged over 45 years who had a stroke for physical functioning (18%), role physical (54%), vitality (10%), social functioning (17%), and role emotional (35%)[168]			No floor effects on the subscales for MS subjects[141]	No floor effects[163]
Evidence of Floor Effects *In older adults*	Developers suggested that older adults may have more floor effects because they may have more sickness than the general population. Floor effects in excess of 20% were reported for Role Limitations Emotional and Physical by 12 studies[127]	Minimal floor effects for the total score for multiple sclerosis, stroke, and spinal cord injury patients upon admission to a neurorehabilitation unit. Seven of 10 individual items had floor effects[169] Floor and ceiling effects may lead to underestimating problems in a third of stroke patients[170]			

	SF-36	Barthel ADL	Lawton IADL	HUI3	CRQ
Evidence of Ceiling Effects *In general*	No ceiling effects for the Physical and Mental component summary scores observed in the general U.S. population. 40% in physical functioning, 71% in role-physical, 32% in bodily pain, 52% in social functioning, and 71% in role emotional were observed to have the highest scores[167] Ceiling effects for people aged over 45 years who had a stroke for role physical (16%), bodily pain (25%), social functioning (18%), role emotional (51%), and mental health (12%)[168]			Ceiling effects were present in only 3% of MS subjects for the overall utility and each of the subscales[141] May be problematic for population screening and long-term followup studies[23]	No ceiling effects[163]

	SF-36	Barthel ADL	Lawton IADL	HUI3	CRQ
Evidence of Ceiling Effects *In older adults*	Ceiling effects in excess of 20% for Role Limitations Emotional and Physical and Social Functioning (includes 16 studies)[127]	Minimal ceiling effects for patients upon admission to a neurorehabilitation unit for multiple sclerosis, stroke, and spinal cord injury. Ceiling effects were present at discharge. Nine of 10 individual items had ceiling effects upon admission and ceiling effects increased at discharge[169] Unacceptable ceiling effects for older adults in the acute care setting[129] Ceiling effects among patients recovering from a stroke or transient ischemic attack[171] Various studies have shown that the index has ceiling effects among stroke patients[172]	20% of dementia patients obtained the highest score[125] In clinic patients, most achieved a high IADL score[154]		

	SF-36	Barthel ADL	Lawton IADL	HUI3	CRQ
Interpretation	The smallest amount that the SF-36 score can change if patients move up or down one response level varies from 5-12.5, although the clinically important differences are higher for asthma, COPD, and heart disease patients[173] The SF-36 ranges from 0-100, but it is not an interval scale. For example, a change of 10 points means something different when going from 40 to 50 versus 85 to 95[174]	A change of 1.85 in the total score can be considered the minimally important difference for stroke patients[175] The Barthel index ranges from 0-20 or 0-100, but it is not an interval scale. Equal changes in scores for individual items do not correspond to equal changes in functioning[176]		Changes of 0.03 in overall scores are important and in some situations, 0.01 may be meaningful. Within attributes, changes of 0.05 are meaningful[9]	Mean clinically important difference of 0.5[177]

	SF-36	Barthel ADL	Lawton IADL	HUI3	CRQ
Comments		BI has considerable imprecision (95% CI of ±4 points; 20 point scale)[133] Group level indicators of responsiveness (e.g. effect sizes, standardized mean differences) are potentially misleading for the BI[178] Designed for use with long-term hospital patients with neuromuscular or musculoskeletal disorders; only suitable for the institutionalized populations for which it was designed[176] BI has been extensively studied in stroke populations, but less studied in the general older adult population.			Normal distribution[131]

Note: This table is intended to be illustrative, rather than a comprehensive review.
Abbreviations: ADL = activities of daily living; BI = Barthel Index; CI = confidence interval; COPD = chronic obstructive pulmonary disease; CRQ = chronic respiratory questionnaire; HUI = Health Utilities Index; IADL = instrumental activities of daily living; ICC = intraclass correlation; SF = short form; U.S. = United States.
*References are located in the reference list for the body of the paper.