A Primer for Systematic Reviewers on the Measurement of Functional Status and Health-Related Quality of Life in Older Adults
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. 290-2007-10057-1

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

Stephanie Chang, M.D., M.P.H.  Suchitra Iyer, Ph.D.
Director  Task Order Officer
Evidence-based Practice Program  Evidence-based Practice Program
Center for Outcomes and Evidence  Center for Outcomes and Evidence
Agency for Healthcare Research and Quality  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 ........................................................................................................................................ 1

Patient Reported Outcomes, Health-Related Quality of Life, and Function: An Overview of Measurement Properties ................................................................................................................................. 2
  Terminology .................................................................................................................................... 2
  Classification of Health-Related Quality of Life Measures ........................................................... 2
    Quality-Adjusted Survival ............................................................................................................. 3
  Definitions ...................................................................................................................................... 4
    The Distinction Between Predictive Validity and Responsiveness .............................................. 5
  How Should Reviewers Approach These Measurement Properties? ........................................ 5
  Reviewing Measures of Health-Related Quality of Life: Special Considerations for Older Adults Populations .......................................................................................................................... 6
  Floor and Ceiling Effects .............................................................................................................. 7
  Proxy Respondents ...................................................................................................................... 7

What To Assess and How To Interpret Measures of Functional Status and Health-Related Quality of Life ................................................................................................................................. 9
  Conceptual Framework ................................................................................................................ 9
  Minimum Important Difference ................................................................................................. 9
  Absolute or Relative Change? .................................................................................................... 10
  Observational Data .................................................................................................................... 11

Implications for Researchers ........................................................................................................ 12
  Assessing Health-Related Quality of Life .................................................................................. 12

Implications for Systematic Reviewers ......................................................................................... 14

Summary and Conclusions ........................................................................................................... 15

Brief Definitions of Important Measurement Properties ................................................................. 16

References ...................................................................................................................................... 16

Tables ............................................................................................................................................ 17
  Table 1. Brief definitions of important measurement properties ................................................ 16

Figures .......................................................................................................................................... 9
  Figure 1. Conceptual framework ................................................................................................ 9

Appendix ....................................................................................................................................... 16
  Appendix A. Measurement Properties
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. 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. 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.

Functional Status: Starfield: “The capacity to engage in activities of daily living and social activities”.

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. 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). “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. 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: health profiles such as the Short-Form 36 and preference-based measures such as the Health Utilities Index. 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\textsuperscript{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.\textsuperscript{12}

Measures can also be classified by their intended purpose.\textsuperscript{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.\textsuperscript{15} The psychometric paradigm draws from psychology and typically relies on a latent-variable model.\textsuperscript{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.\textsuperscript{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.\textsuperscript{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.\textsuperscript{20} In the multi-attribute approach, the subject completes a questionnaire based on the measure; examples of multi-attribute measures include the EQ-5D\textsuperscript{21} and Health Utilities Index (HUI).\textsuperscript{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.\textsuperscript{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.\textsuperscript{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.\textsuperscript{29} Cohen provides a scheme to interpret the magnitude of ES: small (0.20); moderate (0.50); or large (≥ 0.80) change.\textsuperscript{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.\textsuperscript{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.\textsuperscript{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.\textsuperscript{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 QWB52. 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. 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-5D21 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. 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. 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. 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\textsuperscript{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

Another prominent framework is provided by the World Health Organization International Classification of Functioning, Disability, and Health (ICF).\textsuperscript{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”. 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.”

MIDs are estimated using anchor-based or distribution-based approaches. 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. 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. 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, a change of 10 is equivalent to an ES of 0.43, moderate in the scheme proposed by Cohen. 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. 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.

### 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. 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. 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. 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.

**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. 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. 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. 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. Respondents have been followed since, most recently surveyed at 15 years of followup (from 2003 to 2005). HRQL measures used in the Beaver Dam study included the SF-36, the QWB, and the time-trade off. 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.” 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.” 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. 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. 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. 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). 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.\textsuperscript{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.\textsuperscript{15} In particular Haywood and colleagues report that measures acceptable to general adult populations are sometimes not acceptable to older respondents.\textsuperscript{45}
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

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliability</td>
<td>A reliable measure is consistent and reproducible.</td>
</tr>
<tr>
<td>Internal Consistency</td>
<td>The extent to which items are measuring the same concept.</td>
</tr>
<tr>
<td>Intra- and Inter-Observer Reliability</td>
<td>The extent of agreement across assessments or among individuals.</td>
</tr>
<tr>
<td>Validity</td>
<td>The measure accurately reflects the concept it is intended to measure.</td>
</tr>
<tr>
<td>Content Validity</td>
<td>The extent to which the measure covers the full range of meanings included in the concept.</td>
</tr>
<tr>
<td>Criterion Validity</td>
<td>The extent of agreement between the measure and a gold standard measure of the same concept.</td>
</tr>
<tr>
<td>Construct Validity</td>
<td>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.</td>
</tr>
<tr>
<td>Convergent Validity</td>
<td>Convergent validity refers to evidence of a moderate or strong relationship between measures of the same concept or construct.</td>
</tr>
<tr>
<td>Discriminant Validity</td>
<td>Discriminant validity refers to evidence of the lack of relationship between measures of a different concept or construct.</td>
</tr>
<tr>
<td>Cross-Sectional Construct Validity</td>
<td>Evidence of construct validity based on comparisons at a point in time.</td>
</tr>
<tr>
<td>Responsiveness (Longitudinal Construct Validity)</td>
<td>The ability of a measure to capture meaningful change when it occurs.</td>
</tr>
<tr>
<td>Interpretation</td>
<td>The ability to attach meaning to the scores provided by a measure.</td>
</tr>
</tbody>
</table>

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


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


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**

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Content Validity</th>
<th>Internal Consistency</th>
<th>Inter-Observer Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>SF-36</td>
<td>Built upon previous measures. Represents health concepts most frequently included in health surveys and additional concepts strongly supported by empirical evidence.</td>
<td>Moderate to high levels for all domains, Cronbach's alpha ranging from 0.49 (Social Functioning) to 0.96 (Physical Functioning). Factor analysis indicated that the instrument was unidimensional for stroke patients, but multidimensional for geriatric and hip-fracture patients. Rasch analysis indicated that using a total score was not appropriate for older adults in the acute care setting. Cronbach's alpha of 0.84 upon admission and 0.85 upon discharge for stroke inpatients receiving rehabilitation.</td>
<td>In general</td>
</tr>
<tr>
<td>Barthel ADL</td>
<td>Developed to determine the amount of nursing care hospital patients undergoing rehabilitation would need.</td>
<td>Acceptable internal consistency (Cronbach's alpha &gt; 0.7). Two subscales found with factor analysis with Cronbach alpha of 0.91 and 0.78.</td>
<td></td>
</tr>
<tr>
<td>Lawton IADL</td>
<td>No information on how items were selected.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HUI3</td>
<td>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. Theoretically and empirically evidence have guided the creation of HUI3; selection of attributes was guided by the importance the general population placed on each attribute.</td>
<td>Little overlap among attributes, ranging from 0.02 (vision and speech correlations) to 0.35 (emotion and cognition correlations).</td>
<td></td>
</tr>
<tr>
<td>CRQ</td>
<td>The items were generated from a literature review, consultations with health professionals, and interviews with patients about concerns/problems most important to them.</td>
<td>Cronbach's alpha ranges from 0.76 (Mastery) to 0.90 (Emotional function) and 0.93 total. Cronbach's alpha ranges from 0.51 (Dyspnea) to 0.88 (Mastery).</td>
<td></td>
</tr>
<tr>
<td>SF-36</td>
<td>Barthel ADL</td>
<td>Lawton IADL</td>
<td>HUI3</td>
</tr>
<tr>
<td>-------</td>
<td>-------------</td>
<td>-------------</td>
<td>------</td>
</tr>
<tr>
<td><strong>Inter-Observer Reliability</strong>&lt;br&gt; <em>In older adults</em></td>
<td>\begin{itemize} \item Fair-moderate agreement for individual items, high percentage of agreement for the total score \end{itemize}</td>
<td>\begin{itemize} \item 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 \end{itemize}</td>
<td>\begin{itemize} \item Among caregivers and stroke patients, caregivers with pain overestimated patient pain and depressed caregivers underestimated patient pain \end{itemize}</td>
</tr>
<tr>
<td></td>
<td>\begin{itemize} \item 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 \end{itemize}</td>
<td>\begin{itemize} \item 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 \end{itemize}</td>
<td>\begin{itemize} \item Overall score had ICC=0.70 for patient-proxy responses at baseline and ICC=0.86 at 6 months \end{itemize}</td>
</tr>
<tr>
<td>Test-Retest Reliability</td>
<td>SF-36</td>
<td>Barthel ADL</td>
<td>Lawton IADL</td>
</tr>
<tr>
<td>------------------------</td>
<td>-------</td>
<td>------------</td>
<td>-------------</td>
</tr>
<tr>
<td><strong>In general</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Generally good test-retest (ICC ranging from 0.57 to 0.80, except Mental health at 0.28) among stroke patients</td>
<td></td>
<td>For patients retested after 3 weeks, scores for 35 of 41 patients were within 10 points of the original score</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test-retest for rheumatoid arthritis patients, ICC: 0.81 (0.66-0.90)</td>
<td></td>
<td>Good test-retest for breast hypertrophy patients, ICC=0.84</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Test-retest among epilepsy patients, 0.87 +/- 0.3 (95%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ICC of 0.77 from a population survey</td>
<td></td>
</tr>
<tr>
<td><strong>In older adults</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>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</td>
<td></td>
<td>No studies of test-retest in general older adult population</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Among stroke patients, agreement was &gt;75% for individual items</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>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)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A-3
<table>
<thead>
<tr>
<th>Cross-sectional Construct Validity</th>
<th>SF-36</th>
<th>Barthel ADL</th>
<th>Lawton IADL</th>
<th>HUI3</th>
<th>CRQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>In general</td>
<td>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&lt;sup&gt;140&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td>Demonstrated validity with childhood cancer, adult oncology, population health survey&lt;sup&gt;22&lt;/sup&gt;, colorectal cancer, stroke, arthritis&lt;sup&gt;23&lt;/sup&gt;, neurological disability&lt;sup&gt;141&lt;/sup&gt;, and MS&lt;sup&gt;151&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

| Cross-sectional Construct Validity | Items correlate more highly with the proposed domain than with other domains<sup>127</sup> | Rank correlation coefficient with SF-36 in stroke patients ranged from 0.22 for Role Limitations – Emotional and 0.81 for Physical Functioning subscales<sup>152</sup> | 4 IADL items (telephone, medications, finances, and transportation) were associated with cognitive impairment in older community-dwelling adults<sup>153</sup> | | | Demonstrated validity among groups with hearing loss<sup>155</sup>, Alzheimer’s disease<sup>156</sup>, chronic conditions<sup>157</sup>, socioeconomic status<sup>158,159</sup>, type 2 diabetes<sup>160</sup>, coronary heart disease<sup>161</sup>, Parkinson disease<sup>162</sup>, socioeconomic status<sup>163</sup> |
| In older adults                   | Rank correlation coefficient with the Nottingham health profile for stroke patients ranged from -0.19 for Sleep and -0.84 for Physical Mobility subscales<sup>152</sup> | Scores correlated with Berg balance scale and Fugl-Meyer motor assessment at stroke recovery stages<sup>136</sup> | Indeterminate construct validity<sup>125</sup> | | |

|                  | | | | | |
|                  | | | | | |

|                  | | | | | |
|                  | | | | | |

A-4
<table>
<thead>
<tr>
<th>SF-36</th>
<th>Barthel ADL</th>
<th>Lawton IADL</th>
<th>HUI3</th>
<th>CRQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Longitudinal Construct Validity (Responsiveness)</td>
<td>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.</td>
<td>Responsive to treatments of osteoarthritis of the knee and elective total hip arthroplasty for osteoarthritis.</td>
<td>Responsive to changes after respiratory rehabilitation (Guyatt 1997, de Torres 2002), changes 10 days post acute COPD exacerbation (Aaron 2002), improvements and deteriorations in how patients felt. Dyspnea section was responsive to changes after treatment for patients with chronic airflow limitation.</td>
<td></td>
</tr>
<tr>
<td>SF-36</td>
<td>Barthel ADL</td>
<td>Lawton IADL</td>
<td>HUI3</td>
<td>CRQ</td>
</tr>
<tr>
<td>-------</td>
<td>-------------</td>
<td>-------------</td>
<td>------</td>
<td>-----</td>
</tr>
<tr>
<td><strong>Longitudinal Construct Validity (Responsiveness)</strong>&lt;br&gt; <em>In older adults</em></td>
<td>Hypothetical improvement in health states was associated with small to large effect sizes in community-dwelling older women(^{127})</td>
<td>Admission scores predicted mortality, length of hospital stay and subsequent progress among stroke patients(^{124})</td>
<td>Indeterminate responsiveness(^{125})</td>
<td>Measured significant improvement after hearing aid fitting(^{155})</td>
</tr>
<tr>
<td></td>
<td>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})</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Responsive to change from 1 to 3 months in recovering stroke patients(^{165})</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Responsive to change in patients undergoing inpatient neurorehabilitation(^{166})</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Responsive to changes over time in stroke patients(^{136})</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evidence of Floor Effects</td>
<td>SF-36</td>
<td>Barthel ADL</td>
<td>Lawton IADL</td>
<td>HUI3</td>
</tr>
<tr>
<td>---------------------------</td>
<td>-------</td>
<td>-------------</td>
<td>-------------</td>
<td>------</td>
</tr>
<tr>
<td><strong>In general</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>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})</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>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})</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>No floor effects on the subscales for MS subjects(^{141})</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>No floor effects(^{163})</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>In older adults</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>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})</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>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})</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Floor and ceiling effects may lead to underestimating problems in a third of stroke patients(^{170})</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Evidence of Ceiling Effects

<table>
<thead>
<tr>
<th>SF-36</th>
<th>Barthel ADL</th>
<th>Lawton IADL</th>
<th>HUI3</th>
<th>CRQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>In general</td>
<td>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</td>
<td>Ceiling effects were present in only 3% of MS subjects for the overall utility and each of the subscales.41 May be problematic for population screening and long-term followup studies.23</td>
<td>No ceiling effects.163</td>
<td></td>
</tr>
<tr>
<td>Evidence of Ceiling Effects</td>
<td>SF-36</td>
<td>Barthel ADL</td>
<td>Lawton IADL</td>
<td>HUI3</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-------</td>
<td>-------------</td>
<td>-------------</td>
<td>------</td>
</tr>
<tr>
<td>In older adults</td>
<td>Ceiling effects in excess of 20% for Role Limitations Emotional and Physical and Social Functioning (includes 16 studies)</td>
<td>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.</td>
<td>Unacceptable ceiling effects for older adults in the acute care setting. Ceiling effects among patients recovering from a stroke or transient ischemic attack. Various studies have shown that the index has ceiling effects among stroke patients.</td>
<td>20% of dementia patients obtained the highest score. In clinic patients, most achieved a high IADL score.</td>
</tr>
</tbody>
</table>

A-9
<table>
<thead>
<tr>
<th>Interpretation</th>
<th>SF-36</th>
<th>Barthel ADL</th>
<th>Lawton IADL</th>
<th>HUI3</th>
<th>CRQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>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. 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.</td>
<td>A change of 1.85 in the total score can be considered the minimally important difference for stroke patients. 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.</td>
<td>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.</td>
<td>Mean clinically important difference of 0.5.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SF-36</td>
<td>Barthel ADL</td>
<td>Lawton IADL</td>
<td>HUI3</td>
<td>CRQ</td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>------------</td>
<td>-------------</td>
<td>------</td>
<td>-----</td>
<td></td>
</tr>
</tbody>
</table>
| **Comments** | BI has considerable imprecision (95% CI of ±4 points; 20 point scale)\(^1\)\(^3\)\(^3\)

Group level indicators of responsiveness (e.g. effect sizes, standardized mean differences) are potentially misleading for the BI\(^1\)^78

Designed for use with long-term hospital patients with neuromuscular or musculoskeletal disorders; only suitable for the institutionalized populations for which it was designed\(^76\)

BI has been extensively studied in stroke populations, but less studied in the general older adult population.

| | Normal distribution \(^1\)\(^3\) \(^1\) |