

MEASURING HEALTH-RELATED QUALITY OF LIFE

Dennis G. Fryback
Professor Emeritus, Population Health Sciences
University of Wisconsin-Madison

*Workshop on Advancing Social Science Theory: The Importance of Common Metrics
The National Academies, Division of Behavioral and Social Sciences and Education
Washington, D.C., February, 2010*

1. Introduction

Over the past 40 years a small set of preference-based measures of health-related quality-of-life (HRQoL) has appeared. Several of these measures are now in use worldwide. There is controversy, sometimes heated, about which of these is “best” and indeed whether any are in fact good measures for what they purport to do and therefore whether new, better measures should be developed and the old ones discarded. I have long advocated for population-based data collection using these standardized measures.(Fryback, Dasbach, Klein, Klein, Dorn, Peterson and Martin, 1993; Fryback, Dunham, Palta, Hanmer, Buechner, Cherepanov, Herrington, Hays, Kaplan, Ganiats, Feeny and Kind, 2007) In addition I have argued against tweaking the existing measures to “improve” them or casting them aside to develop “better” measures.(Fryback, 2005)

In this paper I present an overview of these measures as a case study in standardized health measures. The paper will first locate these measures in the broad landscape of health measures for readers unfamiliar with this subject, then describe the specific measures in more detail. I will present examples where these measures as standardized measures have contributed to knowledge of health because of standardization, and discuss where they may have failed. Finally, I will discuss what I see as a path forward for retaining the essence of standardization of the measures while simultaneously improving their implementation.

2. Health measures – A Quick Typology.

Health may be measured at the population level, or for a group of individuals, or at the level of the individual person. These measures are not mutually exclusive, as most individual measures can be aggregated in some fashion to represent a group or population. Health measures may be roughly classified as indicators, disease-specific

measures, generic health profiles, and summary health-related quality of life indexes. These are briefly introduced next.

2.1. *Health Indicators.*

Many health data are collected in the form of health indicators. An indicator is a measure focused on one particular aspect of health in a population. For example, smoking rate and prevalence of overweight are indicators. Other indicators are rates of specific diseases and disabilities, or rates of services such as vaccinations or well-baby exams, infant mortality rates, and so forth. Among the oldest and most important of indicators are mortality rates and their associated statistics, life expectancies. The “leading health indicators” in Healthy People 2010, the current decadal health planning and policy exercise of the US government, have been selected because each reflects a major public health or health services concern in the population and is the focus of a directed health policy effort (http://www.healthypeople.gov/Document/HTML/uih/uih_4.htm, accessed Nov. 21, 2009).

Health indicators are important measures. They offer the most detailed view of health and explaining changes in health will almost certainly involve examining changes in single indicators. A population health data system should certainly be built on as comprehensive a set of health indicators as possible. But without higher level aggregate measures a list of indicators is just a list. It describes the details of health but does not offer summarization or evaluation of overall change. If some indicators show better health, some are unchanged, and some show worse health, then we are unable to make aggregate statements about whether in composite we are better off or not.

Michael Wolfson, Assistant Chief Statistician, Analysis and Development, at Statistics Canada, talks about the population health data pyramid as his vision of a health data system to support population health planning (fig 1.) where data allow more aggregative summarization and evaluation as one moves up the data structure and one may drill down to more and more disaggregate data for explanatory and descriptive purposes.

Data Pyramid for Population Health (after Wolfson)

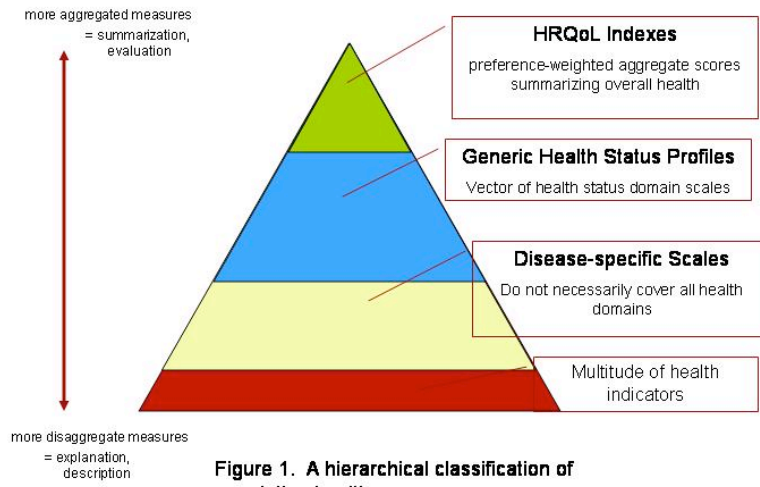


Figure 1. A hierarchical classification of population health measures.

At the bottom of the pyramid are health indicators and vital statistics. These serve as descriptors of health and mortality and perhaps as explanatory variables to help understand change or lack thereof in population health. As in the US Healthy People policy exercise, they can also serve to motivate policy and public action. But to make an assessment about whether on net the population is better or not requires more aggregation of data and some external value judgments about health and well-being. The US National Center for Health Statistics takes as its central charge the collection and distribution of vital statistics and health indicators. Indicators collected from individuals by NCHS include objective measurements such as those made by direct examination in the National Health and Nutrition Examination Survey (NHANES) (<http://www.cdc.gov/nchs/nhanes.htm>) which has been ongoing since the mid 1960s. Other indicators are collected by NCHS using questionnaires about health, health behaviors, and functioning in the National Health Interview Survey (NHIS) (<http://www.cdc.gov/nchs/nhis.htm>) which has been running since the late 1950s.

2.2. Disease-specific scales.

At the next level up from health indicators in the data pyramid are disease- or organ-specific scales. These may be used to assess treatment progress for a particular disease in an individual, or, if administered in population surveys, to help quantify burden of the disease in a population. Disease-specific scales often take the form of questionnaires with items assessing various aspects of symptoms and functioning marking various degrees of disease severity and impact that people may experience. Scores are usually computed by summing categorical responses across items. Three examples of disease-specific indexes are:

- *State-Trait Anxiety Inventory for Adults* (Spielberger, Gorsuch, Lushene, Vagg and Jacobs, 1983). This instrument contains 20 items forming a State scale of anxiety and 20 items forming a Trait

scale of anxiety. Typically these are presented to the individual for self-completion as a paper and pencil questionnaire. Items are statements such as “I feel calm” or “I am presently worrying over possible misfortunes.” (for State anxiety) and “I am a steady person” or “I lack self-confidence” (for Trait anxiety). Responses are categorical: 1=not at all, 2=somewhat, 3=moderately so, 4=very much so. Scales are summed numerical responses.

- *National Eye Institute Visual Functioning Questionnaire-25*. (Mangione, Lee, Pitts, Gutierrez, Berry and Hays, 1998) Twenty-five questions are asked, with response scales ranging from 2-6 categories. The first 4 questions ask respondents to rate their general health, their eyesight, worry about eyesight, and pain or discomfort around the eyes. The next section asks 12 questions concerning how vision problems may limit everyday functioning (e.g., “Because of your eyesight, how much difficulty do you have noticing objects off to the side while you are walking along?” “...do you have seeing how people react to things you say?”). The third section asks respondents 9 questions about how much of the time their ability to do things may be affected by their vision (e.g., “How much of the time do you accomplish less than you would like because of your vision?” “...are you limited in how long you can work or do other activities ...?”).
- *Oswestry Low Back Pain Disability Index* (Fairbank, Couper, Davies and O'Brien, 1980) is a 10-item questionnaire with 6 response categories for each item. Items cover pain intensity, personal care, lifting, walking, sitting, standing, sleeping, work, social life, and traveling. Scores are summed responses normalized to a 0-100 scale. Scores of 0-20 indicate minimal disability, while those with 40 or above are severely disabled by their back pain.

There are scores if not hundreds of disease-specific measures in the medical literature. A very common use of these scales is in clinical trials of new therapeutic interventions. Few (if any) of these scales are in the US national health data bases. The instruments are often proprietary and can be lengthy; and any given disease is rare in the population so the instrument would be irrelevant to the majority of survey respondents. These characteristics mitigate against administration of disease specific indexes in national surveys where competition for space on the protocols can be fierce. So most data using such instruments is in clinical data sets and research data sets.

A disease-specific measure rarely tells about the whole person (Patrick and Deyo, 1989). These scales have been constructed by selecting questions which are indicators of specific aspects of health the developers feel are important with respect to the particular disease or organ system. Aspects of health which may vary greatly from

person to person but which are not thought to be affected by the disease are usually not sampled. Thus an index concerned with diabetes impact may not ask about joint pain. Joint pain would be a major health dimension sampled in an index measuring impact of arthritis. Hearing is rarely sampled in arthritis-specific instruments, but over-the-counter medication for pain may be somewhat ototoxic and result in tinnitus or hearing loss. And, if only an arthritis-specific or an diabetes-specific instrument is administered to someone who has both arthritis and diabetes the single instrument may miss aspects of health very important to the person's daily life and how they feel in general.

2.3. Generic Health Status Profiles.

The purpose of a generic health status profile is to provide a coordinated summary of health of an individual for each important domain of health so that a composite, overall picture is obtained without being limited to one organ system or the likely effects of a particular disease. The overall health status of a person with multiple health conditions can be assessed. This begs the question of what is health.

In the preamble to the World Health Organization constitution in 1948 defined health as “a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity.”(WHO, 1948) This tripartite conceptualization of health into physical, mental, and social components has tended to dominate most frameworks for general health since that time. Of course a conceptual parsing of health into physical, mental, and social components need not stop with those concepts. Physical health may be parsed into mobility and ambulation, limitations on ability to do usual activities, pain, etc. Mental health could be parsed into cognitive function, emotional health and its limits on functioning, and so forth. Social health may involve abilities to interact with friends and family, intimacy, spirituality, and aspects of the social and physical environment in which the person lives. Although the latter, which go “beyond the skin” of the person, are acknowledged as important they are not included by many measures in order to keep the measure focused on a person as an entity with a set of generic capabilities for a general, unspecified environment. (The World Health Organization's quality of life measure, the WHOQOL (WHOQOL, 1998), is an exception as it has very broad coverage of domains both within and beyond the skin.)

Perhaps the most widely used health status profile in the world is the SF-36. It began as a 250-item questionnaire to assess the general health of people participating in the RAND Corporation's Health Insurance Experiment (HIE) in the 1970s.(Lohr, Brook, Kamberg, Goldberg, Leibowitz, Keesey, Reboussin and Newhouse, 1986; Lohr, Kamberg, Keeler, Goldberg, Calabro and Brook, 1986) People were assigned to different forms of health insurance and followed for a number of years. Researchers were interested in the costs of health care and health care utilization behavior under each form of insurance. They also needed to know whether health differed or not under the various insurance schemes. A long questionnaire was developed for this latter purpose. It was generated to comprehensively assess multiple dimensions of health generally following the WHO conceptualization.

Subsequent to the Health Insurance Experiment, researchers collaborating at RAND and in Boston wished to demonstrate a general model for evaluating the outcomes of medical care.(Tarlov, Ware, Greenfield, Nelson, Perrin and Zubkoff, 1989) They wanted to use a general health questionnaire, but the questionnaire in the HIE was much too long. Ware and colleagues developed a short form of the questionnaire using only 36 questions to cover 8 scales (Ware and Sherbourne, 1992): physical function (PF), role function as limited by physical health (RP), bodily pain (BP), social functioning (SF), mental health (MH), role function as limited by emotional health (RE), vitality (VT), and general health perception (GH) which is self-rated health. This was formally known as the Medical Outcome Study Short Form-36, or “MOS Short Form-36,” and now is just the “SF-36” for short.

Ware and colleagues copyrighted the SF-36 and distributed licenses for its use, free to academics and at some charge to insurance companies, clinics, and hospitals. He formed a company, QualityMetric, Inc., to distribute the SF-36, to provide consulting in its use, and which commissioned a private survey to establish population norms for the SF-36 scales. Because of some dispute over copyright among investigators, an unrestricted public version was distributed as the RAND-36.(Hays, Prince-Embury and Chen, 1988) Later, Ware and colleagues changed the instructions to respondents, changed the wording and responses for a number of the questions, and changed the formatting of the questionnaire to improve its psychometric properties. Also they changed scoring on the 8 scales from 0-100 scales to norm-based T-scores, where the population norm is 50 and the population standard deviation is 10. This version was copyrighted and trademarked by QualityMetrics and is now distributed in many languages and as self-completed questionnaire or as interviewer-administered questionnaire under the name SF-36v2™.(Ware, 2000) (<http://www.sf-36.org/>, accessed 20 Nov 2009)

The SF-36 was shortened further into a proprietary, 12-item questionnaire, the SF-12, now also in version 2, SF-12v2™. The Veterans Administration commissioned development of a 12-item version for its use, now known as the Veterans RAND 12 item health survey (VR-12).(Selim, Rogers, Fleishman, Qian, Fincke, Rothendler and Kazis, 2009)

I dwell on the SF-36 here because it is used worldwide in literally thousands of studies and practice settings and because it has undergone changes in questions since it was developed, an issue pertaining to standardizing questionnaires. Its 8 scales group into 2 clusters, generally denoted as physical health and mental health and (based on a factor analytic approach) scales may be computed for these two components, the physical component score (PCS) and the mental component score (MCS). These are proprietary scales derived from orthogonal factors and the scoring algorithms are distributed via a licensed user manual. Whether or not a measurement scale is proprietary is another issue in the health measures debates about using standardized instruments.

The SF-36 is the prototypical generic health profile since it summarizes health as a profile, or vector, of either 8 or 2 scores: (PF, RP, BP, GH, VT, SF, RE, MH) or (PCS,

MCS). An image in Figure 2 from the SF-36 website (<http://www.sf-36.org/tools/sf36.shtml>) shows an example of norm-based profiles for adult asthma patients before and after treatment. The line labeled “norm” is the age-relevant population average score on the scale, which is by design set to equal 50. Scores are scaled so that the standard deviation in the population is 10. The black bars show the profile before treatment and the gray bars show profiles after treatment. These results are interpreted to show that adults with asthma in this study were essentially average in amount of bodily pain, but nearly a standard deviation below average in physical function, physical limitations of role function, and self-assessed general health. It appears that asthma primarily affects physical health compared to mental health, although there are components of mental health (vitality, social function and emotional limits on role function) that appear to be somewhat affected. Treatment primarily remediates physical health functions.

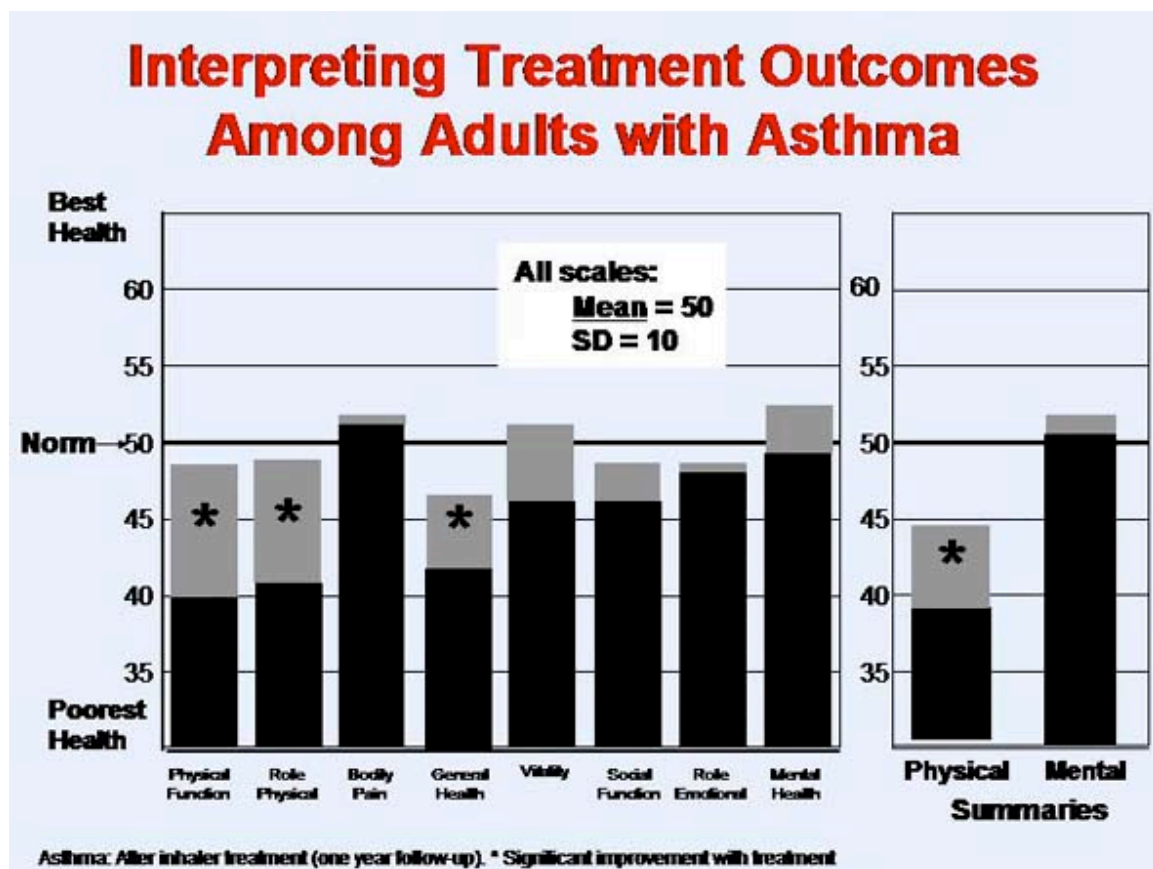


Figure 2. Image from SF-36.org showing an example of norm-based scoring of the SF-36 profile and interpretation of changes.

With population norms developed by QualityMetrics and norms in various patient populations established by the Medical Outcomes Study, availability of the single measure allows health providers (hospitals, clinics, etc.) and researchers to compare their specific population and results to general results in the population or other patient populations. The chances are good that readers of this paper aged 50 or older have

completed the SF-36 in association with medical care they have received. Ware and colleagues, by dogged determination and entrepreneurial spirit, brought health measurement to the broad stage of health care and population health.

That this discussion has focused on the SF-36 should not be interpreted to mean it is the only generic health profile. Examples of other generic health status profiles in the literature are the Sickness Impact Profile (SIP) (Bergner, Bobbitt, Carter and Gilson, 1981), the Nottingham Health Profile (Hunt, McKenna, McEwen, Williams and Papp, 1981), and the Duke Health Profile (Parkerson, Broadhead and Tse, 1990). However none has been as widely used and as influential as the SF-36.

To this point the discussion has centered on indexes, disease specific and generic health profiles, formed by summing categorical ratings. These scales are developed under aegis of classical test theory. Health is conceived of as being comprised of multiple domains – physical functioning, mental functioning, social functioning, self-perceived health, etc. Items (questions) are selected to sample targeted domains of health. The scales for each domain are functions of sums across responses to the germane items for the scale. Although the scales derive some ordinal meaning from the response categories, and generally (although not necessarily) people agree that larger scores represent better health in each domain, there is no way to compare two health states where the profile of one does not dominate the profile of the other health state (i.e., every component score is larger for one health state compared to the other). If the (PCS, MCS) vector for one person is (40, 55) and for another person is (55, 40) we cannot say the first person is either more or less healthy than the other – the two people are simply different. So the scores *describe*, but do not *value* health.

2.4. Health-Related Quality-of-Life Indexes—historical development of QWB, HUI2, HUI3, EQ-5D, and SF-6D.

The RAND health insurance experiment in the 1970s led to knowledge about behavioral influences of different insurance arrangements. In a parallel effort, in the late 1960s and early 1970s, other researchers also began to think about how to measure the health output of medical care in order to optimize use of health resources. They approached the problem using techniques of operations research and systems analysis. They noted that health care addressed two fundamentally different aspects of health outcome: mortality and morbidity. A central problem from this point of view was that there was a need for a measure that combined both mortality and morbidity into one single summary number.

Sullivan, at the Bureau of the Census, suggested disability-weighting life years to compute a health-adjusted life-expectancy index to summarize national population health. Changes in this index would reflect both changes in morbidity and in mortality. (Sullivan, 1966; Sullivan, 1971) Fanshel (a systems scientist in operations research) and Bush (a family physician) proposed an 11-category categorical index of functional health and well-being:

S_A-Well-being. This is a theoretical state analogous to the mathematical asymptote line. It corresponds to the World Health Organization's "positive physical, mental, and social well-being."

S_B-Dissatisfaction. In this state, all the subjective and social behavioral indicators are within acceptable limits, but there are undesirable conditions like dental caries, or air pollution. It includes much of the population at large not in a lower state, since almost everyone has some type of unsatisfactory condition that must receive some weight. It is a very slight, but significant, deviation from well-being.

S_C-Discomfort. This state arises from symptoms, such as colds, mild headaches, itches, irritabilities. Daily activities (work, school, family care) are continued with no significant reduction of efficiency.

S_D-Disability, minor. This state includes illness from whatever cause and/or emotional disturbance. Daily activities are continued but with significant reduction of efficiency.

S_E-Disability, major. This state includes persons who can carry on only in a restricted way the activities usual for their ages and sexes, such as special schools for the mentally retarded or sheltered workshops for adults. Therefore, there is a severe reduction of efficiency in the performance of their expected functions.

S_F-Disabled. Persons in this state are unable to go to school, to work, or to the equivalent, but are ambulatory and able to move about the community.

S_G-Confined. Here they are not bedridden, but very likely institutionalized.

S_H-Confined, bedridden. Whether the bed is at home, in a hospital or a nursing home is a matter of professional judgment, based on the cause of the illness and its prognosis; but in each instance here the person's functional status is confinement to bed.

S_I-Isolated. This state requires separation from family, friends, and activities, such as confinement to a special-care unit, operating room, delivery room, psychiatric security ward, or comparable isolation.

S_J-Coma. This state contains those with no significant functional distinction from death, except a nonzero probability of transition to a higher state.

S_K-Death. This state implies absolute dysfunction, with a zero transitional probability to a higher state.

NOTE: If a person is in surgery for several hours, but confined to bed before and after, he is in state S_I , i.e., we take the worst state of the day as the definition for the day.

It is important to note that no statement has been made about the cause for lack of well-being, whether it is due to air pollution, job tensions, poverty, or chronic disease. Thus, a person may be unable to work because of tuberculosis, alcoholism, mental retardation, or heart disease, but if he is not in a lower state, he is in state S_F (disabled). Likewise, air pollution may yield simply S_B (dissatisfaction) or be the cause of a respiratory illness requiring confinement to bed (S_H), or even death (S_K). Schizophrenia can vary from mild behavior disturbance (S_D), to hallucinations and confinement (S_I), or to suicide (S_K).

Similarly, function/dysfunction is independent of social role. A child may be inhibited in play (S_D), bedridden (S_H), or comatose (S_J). Old people may be symptomatic (S_C), shut-in (S_G), or bedridden (S_H), or in other states.

Treatment itself may be dysfunctional, even in the absence of illness. A person who goes into the hospital for a diagnostic evaluation is away from work and family, even if the work-up is negative. ... (Fanshel and Bush, 1970, pp.1029-30)

Fanshel and Bush proposed using judgments elicited from public officials to represent societal values to obtain weights for the states. They elaborated two different methods by which cardinal weights for the states S_B through S_J could be derived through the method of equivalence judgments once state S_A was assigned weight of 1.0 and S_K

assigned 0.0, i.e., scaled from total well-being to dead. With these cardinal weights, health experienced by a population over the course of a year could be computed by weighting each day with the weight of the worst state experienced that day for each person, and averaging this over people each day, and over days for the year. The ideal outcome would be 1.0 or an average of 1 dysfunction-free year for each person. By construction, a fractional average of, say, 0.6, would mean the average outcome is 6/10ths of a dysfunction-free year. The health policy optimization model proposed by Fanshel and Bush sought maximize the average health outcome per health dollar invested using a metric such as this to measure outcomes.

By 1976 the Fanshel and Bush list of 11 functional states had been elaborated into a 4-factor classification scheme termed the Index of Well-Being (IWB). (Patrick, Bush and Chen, 1973b; Patrick, Bush and Chen, 1973a; Kaplan, Bush and Berry, 1976) There were 3 dimensions concerning functioning: mobility, physical activity, social activity. These 3 factors generated 100 theoretical combinations, of which the authors had observed 43 in an observational sample of over 10,000 people. The fourth factor, “symptom/problem complex”, listed thirty-six different symptoms or problems a person might have. Together, the 43 observed combinations of the functioning factors and the list of symptoms and problems formed a descriptive system of health states. To score the health states the index used social utility weights elicited using a method of equivalence judgments from a probability sample of 867 people from San Diego, California. These people were presented with subsets of hypothetical health states drawn from the 43 x 36 potential health states defined by the IWB and gave systematically elicited numerical ratings for how undesirable was each health state. These ratings were analyzed by regression analysis using as independent variables a number of indicator variables representing levels of functioning on each the 3 functional factors and the 36 groups of symptoms and problems as additional indicator variable, processed their valuations into weights for the scale levels.

When the IWB regression function was rescaled using anchors of dead=0.0 and perfect health=1.0, then a person’s lifetime health could be computed in terms of well-years which integrated the person’s instantaneous well-being on the 0-to-1 scale over time. By the early 1980s the IWB was refined and evolved into the Quality of Well-Being index (QWB). (Kaplan and Bush, 1982) and well-years became quality-adjusted life years. (There is some dispute about who first used the term “quality-adjusted life year” and its acronym, QALY. It could have been the QWB developers in San Diego, it may have been researchers at Harvard (Weinstein and Stason, 1977), or it may have been Torrance, a Canadian health researcher.)

Building on his dissertation work about the same time as the IWB was being developed, Torrance also described a complete model for scaling health outcomes based on the concept of utility as formulated in economic decision theory. (Torrance, Thomas and Sackett, 1972) A utility scale is an interval scale which may be used to compute expected value of outcomes for the purpose of guiding decisions – decisions based on maximizing expected utility. It is generally accepted that the axiomatic method to elicit utility weights is to use what is called the “standard gamble.” The respondent is offered a

hypothetical choice between life in a fixed health state for a given period of time, or a two-outcome gamble whose outcomes are life for the same period of time in the best possible health state (Torrance defined this as good health, the absence of physical, mental and social disabilities and symptoms) or immediate death, the worst outcome. The gamble was offered with probability p of the best outcome and $1-p$ of the worst outcome; p was varied up and down until a probability, p^* is found where the respondent determines the two choices to be equivalent. Under suitable assumptions about independence of preference for time and health state, and linearity of preference for time, the utility of the fixed health state may be shown to be p^* .

In the operations research literature and the management science literature, Keeny and Raiffa published a landmark book in 1976, describing theory and fully operational procedures to assess utilities for outcomes involving multiple dimensions of value. The particular contribution of this work was to give analysts a method to decompose utilities for multidimensional outcomes into a series of one-dimensional utility functions which were much easier to elicit. Multiattribute utility theory (MAU) seemed tailor made for health indexes. A person's health state at any given time was multidimensional. If the overall utility of a health state could be decomposed as a simple function of utility for physical function, utility for mental function, utility for pain, etc., then a roadmap for developing a health utility index was clear. First, a health status classification scheme would be developed, expressing a person's health state as a set of levels on individual health dimensions. The set of dimensions would need to be comprehensive to encompass generic health. Then the MAU procedures could be used to assess a mathematical utility function by which the utility of the joint health state could be computed from utilities of the component dimensions.

Torrance, Boyle, and Horwood used this roadmap to construct the predecessor to the current Health Utility Indexes, the Health Utility Index, Mark I (HUI1). (Torrance, 1976) The index was directed to scaling health states for children aged 2 or older. Torrance, et al, used modified MAU techniques, employing the method of time tradeoffs Torrance had developed in his earlier paper to elicit utility functions from a sample of 128 parents of school children in Ontario, Canada. The health state classification scheme defined 4 dimensions. First was physical function with 6 levels ranging from level 1, no limitations, to level 6, "NEEDING HELP from someone else to get around the house, yard, neighborhood or community AND NOT being able to use or control the arms and legs." It similarly defined 5 levels of self-care and role activity, 4 levels of social-emotional function, and a dimension labeled "health problem", with 8 levels: level 1, "having no health problem", level 2, "having a minor physical deformity or disfigurement...", level 3, "needing a hearing aid", level 4, "having a medical problem which causes pain or discomfort for a few days in a row every two months", level 5, "needing to go to a special school because of trouble learning or remembering things", level 6, "having trouble seeing even when wearing glasses", level 7, "having trouble being understood by others", and level 8, "being blind OR deaf OR not able to speak."

After the HUI1, which is no longer used, two more extensive versions of the Health Utilities Index were developed, HUI Mark II (HUI2) and HUI Mark III (HUI3). HUI2

has 7 health attributes varying from 4 to 6 levels each, and HUI3 has 8 attributes with either 5 or 6 levels each.(Feeny, Furlong, Boyle and Torrance, 1995) Weights for these indexes were derived from assessments elicited from community population samples in Hamilton, Ontario.

Two more HRQoL indexes were developed about the same time. The first was developed by an international collaboration of European health researchers. The EuroQoL group was established in 1987 to design a simple yet comprehensive standardized non disease-specific measure that could be used throughout the European community to standardize health outcomes research.(www.euroqol.org) The result was the 5 dimension EQ-5D. The EQ-5D has 5 questions, each addressing a separate dimension of health, mobility, ability to do usual activities, pain, anxiety and depression, and self-care. Each question has 3 potential response categories, no problem with the dimension, some problems with the dimension, and extreme problems with the dimension. Five questions with 3 levels each yields a descriptive system with $3^5 = 243$ possible health states. The EQ-5D is available in more than 100 translated versions if one counts alternative administration modes as versions (face-to-face, self-complete, telephone interview, interactive voice response versions, and proxy versions). Population norms are now available for Armenia, Belgium, Canada, Finland, Germany, Greece, Hungary, Japan, Netherlands, New Zealand, Slovenia, Spain, Sweden, United Kingdom, and Zimbabwe through the EuroQol group. In 2005 utility weights for EQ-5D health states were derived for the United States, adding to over a dozen other nation's sets of weights that have been derived from population-based samples for the EQ-5D.(Brooks, Rabin and de Charro, 2003; Shaw, Johnson and Coons, 2005) Although often criticized for poor psychometric properties, the EQ-5D has persisted over 20 years and is very widely used today.

The last HRQoL index to be introduced in this section is the SF-6D. Brazier and colleagues, health economists at Sheffield University in England, noted that in spite of its widespread use as a health outcome measure the SF-36 was not useful for economic evaluation purposes because it was not a utility measure. It did not produce a utility valuation for the health states it defined.(Brazier, Usherwood, Harper and Thomas, 1998) They used 11 of the 36 questions in the SF-36 to define a 6-dimension health status description system. They reduced the 8 scales of the SF-36 into 6 dimensions by combining the two role-function scales into a single dimension of health-related restriction on role function, and they dropped the self-rated health dimension the SF-36 developers called "general health perceptions" because Brazier et al reasoned that health states should be assigned values by social preferences and not individuals experiencing the states.(Brazier, Roberts and Deverill, 2002) A sample of hypothetical health states was constructed using this 6 dimensional classification. These were presented to a population sample in England and each person assigned utilities to a subset of the health states. Regressing the utility assessments on the domain levels in the health states, Brazier et al constructed a scoring function mapping the 6-dimension health states into a 0-1 utility scale. The result was a summary utility scoring for SF-36 data, making it possible to compute QALYs as an outcome measure using the SF-36. Somewhat earlier Fryback et al (Fryback, Lawrence, Martin, Klein and Klein, 1997) had made a first step

this direction by regressing QWB scores for a community sample of older people on SF-36 data from the same people in Beaver Dam, Wisconsin. However there were many problems with the Beaver Dam equation for mapping SF-36 data to a utility scale, and the SF-6D was an immense step forward in this regard.

3. The case for a standardized HRQoL measure for population health

The previous section has detailed the nature and history of five leading HRQoL indexes. The past 40 years has been productive in this regard. Perhaps it has been too productive as we now have competing measures, each of which scales health somewhat differently. This section reviews several uses for a standardized summary HRQoL measure:

- Establishing HRQoL norms for the population;
- enabling computation of quality-adjusted life expectancy for the population and population subgroups, and
- use in longitudinal tracking of HRQoL in the population.

Current data sources for these purposes will be noted.

3.1. Norms for HRQoL in the US population.

As the term will be used here, a norm describes what is typical for a population. Norms, e.g. those for the SF-36v2TM indicated on the ordinate in Figure 2, are averages computed from representative population surveys—in fact the scales in the SF-36v2TM health profile are computed with norm-based scoring expressly to facilitate comparison to population norms. As was demonstrated above, the impact of a disease and treatments of a disease can be profiled by showing scores for untreated and treated patients.

Norms provide a point of comparison to understand how an individual or group differs from what is typical for the population. Because a summary HRQoL measure results in a single number representing social valuation of health on a scale anchored by dead=0 and full health=1, it allows evaluation of an individual's or group's health relative to the population norm: if the HRQoL score for the individual (or the group's average HRQoL score) is higher than the norm then he/she/they are on average more healthy than the comparison population; if the score is lower then they are less healthy. The magnitude of the difference may also be meaningful as for many of the HRQoL indexes a difference of approximately 0.03 is considered to be clinically meaningful.

It is important to note that comparing the individual or group to the population using health indicators on an indicator-by-indicator basis can determine whether they differ from the population but cannot serve to say whether on net their health is better or worse than what is typical for the population. The critical information added by the HRQoL index score comes from the societal utility or health preference assessments made during construction of the HRQoL index scoring function where people are asked to make tradeoffs among different domains of health function and symptoms when they assign values to health states.

The US has sporadic data to establish HRQoL norms in the population. Neither of the two large, recurring national population health surveys, the NHIS and the NHANES, directly includes a standardized preference-based summary HRQoL index in its questionnaires. However, two federally conducted surveys have included HRQoL instruments on an *ad hoc* basis. The Joint Canada United States Survey of Health (JCUSH), conducted by the US National Center for Health Statistics and Statistics Canada, used the HUI3 to survey Canadian and US adults in a one-time population survey in 2002-3 (<http://www.cdc.gov/nchs/nhis/jcush.htm>). The US Agency for Healthcare Research and Quality conducts the recurring Medical Expenditure Panel Survey (MEPS) on a subset of NHIS respondents. MEPS included the EQ-5D in its 2000 and 2001 surveys, and the SF-12 from 2000-present, although the version of the SF-12 used has changed in that time (<http://www.meps.ahrq.gov/mepsweb/>). Additionally, two research projects conducting US population surveys have included various HRQoL measures in the past 8 years. The US Valuation of the EuroQol EQ-5D (USVEQ) surveyed a population sample of 4048 adults using EQ-5D, HUI2, and HUI3.(Luo, Johnson, Shaw, Feeny and Coons, 2005) The US National Health Measurement Study (NHMS) collected EQ-5D, SF-6D, HUI2, HUI3, and QWB-SA in a national sample of 3844 older adults in 2005-6.(Fryback, Dunham, Palta, Hanmer, Buechner, Cherepanov, Herrington, Hays, Kaplan, Ganiats, Feeny and Kind, 2007)

These four surveys have been used to compute age and gender-based norms for HRQoL for the years and population segments for which the data were available.(Luo, Johnson, Shaw, Feeny and Coons, 2005; Fryback, Dunham, Palta, Hanmer, Buechner, Cherepanov, Herrington, Hays, Kaplan, Ganiats, Feeny and Kind, 2007; Hanmer, Hays and Fryback, 2007) However JCUSH, USVEQ, and NHMS were one-time studies and their data will become dated in the US. Statistics Canada is committed to ongoing administration of HUI3 in its Canadian Community Health Survey, a large on-going population survey. The US National Center for Health Statistics does not administer a HRQoL instrument in any of its on-going population surveys. The study directors for MEPS do not guarantee inclusion of either of the two HRQoL instruments in future surveys.

3.2. Computation of quality-adjusted life expectancy.

As briefly discussed earlier, a unique role of a summary HRQoL measure is to compute quality-adjusted life years, or QALYs. Figure 3 shows

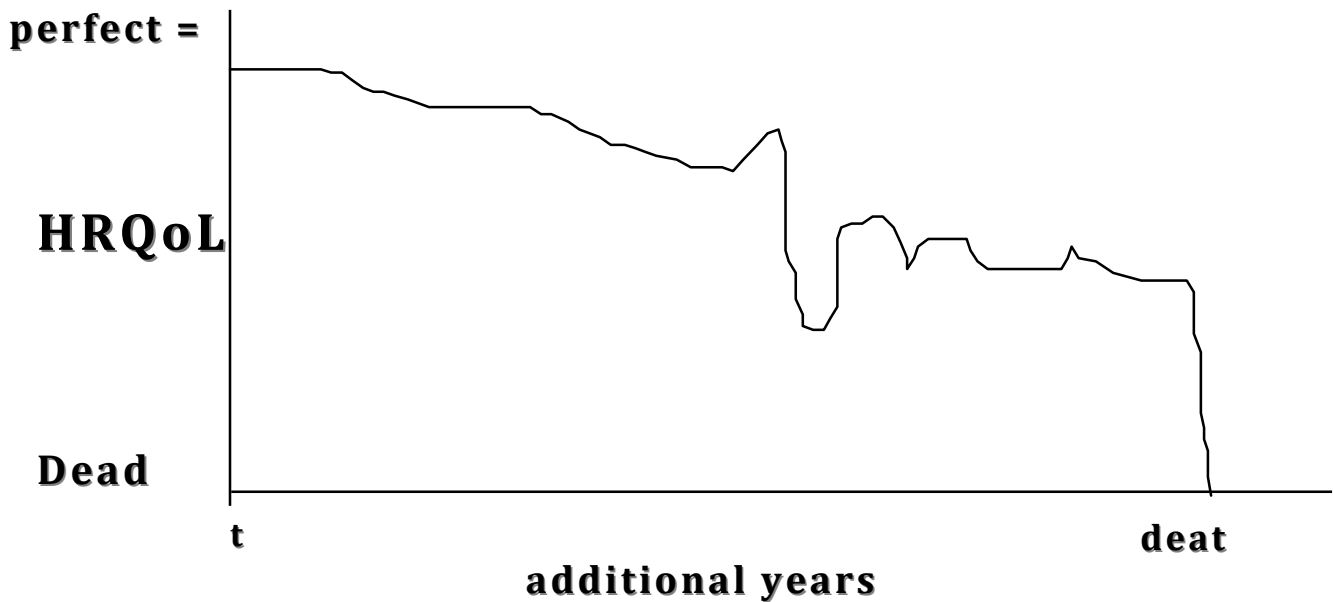


Figure 3. Health-related quality of life as a function of time over remaining life from time t_0 .

Fig. 3 represents a person's HRQoL as a function of time. Starting at time t_0 the person's HRQoL is very good, nearly 1, but slopes downward as minor health problems accumulate. Perhaps realizing this decline he began a biking program and suffered a serious injury soon after accounting for the drop in the center of the plot. Rehabilitation recovers some health, but leaves him on the same general slope of decline albeit with more variance than before. Death was relative swift when it came.

In retrospect for this individual the area under the curve—the integral of HRQoL over time—is the number of QALYs the person lived from time t_0 . Prospectively, conditioned on sex, age at t_0 , and health state at t_0 , the statistical expectation of QALYs lived is the person's quality-adjusted life expectancy (QALE). The health policy model proposed alongside the development of the QWB used average QALYs projected under differing health policies to guide those policies (Kaplan and Anderson, 1996). Expected QALYs are used as the numeraire to maximize in cost-effectiveness analysis of health interventions.(Weinstein and Stason, 1977; Gold, 1996)

We almost never have data to estimate QALE for a given individual. But as a population health measure, a number similar to QALE conditioned on sex and age can be computed using life tables and HRQoL from population surveys. In essence the average HRQoL reported by people at each year of age is used to weight a life-table survival function to get a weighted life expectancy at each year of age, an idea which has been developed and refined over time.(Sullivan, 1971; Erickson, Wilson and Shannon, 1995; Rosenberg, Fryback and Lawrence, 1997)

Although it is possible to generalize the concept of QALYs by using any summary health measure, there are strong reasons beyond the scope of this paper to support using a utility-theoretic, preference-based standardized HRQoL index for the computation if it is to guide health policy.(Gold, 1996; Gold, Stevenson and Fryback, 2002)

3.3. Longitudinal summary HRQoL data for the US population.

An important use of standardized measures is to track changes in a population over time. Changes in population health can be measured by changes in indicators. But without a summary HRQoL measure, we cannot say whether overall health is improving or not.

Each decade US health policy community, led by the Surgeon General and the Secretary of Health and Human Services, establishes 10-year national goals for health achievement. This priority-setting exercise was first completed in 1990, and titled *Healthy People 2000* (<http://odphp.osophs.dhhs.gov/pubs/HP2000/>). In 2000, new goals were set by *Healthy People 2010* (<http://www.healthypeople.gov/>). The exercise is nearly completed for *Healthy People 2020* (<http://www.healthypeople.gov/hp2020/Comments/default.asp>).

Goals for *Healthy People* are expressed in the form of target changes in many leading indicators for health. But with the first *Healthy People* the need for a summary measure of health was recognized since one goal of *Healthy People 2000* was to improve quality-adjusted life expectancy (QALE) in the US. An ad hoc HRQoL summary measure, the HALex (Health and Activity Limitation index), was constructed from data elements collected in the National Health Interview Survey for use in *Healthy People 2000* to compute “Years of Healthy Life” (YHL), a life-table based measure computed like QALE, but using the HALex instead of an existing HRQoL index.(Erickson, Wilson and Shannon, 1995; Erickson, 1998)

HALex has two components, self-assessed categorical health (“Would you say your health in general is excellent, very good, good, fair, or poor?”), and a six category activity limitation domain, ranging from no limitations in usual activities to unable to carry out activities of daily living (ADLs) for oneself (bathing, dressing, etc.). HALex thus has 30 “levels”, 5 categories of self-assessed health crossed with 6 levels of activity limitation, ranging from self-assessment “excellent” without any activity limitations to self-assessment “poor” with total limitation in ADLs. Each of the 30 levels was assigned a utility value using a multiplicative multiattribute utility function (Keeney and Raiffa, 1976) and using the HUI1 single attribute utility functions as correspondence functions.

Since the *Healthy People 2000* plan, the use of QALE as a measure appears to have fallen from favor. Instead of YHL, three other measures are substituted for *Healthy People 2010* and *Healthy People 2020*:

- *Expected years self-rated “good”, “very good”, or “excellent” health.* This is computed using life tables and a single self-rated health question which asks people to rate their overall health using a 5-category response scale: excellent, very good, good, fair, or poor. An indicator variable is constructed which is equal to 1 if the response is in the first three categories and otherwise is 0. This indicator is used to compute weighted life-expectancy using the same method as described for QALE above.
- *Expected years free of activity limitation.* The NHIS asks a series of questions to determine if a person’s health limits their role activities. Major social role activity can be work, school, homemaking as usually associated with the person’s particular age. Based on their responses each person is classified into one of six groups: (1) no limitations; (2) no limitations in major activity but limited in other activities; (3) limited in major activity; (4) unable to perform major activity; (5) unable to perform instrumental activities of daily living without help of another person; and (6) unable to perform self-care activities of daily living without help of another person. Similar to the indicator above, an indicator is constructed which is equal to 1 if the activity classification is at the first level, and 0 otherwise. This indicator variable is used to weight life table-computed life expectancies. This measure differs from the first in that people at the very lowest activity level may well self-rate their health as excellent or very good, and, vice versa, there are people without activity limitations who self-rate their health as only fair or poor. In the NHIS conducted in 1990, 83.2% of adults aged 18 and reported no activity limitations; 5.3% of these rated their health as fair or poor. In the NHMS, surveying adults aged 35-89 in 2005-6, the corresponding figures were 65.3% and 11.1%. So expected years without activity limitations is not identical to expected years in good to excellent health.
- *Expected years free of selected chronic diseases.* An indicator is constructed which is 0 if the person has one or more of arthritis, asthma, cancer, diabetes, heart disease, high blood pressure, kidney disease, or stroke, and otherwise is 1. This indicator, where “1” indicates absence of the selected diseases (*i.e.*, a year weighted “1” is a year free of these diseases) is used to compute weighted life expectancies.

In essence then, three indicators have been substituted for an overall measure of HRQoL by Healthy People. Representative results using these indicators are shown at (<http://www.healthypeople.gov/Data/midcourse/html/execsummary/Goal1.htm>).

Analysts who wish to compute QALE for the US population can compute HALex from annually collected NHIS data, then combine this with updated life tables as they become available to compute QALE for age- and sex-specific subgroups.

Because NHIS does not collect HRQoL measures other than HALex, QALE cannot be computed with any other HRQoL weighting using national data. Although HALex is an interesting measure of HRQoL, it is controversial for several reasons. First, it is the only measure to include self-rated health as a domain; all other HRQoL measures have explicitly excluded self-rated health under the presumption that this conflicts with using population preferences for health states and not the affected individual's rating. Second, it is not comprehensive in domain coverage as it has only activity limitations and self-rated health. Third, the weighting of health states in HALex was developed ad hoc and not derived by actual elicitation of health state utilities for the measure by a representative population sample.

Computation of QALE is also limited by available life tables which may not be available by year of age for sociodemographic or socioeconomic subpopulations of interest. Developing detailed life tables for such subpopulations is a challenge for national data systems.

4. Six competing measures; why hasn't the US selected one to use as a common measure in national data sets?

Six HRQoL measures— QWB, HUI2, HUI3, EQ-5D, SF-6D, and HALex—have been discussed above. Each is a candidate to standardize the measurement of HRQoL in US population data sets.

If all these indexes scaled health the same way there would not be a problem. But they do not scale health entirely the same. Differences between the HALex and the other 5 have just been mentioned. There are other differences too. All of the indexes use the same anchor points for their utility scales, with 1.0 representing full health and 0.0 being dead. But anchor points need not be *end*-points. Three indexes, HUI2, HUI3, and EQ-5D, identify some health states as worse than being dead and assign them utilities less than 0.0. The remaining two indexes, far from allowing states worse than dead, have minimum scores of 0.09 (QWB) and 0.3 (SF-6D).

It may seem from the scaling that the six indexes are incompatible. This seems not to be the case – although the scales differ on their faces, they may be related somewhat like Fahrenheit, Celcius, and Kelvin, albeit in a more complex fashion. Fryback, et al (Fryback, Palta, Cherepanov, Bolt and Kim, 2009), have demonstrated the six indexes order health states in approximately the same fashion, and are roughly related by two-part linear transformations and with flexibility regarding the definition of zero. However there is considerable statistical noise in these transformations, particularly for health states which are above average for the population. Principal components analysis shows about 70-75% common variance among the indexes. Unique variance for each index in a population sample appears to contain useful information; perhaps each index has better resolution in a particular range of overall health or measures some domain at better resolution or with less noise than other indexes do.

Even though each index may have strengths and weaknesses, there are strong advantages to deploying a single measure as a common measure across national data sets to allow computation of meaningful norms, to facilitate longitudinal tracking of population health across time, or to make comparisons across data sets. In Europe, Australia, and a number of other countries the EQ-5D is consistently used as a common measure in national data sets. In Canada the HUI3 is consistently collected in major national health surveys.

Why isn't a common measure being used in the US? All six measures are good measures, but not perfect measures. I believe the fact that no one of the measures is so obviously better than the others has left room for what I'll characterize as "politics" and "Politics."

4.1. Little "p" politics.

I include under this heading the competing interests behind the various indexes.

- *Competing index developers innovating and promoting their work.* It takes a considerable amount of work – sometimes an entire career – to develop and promulgate a HRQoL index. It is no wonder that the teams of researchers who have been identified with different indexes wish to see their work take prominence.

An index does not get picked up and used simply because it exists. I am convinced there is a cycle of innovation with these instruments just as there is for any set of competing innovations. The cycle involves a long prodromal stage of development and early use of the instrument with a few publications appearing. For a newcomer index to be adopted by researchers it must offer some advantage over existing indexes.

The SF-36 has come to dominate the health status profile field. Part of its success is due to its innovations compared to extant measures when it was introduced—it was shorter, its developers devised an attractive format for self-completed forms used to administer the index, and they promised norms against which results with the SF-36 results for individual patients could be compared. They worked hard to develop a user community among health institutions.

The EuroQoL group which developed the EQ-5D wanted a minimal response burden instrument to standardize health outcomes measurement. Their 5-question index along with an auxiliary self-rating filled this bill exceptionally well. The SF-36 was not available in translation for European use until later, so the EQ-5D gained the early lead. It was easy to translate and required less psychometric validation than the SF-36 did

for the European “market.”

The developers of the HUI2 and HUI3 took a lesson from the SF-36. They worked to develop relatively brief questionnaires to reduce response burden and developed a web site to present work using the HUI2 and HUI3.

The QWB developers, although early on the scene with a clear field for a number of years, found their instrument, which took 12-15 minutes to administer by a trained interviewer, was at a disadvantage when a number of self-completed and less burdensome instruments appeared. Use of the QWB began to drop off. The developers then produced a self-administered version of the index, the QWB-SA (Andresen, Rothenberg and Kaplan, 1998), and this is seeing some use now although not often beyond the development collaborators.

It seems clear that without a base of reports in the literature, a set of collaborators who might serve as formal or informal sources of expertise to users, and an instrument having relatively little response burden, an index does not get used much if at all.

Developers have pride in their progeny. They often see their index as the best available (why else to commit so much time and energy to it?). So they tend to promote their work over others’.

- *Proprietary interests.* There is a long history of proprietary interest in the field of psychological testing and for many disease-specific health scales. Developers charge for use of the index and users pay either a per-study or per-respondent license fee. The fee may be nominal or substantial.

The SF-36, now the SF-36v2™, initially was available at no charge for research use, and there was a licensing fee for use by health care providers. QualityMetrics made money from selling mark-sense administration forms and providing scoring services to hospitals, clinics, and HMOs who administered the SF-36. Now there is a licensing fee to all users. Users are able to purchase ready-made forms and scanning services for scoring, or they may make their own forms and create their own statistical programs to score the health profile scales. The SF-6D is available for use as part of this licensing, or is licensed separately and is still at no cost to academic users.

The HUI2 and HUI3 domains and scoring algorithms are published in the open literature. However the specific questionnaires completed by respondents and the algorithms which map questionnaire responses onto the public domain variables are maintained as confidential, licensed products. Researchers are charged \$4000 to \$6000 per study to use the

questionnaires.

The QWB and QWB-SA are available under a no-cost licensing agreement to any entity which wishes to use them. The EQ-5D is also available simply through registering at the EuroQol web site. The HALex also is openly published and can be used by anyone.

There are serious concerns about using proprietary instruments in US national data sets. If a proprietary instrument is selected as the common metric across data sets a huge competitive advantage, and perhaps monetary gain, is ensured to the selected instrument and its proprietors. This has been a barrier to selecting a single instrument.

- *Competition for space on national protocols.* The large US national studies, NHIS, NHANES, MEPS, and others, involve long protocols. Many investigators have contributed questions to these protocols. Because of the length, the response burden is large on respondents. There is a tension between the study administrators who wish to keep response burden low, and investigators who wish to have as many questions as possible.

Because standardized summary HRQoL indexes are intact instruments, if they are included in a national survey protocol the entire questionnaire must be included. It is not possible to include only a subset of the questions – it must be all or none. This seems to “break the bank” with the survey administrators who routinely ask investigators to limit addition to the protocol to a few questions at most. The SF-36 has 36 questions; the HUI2/3 has 18 questions for self-administration; the QWB-SA has over 50 questions to complete; the EQ-5D has 5 questions.

It has also been my experience that survey administrators want to change the questions to improve the wording (from the administrators’ point of view). Critiques of all the instruments have been published and every instrument has questions that could be improved (e.g., see (Mallinson, 1998)). But changing the questions will result in a questionnaire that is not comparable to the “official” version.

Given these problems, only the shortest of instruments, EQ-5D and SF-12 with 5 and 12 questions, respectively, have made it onto a national survey in the US, the MEPS, and then without commitment to administration in future cycles of the MEPS. It was only with intense lobbying that MEPS administrators committed to this much. My personal hope is that success in seeing these data used for US national norms (e.g., see (Sullivan, Lawrence and Ghushchyan, 2005; Hanmer and Fryback, 2006; Sullivan and Ghushchyan, 2006)) will encourage continuing commitment to include these instruments periodically in the future as a matter of routine

to allow longitudinal tracking of population health using MEPS data.

MEPS is just 14 years old. As few as five or six years ago it was underutilized even by the health economics and policy community for which it was first targeted. Use by that sector has increased exponentially as the research community has become familiar with MEPS and better tools and protocols to access the data were developed (especially protocols to more easily deal with the confidential, person-level data in MEPS). Researchers interested in HRQoL are beginning to discover MEPS as a source of interesting longitudinal data to relate health care and changes in HRQoL and the near future should see much more use of the data to address interesting question in this regard.

- *Concern about subjectivity of the measures.* Privately, officials with the US National Center for Health Statistics have voiced objections concerning use of preference-based HRQoL measures. The concern is that the scoring is a function of subjective weighting of health domains. NCHS prides itself on objectivity of its data and reporting a measure which combines health domains “using subjective weights” conflicts with that self-perception. The perception that HRQoL measures incorporate arbitrary subjective judgment is a barrier to having them collected by NCHS.

4.2. Big “P” Politics

There are larger, more overtly political issues which mitigate against inclusion of a common HRQoL instrument in national surveys.

- *National Chauvinism in scoring.* Preference-based HRQoL indexes are scored by algorithms which predict average utilities assigned to health states by a sample of people from a population. Early on these were community samples, such as the 800 people drawn from San Diego, California, and surrounds used for the QWB scoring algorithm in the early 1970s. The HUI2 and HUI3 are weighted based on samples of people from Ontario, Canada. With the EQ-5D development, it became the norm to use a national population sample and a systematic study was done to derive population-based weights for the United Kingdom. Later, exactly the same experimental protocol was followed to derive national EQ-5D weights for many other countries so that each could show that the EQ-5D results were applicable to their own national health policies. For cross-national comparison of results the UK scoring was generally used.

In 2005 the US Agency for Healthcare Research and Quality funded a several million dollar project to systematically sample adults to represent

the US population and to elicit EQ-5D weights from them so that the US would have its own set of weights for data pertaining to US health policy (Fryback, 2005; Shaw, Johnson and Coons, 2005).

Other competing indexes are not so lucky. The HUI is still weighted according to the sample from Hamilton, Ontario, Canada. The SF-6D is weighted using a population sample from the UK. The QWB-SA has US roots, but is still based on a community sample from San Diego. And the HALex is ad hoc, not being based on any sample but weighted by correspondence to both QWB and HUI.

US national statistical agencies are reluctant to include instruments that are not based on US scoring – especially if inclusion is for a long-term commitment to use the instrument longitudinally for health policy.

- *Endorsement by a federal agency.* As mentioned briefly above, if a US agency were to adopt one measure for an extended committed time, this adoption could be construed as an endorsement of one index over the others. It would certainly be a *de facto* endorsement if not *de jure*.

Unless there is clearly proof that one index is superior to the others, it is unlikely that one will be selected for long-term use in a US national data set, especially if proprietary interests mean that some gain and some lose financially as a consequence.

- *Are there vested interests in not having a common metric?* Many groups advocate for health research dollars. There are advocacy groups such as the American Cancer Society, the American Heart Association, and the American Diabetes Association to name but a few that come to mind quickly. These groups have vested interests in seeing federal research dollars go to a specific disease or set of diseases, and they solicit funds from the public based in part on the public's perception of how serious that disease is. Most of the institutes of the National Institutes of Health are organized around a small set of diseases or conditions or an organ system. The institutes compete with each other for congressionally mandated dollars and for allocations of NIH discretionary budget by the governing bodies within NIH.

The fact that these organizations are largely attached to diseases or organ systems means that the aspects of health in which they are most interested are usually measured most precisely by disease-specific or organ-specific measures. If generic measures are used in national data systems it is possible that the common metric could create winners and losers among them in the advocacy competition for public dollars. Generally these organizations have advocated for measures of most interest to their own

disease and calls for common metrics which cross diseases have not found a responsive audience with them.

5. Success stories

In spite of the political and Political motivations serving as barriers to adopting a common HRQoL metric in the US, there are success stories in this regard elsewhere and in the US.

5.1. EuroQol and the EQ-5D.

The story of the development of the EQ-5D is one success story. This is not a story of adoption of an existing index, however. The EuroQol group was formed with the explicit purpose of developing a common metric for HRQoL to be used in the European Community.(Brooks, Rabin and de Charro, 2003) A collaborating international group of researchers obtained funding to pursue this goal.

The EQ-5D has some serious psychometric problems, particularly with respect to a large ceiling effect in relatively healthy populations. Further, the 3-category responses for the 5 questions are often deemed to have too little resolution to be meaningful measures at an individual level. In spite of these limitations, the EuroQol group has found the measure to be useful and it has been widely adopted around the world. As of the end of 2009, eight more language versions were made available including Chinese, Vietnamese and Japanese for speakers of those languages in Australia, Macedonian for Macedonia, Russian for Lithuania, Setswana for Bostwana, Swahili for Tanzania, and Farsi for Iran. This brings the total number of translations to over 100. Weighting algorithms have been completed for population samples in 13 different countries.(Szende, Oppe and Devlin, 2007) The EQ-5D is widely used across the European Community to compare health results from one nation to another and for tracking health with national data systems within countries.

5.1. Statistics Canada surveys.

Every two years Statistics Canada surveys the health of some 130,000 Canadians in a population sample termed the Canada Community Health Survey aimed at understanding health and changes in health at the level of some 130 communities throughout Canada. One of the regularly collected measures is the Health Utilities Index. Although some of the questions on the interviewer-based form used by Statistics Canada interviewers differ somewhat from the proprietary questions distributed by the HUI developers, the HUI3 may be scored from a respondent's answers. Based on these data, some interesting reports are now beginning to appear. Orpana et al report 10-year results for a longitudinal cohort of 7000 people.(Orpana, Ross, Feeny, McFarland, Bernier and Kaplan, 2009) Taking account of the impacts of institutionalization and death, this study described the normative trajectories of health-related quality of life in Canada as individuals age from mid-to late life. To my knowledge this is the first report of its kind

with a summary HRQoL index. Using the same data, Kaplan et al replicated an often found result that self-rated health (excellent, very good, good, fair, poor) can predict mortality in time frames up to 10 years. (Kaplan, Berthelot, Feeny, McFarland, Khan and Orpana, 2007) They also showed that the HUI3 is predictive of mortality. Alone, the HUI3 is not as strong a predictor as categorical self-rated health, but it contributes independent information to predicting mortality above and beyond self-rated health.

Documenting and understanding the relationship of HRQoL to mortality is a novel area of research not before possible other than using the categorical self-rated health question.

Since the early 1990s, Statistics Canada has also conducted the National Population Health Survey (NPHS), a representative biennial survey of Canada's adult population with nearly 20,000 people sampled. NPHS also collects the HUI3 on a regular basis. Both NPHS and CCHS have longitudinal subsamples, and both have contributed substantial knowledge about associations of HRQoL and various sociodemographic factors as well as geographical distribution of HRQoL in Canada.

5.2. Medical Expenditure Panel Survey (MEPS).

In several places above the inclusion of EQ-5D and SF-12 in the MEPS has been mentioned. This has provided a data set to quantify impact of many health conditions and diseases on HRQoL in the US population. (Sullivan, Lawrence and Ghushchyan, 2005; Sullivan and Ghushchyan, 2006)

5.3. Medicare Health Outcome Study.

The Centers for Medicare and Medicaid Services (CMS) has been tasked by the US Congress to assure that Medicare beneficiaries who are in Medicare Advantage (MA) plans (these are health maintenance organizations contracted to Medicare) do not suffer loss of quality of life because of this insurance arrangement. Accordingly the Health Outcome Study was organized and required by Congress to sample 1000 beneficiaries from every MA plan each year and to administer a comprehensive questionnaire about their health status. As part of this protocol the SF-36 was administered, and from the SF-36 the SF-6D HRQoL index score can be computed.

Because many plans have few Medicare beneficiaries, the requirement to sample 1000 people per year results in an opportunistic longitudinal panel. These data were linked with the National Cancer Institute's SEER (cancer Surveillance, Epidemiology and End Results) dataset. Longitudinal HRQoL trajectories were examined for HOS study participants who initially did not have cancer and later developed cancer. This is one of the very few before and after data sets allowing study of the impact of cancer on quality of life within the same individuals. (Ambs, Warren, Bellizzi, Topor, Haffer and Clauser, 2008; Clauser, Arora, Bellizzi, Haffer, Topor and Hays, 2008; Clauser and Haffer, 2008; Reeve, Potosky, Smith, Han, Hays, Davis, Arora, Haffer and Clauser,

2009) It was possible only because the HOS administered the SF-36 consistently over nearly a decade.

In 2004, the Medicare HOS changed its instrumentation from the original SF-36 to the Veterans SF-36, a revision to the SF-36 developed for the Department of Veterans Affairs health system. (Kazis, Lee, Spiro, Rogers, Ren, Miller, Selim, Hamed and Haffer, 2004) The Veterans SF-36 is comparable to the original in 6 of the 8 scales, with the 2 role-functioning scales “distinctly different” in response options. Based on examination of the content, it appears possible to score the SF-6D from data collected with the Veterans versions of the SF-36 and SF-12. So the ability to track SF-6D scores over time in the HOS should continue in spite of the change in instrumentation.

5.4. *A note about health measurement at the state and sub-state levels*

Recently there has been interest in comparing health across regions, states, and even counties in the US. Now in its 20th year, *America’s Health Rankings*TM, published in 2009 as joint effort of United Health Foundation, the American Public Health Association, and Partnership for Prevention, ranks the 50 states on 22 different measures of health and health determinants (<http://www.americashealthrankings.org/>). Most of these measures are indicators such as the percent of the population who are obese, the smoking rate, percent of the population without health insurance, and the high school graduation rate (more education is a known correlate of better health). The publishers have made an effort to make a summary measure by weighting indicators into a single number and this is used to create a ranking of the 50 states. *America’s Health Rankings*TM picks its indicators in part based on their consistent availability at the state level. If these measures were not available there could be no meaningful comparison of the rankings from one time to another.

State of the USA, Inc. (SUSA) is a non-profit organization established in 2007 to assess and communicate what its name implies (<http://www.stateoftheusa.org/ourwork/introduction.asp>) as high quality and high reliability information about the state of the US as a country and society. In 2008 SUSA commissioned the Institute of Medicine of the National Academies (IOM) to recommend a list of the most important 20 indicators of health by which the state of health in the US could be assessed and tracked. An IOM committee has devised a list (IOM, 2009). Based on this list, SUSA will also track health of the US and communicate this to the public. One of the criteria for the list was availability of data that “...can be broken down by important population subgroups (e.g., age, gender, socioeconomic status [SES], race/ethnicity, and *geographic region (states, cities, communities)*...” [IOM, 2009 #2670; page 2, emphasis added].

Finally, there are beginning to be efforts to track population health across time at the sub-state level. In the state of Wisconsin, the University of Wisconsin-Madison Population Health Institute publishes an annual ranking of 73 regions (72 counties and 1 metropolitan area) in Wisconsin that employs a set of over 30 health indicators very similar to those in both *America’s Health Rankings*TM and the SUSA health data

(<http://uwphi.pophealth.wisc.edu/pha/wchr.htm>). The UW Population Health Institute model is being expanded by researchers at the UW Population Health Institute, with funding by The Robert Wood Johnson Foundation, to cover all counties in the US. Its initial report is scheduled to be released in February, 2010 (<http://www.countyhealthrankings.org/index.html>).

These efforts to conduct population health comparisons and rankings at the state and sub-state levels draw mostly on the only health data that are routinely collected at these small-area levels. This source is the Behavioral Risk Factor Surveillance System (BRFSS) of the US Centers for Disease Control and Prevention (CDC). CDC funds states to collect a core set of health indicators in a standardized fashion year after year. States may elect to collect additional data of their own choosing, but it is the core data that provide consistent year-to-year data for the projects which seek to rank and compare small areas within the US over time.

Unfortunately, none of the preference-based HRQoL measures is collected by BRFSS. The CDC did originate a different measure of HRQoL, a four-component measure called the CDC HRQOL-4. The four items in this measure have been administered as part of the core BRFSS data protocol since 1993. These questions ask:

- Would you say that in general your health is excellent, very good, good, fair, or poor?
- Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?
- Now thinking about your mental health, which includes stress, depression, and problems with emotions, how many days during the past 30 days was your mental health not good?
- During the past 30 days, for about how many days did poor physical or mental health keep you from doing your usual activities such as self-care, work or recreation?

The two questions concerning physical and mental “not-good” days are combined into a summary measure by adding the two answers and truncating the sum at 30 if it is larger (<http://www.cdc.gov/hrqol/methods.htm>). Jia et al (Jia and Lubetkin, 2008) used BRFSS and MEPS data to develop a mapping from the Healthy Days measures to EQ-5D scores. Because they did not have a data set containing both the CDC measures and EQ-5D for the same people, they matched cumulative frequencies in the population for the measures to create a cross-walk between them. In work in progress, Jia is using NHMS data to update this cross-walk since all the measures are collected for the same respondents in NHMS.

The state and county ranking reports are quite influential in motivating health authorities at a local level. Rankings seem to mobilize the American psyche and these projects make the most of it. The UW Population Health Institute has developed a model for this based on experience with the Wisconsin county rankings, and now use this

model—MATCH: Mobilizing Action Toward Community Health—as the framework for their RWJF funded project to rank all counties in the US (<http://uwphi.pophealth.wisc.edu/pha/match.htm>).

Key to all these efforts is availability of routinely collected data to form standardized indicators of population health.

6. New chapter in US health measurement: PROMIS – the Patient-reported outcomes Measurement Information System.

6.1. Standardized measurement using item response theory scales of health—PROMIS.

In 2004 the National Institutes of Health (NIH) began the NIH Roadmap for Medical Research, a plan to “implement new pathways for medical research in the 21st century” (<http://www.nihroadmap.nih.gov/aboutroadmap.asp>). As part of the roadmap a program of research was begun to standardize measurement of patient-reported health outcomes for use in medical research. The Patient Reported Outcomes Measurement Information System (PROMIS) initiative was funded. It comprised six primary research sites and a statistical coordinating center. PROMIS investigators developed a health domain framework by which to organize their work, and set out to develop item banks to measure each domain as well as an internet-based delivery interface by which computerized adaptive testing (CAT) could be used to administer items for each domain. Based on success of the first round of funding, the grant supporting the PROMIS consortium was set for a new round of competitive funding in 2009.

The PROMIS domain framework is hierarchical and guided at the highest level by the World Health Organization’s tripartite model, and represents overall self-reported health as comprised of physical health, mental health, and social health.(Cella, Yount, Rothrock, Gershon, Cook, Reeve, Ader, Fries, Bruce and Rose, 2007) In the PROMIS framework physical health is subdivided at the next level down the hierarchy into physical function, symptoms, sleep/wake function, and sexual function. Mental health is subdivided into emotional distress, cognitive function, and positive psychological function. Social health is comprised at the next level of social function and social relationships. In the PROMIS framework many of the second-level subdomains are further subdivided into subdomains to a depth of five levels in the domain hierarchy (counting overall self-reported health as the first level).

The mathematical foundation for the measurement model underlying PROMIS is quite different from that for the HRQoL indexes described above. Where the HRQoL index scales are developed within an econometric/preference-based model, PROMIS is built on the psychometric measurement model known as item response theory (IRT). In PROMIS each domain or subdomain is conceived of as a latent health dimension. Each latent dimension has an associated item bank – a pool of questions about function or experience on that dimension. The questions in the item bank sample degree or levels of functioning across the dimension. For example, the physical function item bank currently

has 125 items, each beginning with the stem “Are you able to...” followed by some aspect of physical function such as “...reach into a high cupboard”, “...exercise for an hour”, or “trim your fingernails.” A five category Likert response scale is used for all items, varying from “without any difficulty” to “unable to do.” Items in the item bank have been previously evaluated using IRT analysis and a cumulative probability function for the 5 categorical responses to each item has been estimated as a function of underlying “true” scale score. A respondent is administered a selection of items and based on their pattern of responses to the items an estimated scale score for the respondent’s location on the latent domain is computed using the IRT parameters.(Reeve, Hays, Bjorner, Cook, Crane, Teresi, Thissen, Revicki, Weiss, Hambleton, Liu, Gershon, Reise, Lai and Cella, 2007) Computerized adaptive testing (CAT) is a method within the IRT paradigm to compute scores sequentially and to modify sequencing of items as a person responds in order to minimize the number of items to which the person responds to reach a preset level of precision in estimating that person’s “true” score on the domain.

PROMIS as now in process of being implemented is best described as a generic health status profile, much like the SF-36. As conceived, a respondent using an internet application would be administered CAT-based tests for each of a spanning set of health domains in the PROMIS framework to obtain a scale score for each domain. This set of scores profiles that person’s generic self-reported health. The computer interface is being actively engineered by PROMIS researchers to accommodate a range of disabilities.

6.2. One problem is solved.

The 6 HRQoL indexes are each associated with a fixed questionnaire (although the HUI2/3 questionnaire differs for self-administration versus interviewer administration). Technically, changing the questionnaire can affect how people answer the questions and thus affect the manner in which people are assigned scores. The EQ-5D is often criticized for having only three-category responses and being particularly insensitive at the top of the scale.(Insinga and Fryback, 2003) The SF-36 (and thereby the SF-6D) is criticized for not describing very sick people (Blanchard, Feeny, Mahon, Bourne, Rorabeck, Stitt and Webster-Bogaert, 2004; Brazier and Roberts, 2004) and for having poor or confusing response choices for elderly people.(Mallinson, 1998; Mallinson, 2002) But, changing any of these questionnaires to address such problems may disconnect legacy data collected with the instruments in the past from data collected in the future with new, improved questionnaires.

So the researchers using the existing HRQoL indexes fight the constant tension between a desire to remain fixed to allow comparisons to legacy data and the desire to improve the questionnaires.

PROMIS has overcome this concern. IRT lets us swap out items to update and improve scales without losing backwards comparability. A major strength of IRT is that

the scales for the domains are not defined by a fixed set of questions—the questions *per se* are not the scale as with classical measurement theory.(Nunnally and Bernstein, 1994) The questions sample from the underlying domain and IRT can be used to place each question's responses along that domain. A new question can be integrated with existing questions, and an existing question or its responses can be improved. IRT allows the constant updating of the item bank of questions without losing the underlying domain scaling. Indeed PROMIS is committed to continually improving the item banks.

What needs to happen to be able to apply PROMIS-like updating to HRQoL measurement?

6.3. PROMIS and preference-based HRQoL scoring.

The PROMIS measurement system describes health states in a multidimensional space of latent continua underlying the health domains. An individual's health state is described as a vector of estimated scores for these latent dimensions. Given that the dimensions span the domain hierarchy – *i.e.*, they cover the concept of self-reported health with no path down the hierarchy unmeasured– then in principle we could use econometric elicitation techniques to have people assign utilities to this multidimensional structure and derive a preference-based scoring function as a summary HRQoL score for self-reported health.

This is the process that was used to develop scoring functions for EQ-5D, SF-6D, and QWB. The SF-6D development is particularly germane. Brazier and colleagues used the questions of the SF-36 to define explicit ordinal scales for 6 different domains of health as covered by the SF-36 questionnaire. The scales were physical functioning, role limitations, social functioning, pain, mental health, and vitality, with 6, 4, 5, 6, 5, and 5 levels defined, respectively, for a total possible space with $6 \times 4 \times 5 \times 6 \times 5 \times 5 = 18,000$ possible discrete health states. A strategic sample of 249 of these states was constructed and each of 836 people sampled from the general population ranked then assigned utilities to a subset of 6 of these states. Mathematical modeling of the responses allowed construction of a utility function covering the entire space.(Brazier, Roberts and Deverill, 2002)

The HUI2 and HUI3 scoring functions were developed using multiattribute utility assessment techniques which are more structured, but not entirely dissimilar in approach.(Feeny, Furlong and Barr, 1998; Feeny, Furlong, Torrance, Goldsmith, Zhu, DePauw, Denton and Boyle, 2002)

This process is complicated for PROMIS by the fact that the domain variables are abstract latent scores. How do we do utility assessment of the latent domains covered by PROMIS? All of the HRQoL indexes, and indeed multiattribute utility assessment in other application areas (Keeney and Raiffa, 1976) exploit the fact that domains are defined by explicitly described scales, either categorical or continuous. I've never seen comparable assessment done for latent constructs that do not have a fixed semantic scale or objective scale to work from.

Deriving a summary preference-based HRQoL scoring function for the PROMIS health profile thus presents an interesting applied psychometric problem. IRT results in scales referenced to the population, with mean zero and standard deviation 1 in the population of reference. But a particular score, say -1.3 on a latent scale of physical function, would be an unfamiliar abstraction to a subject in a valuation study who was asked to assign utilities to a vector of such scores. Researchers setting out to do a valuation experiment for PROMIS to parallel that done for SF-6D would need to find a way to make these scores much more concrete and realistic for respondents.

This is a difficult problem, but is not insurmountable in my opinion. With the PROMIS latent scales made more concrete for presentation to subjects, a valuation experiment is possible. This effort could also apply to the problem of making PROMIS scales more concrete for physicians and researchers who wish to use them to track health status of patients.

PROMIS is not ready for this experiment yet. The item banks that exist at present do not completely cover the domain structure – there are paths down the hierarchy without item banks if the PROMIS website is current (<http://www.nihpromis.org/Web%20Pages/Domain%20Framework.aspx> ; accessed Feb. 1, 2010). For example, work is still in progress developing an item bank for cognitive function, a major component of mental health; for social isolation, a component of social health; and for sexual function, a component of physical health. But progress is expected and a summary HRQoL scoring function should be possible in the next few years.

6.4 The promise of PROMIS

At this writing, PROMIS still has a long way to go. The researchers have delivered a dozen item banks so far and are at work on others. A viable internet-based CAT administration interface has been developed.

At NIH, PROMIS is seen as an adjuvant to clinical research. It will help standardize patient-reported outcomes measurement for clinical research as NIH moves to bring research “from the lab to the bedside.” Accordingly, much of the focus of PROMIS has been on the “lower” part of the health spectrum to enable measuring impact of treatments on patients with disease and disability.

I see at least part of the future of PROMIS to be in population health measurement as well as in clinical research and treatment. It is fortunate that IRT allows improving item banks with questions added to increase precision of measurement in the better health part of the health spectrum. This will allow tracking changes in good health as well as poor health in the population.

The younger generations in our population are now growing up with easy access to, even dependence on, the internet. With PROMIS available in CAT format administered via the internet it seems inevitable that PROMIS is poised to be a central

feature in population health assessment in the future. Given the statistical and practical difficulties of representative population sampling via the internet, this future may not be immediately realized. But it is coming and one can almost see the day when efforts such as the state and county rankings described in section 5.4 will be supported by internet-based collection of HRQoL data.

7. Conclusion

In 1996, the US Panel on Cost-Effectiveness in Health and Medicine called for use of preference-based HRQoL measures to compute changes in QALYs experienced by patients as the standard outcome measure for cost-effectiveness analyses to guide health policy (Gold, Patrick, Torrance, Fryback, Hadorn, Kamlet, Daniels and Weinstein, 1996). In spite of a large and growing body of analyses conforming to this standard, cost-effectiveness analyses are rarely acknowledged in the US to guide health policy decisions. Although such analyses find little favor in the US public where health care rationing is an epithet, they have gained ground in the United Kingdom, Canada, and Australia.(Neumann and Greenberg, 2009) Aside from the political problems encountered when thinking about limiting health care access and utilization that might be unique to the US, I think the research community has lost out by not picking one or two HRQoL measures and pushing to have them collected regularly in the population at the national, state, and local levels. We have been hamstrung by the “little p” and “big P” politics of making such a choice. The success stories for health-related quality of life measurement have come from applying a small set of measures consistently over a long span of time at as local a level as possible. Canada probably leads the way in this regard as their use of the same measure over time in large-scale population and cohort studies has led to insights about how HRQoL evolves over the course of a lifetime and has helped begin sorting out the impacts .

The value of HRQoL measures comes from consistent use over time and in different populations, not from their perfection as measures of HRQoL. As my colleagues and I have recently shown, the scales for these measures are very different and cannot even agree about where the health state “dead” is located, but they *do* order people’s health in a very similar fashion and are approximately linearly related.(Fryback, Palta, Cherepanov, Bolt and Kim, 2009) In short, the indexes seem to do the same job in about the same way. But researchers continue to quarrel about which has the “correct” scale. Because I was not involved in development of any of these measures and because I have advocated simultaneous use of as many as possible of the measures at the same time in the same study, I have generally been regarded as a neutral in debates advocating one or another of these measures.

However I think evidence has now accumulated showing that, at least for the purposes of cost-effectiveness analysis of medical practices and for use in large scale population surveys, it is now time that either one of the measures be anointed as winner, or an improved measure developed (Fryback, Palta, Cherepanov, Bolt and Kim, 2009).

Given the politics weighing against picking one as winner, there will probably have to be a push to develop a one final HRQoL measure that will take the best of all of them. I think that the path I've sketched here for deriving a preference-based HRQoL summary measure from PROMIS's suite of domain-by-domain generic measures may well be the path forward. It offers a relatively small response burden as it is administered by computerized adaptive testing. It promises relatively high precision in measuring the individual domains. It can be administered widely, and presumably inexpensively over the internet and thus we could afford to administer it with good sample size at the local level. It results in a measure where the questions that locate a respondent on a domain variable can be continually improved without needing to amend the summary scoring function. And, using analytic techniques that my colleagues and I have demonstrated (Fryback, Palta, Cherepanov, Bolt and Kim, 2009), cross-walks between the new measure and current HRQoL measures could be developed so that legacy data are not entirely left behind in the transition.

Can it happen? I don't know. Perhaps hanging our hats on PROMIS is putting too many eggs in one basket. Most researchers involved in the PROMIS consortium do not come from the preference-based HRQoL tradition. And most researchers who identify with the HRQoL measures are not familiar with PROMIS. But crossover is beginning to happen, and I think it just might work out. So: fingers crossed.

References

- Ambs A, Warren JL, Bellizzi KM, Topor M, Haffer SC and Clauser SB. Overview of the SEER--Medicare Health Outcomes Survey linked dataset. *Health Care Financ Rev* 2008; 29(4) 5-21.
- Andresen EM, Rothenberg BM and Kaplan RM. Performance of a self-administered mailed version of the Quality of Well-Being (QWB-SA) questionnaire among older adults. *Med Care* 1998; 36(9) 1349-60.
- Bergner M, Bobbitt RA, Carter WB and Gilson BS. The Sickness Impact Profile: development and final revision of a health status measure. *Med Care* 1981; 19(8) 787-805.
- Blanchard C, Feeny D, Mahon J, Bourne R, Rorabeck C, Stitt L and Webster-Bogaert S. Is the Health Utilities Index valid in total hip arthroplasty patients? *Quality of Life Research* 2004; 13(2) 339-348.
- Brazier J, Roberts J and Deverill M. The estimation of a preference-based measure of health from the SF-36. *J Health Econ* 2002; 21(2) 271-92.
- Brazier J, Usherwood T, Harper R and Thomas K. Deriving a preference-based single index from the UK SF-36 Health Survey. *J Clin Epidemiol* 1998; 51(11) 1115-28.
- Brazier JE and Roberts J. The estimation of a preference-based measure of health from the SF-12. *Med Care* 2004; 42(9) 851-9.

- Brooks R, Rabin R and de Charro F. *The Measurement and Valuation of Health Status using EQ-5D: A European Perspective*. Dordrecht, The Netherlands: Kluwer Academic Publishers; 2003.
- Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, Ader D, Fries JF, Bruce B and Rose M. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. *Med Care* 2007; 45(5 Suppl 1) S3-S11.
- Clauser SB, Arora NK, Bellizzi KM, Haffer SC, Topor M and Hays RD. Disparities in HRQOL of cancer survivors and non-cancer managed care enrollees. *Health Care Financ Rev* 2008; 29(4) 23-40.
- Clauser SB and Haffer SC. SEER-MHOS: a new federal collaboration on cancer outcomes research. *Health Care Financ Rev* 2008; 29(4) 1-4.
- Erickson P. Evaluation of a population-based measure of quality of life: the Health and Activity Limitation Index (HALex). *Qual Life Res* 1998; 7(2) 101-14.
- Erickson P, Wilson R and Shannon I. Years of healthy life. *Healthy People 2000 Stat Notes* 1995; (7) 1-15.
- Fairbank JC, Couper J, Davies JB and O'Brien JP. The Oswestry low back pain disability questionnaire. *Physiotherapy* 1980; 66(8) 271-3.
- Fanshel S and Bush JW. A Health Status Index and Its Application to Health Services Outcomes. *Operations Research* 1970; 18(6) 1021-1066.
- Feeny D, Furlong W and Barr RD. Multiattribute approach to the assessment of health-related quality of life: Health Utilities Index. *Med Pediatr Oncol* 1998; Suppl(1) 54-9.
- Feeny D, Furlong W, Boyle M and Torrance GW. Multi-attribute health status classification systems. Health Utilities Index. *Pharmacoeconomics* 1995; 7(6) 490-502.
- Feeny D, Furlong W, Torrance GW, Goldsmith CH, Zhu Z, DePauw S, Denton M and Boyle M. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. *Med Care* 2002; 40(2) 113-28.
- Fryback DG. A US valuation of the EQ-5D. *Med Care* 2005; 43(3) 199-200.
- Fryback DG, Dasbach EJ, Klein R, Klein BE, Dorn N, Peterson K and Martin PA. The Beaver Dam Health Outcomes Study: initial catalog of health-state quality factors. *Med Decis Making* 1993; 13(2) 89-102.
- Fryback DG, Dunham NC, Palta M, Hanmer J, Buechner J, Cherepanov D, Herrington SA, Hays RD, Kaplan RM, Ganiats TG, Feeny D and Kind P. US Norms for Six Generic Health-Related Quality-of-Life Indexes From the National Health Measurement Study. *Med Care* 2007; 45(12) 1162-1170.
- Fryback DG, Lawrence WF, Martin PA, Klein R and Klein BEK. Predicting quality of well-being scores from the SF-36: Results from the Beaver Dam Health Outcomes Study. *Medical Decision Making* 1997; 17(1) 1-9.
- Fryback DG, Palta M, Cherepanov D, Bolt D and Kim JS. Comparison of 5 Health-Related Quality-of-Life Indexes Using Item Response Theory Analysis. *Med Decis Making* 2009.
- Gold MR. *Cost-effectiveness in health and medicine*. New York: Oxford University Press; 1996.

- Gold MR, Patrick DL, Torrance GW, Fryback DG, Hadorn DC, Kamlet MS, Daniels N and Weinstein MC (1996). Identifying and valuing outcomes. Cost-effectiveness in health and medicine. Gold MR SJ, Russell LB, Weinstein MC. New York, Oxford University Press: 82-134.
- Gold MR, Stevenson D and Fryback DG. HALYS AND QALYS AND DALYS, OH MY: Similarities and Differences in Summary Measures of Population Health. *Annu Rev Public Health* 2002; 23 115-34.
- Hanmer J and Fryback DG. Do large national surveys yield equivalent population norms for health related quality of life measures? [Abstract]. *Medical Decision Making* 2006; 26(1) E45.
- Hanmer J, Hays RD and Fryback DG. Mode of Administration Is Important in US National Estimates of Health-Related Quality of Life. *Med Care* 2007; 45(12) 1171-1179.
- Hays RD, Prince-Embury S and Chen H. *RAND-36 Health Status Inventory*. San Antonio (TX): The Psychological Corporation; 1988.
- Hunt SM, McKenna SP, McEwen J, Williams J and Papp E. The Nottingham Health Profile: subjective health status and medical consultations. *Soc Sci Med [A]* 1981; 15(3 Pt 1) 221-9.
- Insinga R and Fryback D. Understanding differences between self-ratings and population ratings for health in the EuroQOL. *Quality of Life Research* 2003; 12(6) 611-619.
- IOM. *State of the USA Health Indicators: Letter Report*. Washington, DC: The National Academies Press; 2009.
- Jia H and Lubetkin EI. Estimating EuroQol EQ-5D scores from Population Healthy Days data. *Med Decis Making* 2008; 28(4) 491-9.
- Kaplan MS, Berthelot JM, Feeny D, McFarland BH, Khan S and Orpana H. The predictive validity of health-related quality of life measures: mortality in a longitudinal population-based study. *Qual Life Res* 2007; 16(9) 1539-1546.
- Kaplan R and Anderson J (1996). The general health policy model: An integrated approach. Quality of Life and Pharmacoeconomics in Clinical Trials. Spilker B. New York, Raven: 309-322.
- Kaplan R and Bush J. Health-Related Quality of Life Measurement for Evaluation Research and Policy Analysis. *Health Psychology* 1982; 1 61-80.
- Kaplan RM, Bush JW and Berry CC. Health status: Types of validity and the Index of Well-being. *Health Services Research* 1976; 11(Winter) 478-507.
- Kazis LE, Lee A, Spiro A, 3rd, Rogers W, Ren XS, Miller DR, Selim A, Hamed A and Haffer SC. Measurement comparisons of the medical outcomes study and veterans SF-36 health survey. *Health Care Financ Rev* 2004; 25(4) 43-58.
- Keeney RL and Raiffa H. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. New York: John Wiley & Sons; 1976.
- Lohr KN, Brook RH, Kamberg CJ, Goldberg GA, Leibowitz A, Keesey J, Reboussin D and Newhouse JP. Use of medical care in the Rand Health Insurance Experiment. Diagnosis- and service-specific analyses in a randomized controlled trial. *Med Care* 1986; 24(9 Suppl) S1-87.
- Lohr KN, Kamberg CJ, Keeler EB, Goldberg GA, Calabro TA and Brook RH. Chronic disease in a general adult population. Findings from the Rand Health Insurance Experiment. *West J Med* 1986; 145(4) 537-45.

- Luo N, Johnson JA, Shaw JW, Feeny D and Coons SJ. Self-Reported Health Status of the General Adult U.S. Population as Assessed by the EQ-5D and Health Utilities Index. *Med Care* 2005; 43(11) 1078-1086.
- Mallinson S. The Short-Form 36 and older people: some problems encountered when using postal administration. *J Epidemiol Community Health* 1998; 52(5) 324-8.
- Mallinson S. Listening to respondents: a qualitative assessment of the Short-Form 36 Health Status Questionnaire. *Soc Sci Med* 2002; 54(1) 11-21.
- Mangione CM, Lee PP, Pitts J, Gutierrez P, Berry S and Hays RD. Psychometric properties of the National Eye Institute Visual Function Questionnaire (NEI-VFQ). NEI-VFQ Field Test Investigators. *Archives of Ophthalmology* 1998; 116(11) 1496-504.
- Neumann PJ and Greenberg D. Is the United States ready for QALYs? *Health Aff (Millwood)* 2009; 28(5) 1366-71.
- Nunnally JC and Bernstein IH. *Psychometric theory*. New York: McGraw-Hill; 1994.
- Orpana HM, Ross N, Feeny D, McFarland B, Bernier J and Kaplan M. The natural history of health-related quality of life: a 10-year cohort study. *Health Rep* 2009; 20(1) 29-35.
- Parkerson GR, Jr., Broadhead WE and Tse CK. The Duke Health Profile. A 17-item measure of health and dysfunction. *Med Care* 1990; 28(11) 1056-72.
- Patrick DL, Bush JW and Chen MM. Methods for measuring levels of well-being for a health status index. *Health Services Research* 1973a; 8(3) 228-245.
- Patrick DL, Bush JW and Chen MM. Toward an operational definition of health. *Journal of Health & Social Behavior* 1973b; 14 6-23.
- Patrick DL and Deyo RA. Generic and disease-specific measures in assessing health status and quality of life. *Medical Care* 1989; 27(3 Suppl) S217-32.
- Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, Thissen D, Revicki DA, Weiss DJ, Hambleton RK, Liu H, Gershon R, Reise SP, Lai JS and Cella D. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Med Care* 2007; 45(5 Suppl 1) S22-31.
- Reeve BB, Potosky AL, Smith AW, Han PK, Hays RD, Davis WW, Arora NK, Haffer SC and Clauser SB. Impact of cancer on health-related quality of life of older Americans. *J Natl Cancer Inst* 2009; 101(12) 860-8.
- Rosenberg MA, Fryback DG and Lawrence WF. Population-based estimates of health-adjusted life expectancy: Comparison of alternative methods for computation. {Abstract}. *Medical Decision Making* 1997; 17(4) 527.
- Selim AJ, Rogers W, Fleishman JA, Qian SX, Fincke BG, Rothendler JA and Kazis LE. Updated U.S. population standard for the Veterans RAND 12-item Health Survey (VR-12). *Qual Life Res* 2009; 18(1) 43-52.
- Shaw JW, Johnson JA and Coons SJ. US valuation of the EQ-5D health states: Development and testing of the D1 valuation model. *Medical Care* 2005; 43(3) 203-220.
- Spielberger CD, Gorsuch RL, Lushene R, Vagg PR and Jacobs GA. *State-Trait Anxiety Inventory for Adults. Sampler Set: Manual, Test, Scoring Key*. Redwood City, California: Mind Garden; 1983.

- Sullivan DF (1966). Conceptual problems in developing an index of health. Washington, D.C., U.S. Department of Health, Education, and Welfare.
- Sullivan DF. A single index of mortality and morbidity. *HMSHA Health Reports* 1971; 86(4) 347-355.
- Sullivan P and Ghushchyan V. Preference-Based EQ-5D Index Scores for Chronic Conditions in the United States. *Medical Decision Making* 2006; 26(4) 410-420.
- Sullivan P, Lawrence W and Ghushchyan V. A National Catalog of Preference-Based Scores for Chronic Conditions in the United States. *Medical Care* 2005; 43 736-749.
- Szende A, Oppe M and Devlin N, Eds. (2007). EQ-5D Value Sets: INventory, Comparative Review and User Guide. EuroQol Group Monographs, Vol. 2. Dordrecht, The Netherlands, Springer.
- Tarlov AR, Ware JE, Jr., Greenfield S, Nelson EC, Perrin E and Zubkoff M. The Medical Outcomes Study. An application of methods for monitoring the results of medical care. *Jama* 1989; 262(7) 925-30.
- Torrance G. Social Preferences for Health States: An Empirical Evaluate of Three Measurement Techniques. *Socio-Economic Planning Sciences* 1976; 10(3) 129-136.
- Torrance GW, Thomas WH and Sackett DL. A utility maximization model for evaluation of health care programs. *Health Services Research* 1972; 7(2) 118-133.
- Ware JE, Jr. SF-36 health survey update. *Spine* 2000; 25(24) 3130-9.
- Ware JE, Jr. and Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care* 1992; 30(6) 473-83.
- Weinstein MC and Stason WB. Foundations of cost-effectiveness analysis for health and medical practices. *N Engl J Med* 1977; 296(13) 716-21.
- WHO (1948). New York, World Health Organization.
- WHOQOL. The World Health Organization Quality of Life Assessment (WHOQOL): development and general psychometric properties. *Soc Sci Med* 1998; 46(12) 1569-85.