TO DEVELOP A RESEARCH AGENDA AND RESEARCH RESOURCES FOR HEALTH STATUS ASSESSMENT AND SUMMARY HEALTH MEASURES

Workshop Report

March 2003

Table of Contents

Objective

Donald L. Patrick, Ph.D., M.S.P.H.

The objective of this meeting was to address the measurement of population health status in the United States. Over the last 75 years, interest in population health status assessment has risen steadily. Many forces have contributed to this increased attention to population health. Foremost is the interest in how population health levels are affected by technological advances, such as the development of vaccines or surgical advances. Thomas McKeown discussed this issue in his classic book, *The Role of Medicine*,[1] in which he asserted that population health improvements can be attributed more to public health measures and improvements in the environment than to technological innovation.

Moreover, Ernest Gruenberg brought our attention to what he called "the failures of our success" in the saving of lives through technology that can increase the disability level in the population and lower population health status.[2] Also, the shift from infectious diseases to chronic diseases as major causes of death over the last century and the aging of the population have influenced population health, leading to a focus on the prevention of chronic conditions.

Thirdly, improvements in preventive medicine and health promotion coupled with environmental and behavioral modifications have sparked a debate on whether the period of illness and morbidity at the end of life can be shortened or compressed such that persons are healthy until just shortly before death.[3] Finally, one of the seldom-acknowledged reasons we began assessing population health was that during World War II, the U.S. Congress observed data showing that many young potential military inductees were rejected from service because of health problems. The number of rejections prompted many to ask, What is the health status of the country, and why is it that such a large number of people in the prime of their lives were not considered healthy?[4]

Over the past four decades, researchers have worked to combine mortality with other measures of health-related quality of life to form health status indexes. Sullivan, who worked at the National Center for Health Statistics (NCHS) in the 1960s, was among the first to identify the

conceptual problems in developing such indexes. In 1966, he published a very seminal piece on this matter in the *Rainbow Series*—the NCHS Vital and Health Statistics series (http://www.cdc.gov/nchs/products/pubs/pubd/series/ser.htm), so named because of the color of the documents.[5]

Earlier, in 1964, Sanders proposed combining measures of "functional adequacy" (the number of days each year individuals could fulfill their social roles) with mortality rates to create a modified life table.[6] Sanders saw the use of life-table methodology as one way of allowing health professionals to assess mortality and morbidity simultaneously. He proposed a measure of "effective" life years that reflects the current health of the population in terms of mortality and the effects of morbidity. Although he did not work out this method in detail, his idea of combining these data through survival analysis or life-table procedures led to further efforts to develop single indexes of health. Linder, as director of the NCHS in 1966, called for an overall index of health similar to the gross national product.[7] The proposed "gross national health deficit" was to blend disability days with days of life lost through death and/or lack of intervention.

Sullivan, in 1971,[8] was the first to apply Sanders' suggestion and demonstrate the usefulness of a concept of combined mortality and morbidity that fulfilled the desirable properties of a composite measure as recommended by Moriyama in 1968.[9] Sullivan described two related indexes based on a life-table model: the expectation of life free of disability and the expectation of disability. These indexes are computed by subtracting the duration of bed disability and inability to perform major activities from life expectancy using data collected from the National Health Interview Survey. Based on a 1965 current life-table, the average life expectancy at birth was 70.2 years. Adjusting this for time lost due to disability, where all types of disability were assigned a weight of 1.0, Sullivan estimated that a person born in 1965 would have 64.9 years of life free of disability and 5.3 years with disability.

For the Indian Health Service, Miller developed the Q index for ranking diseases affecting the American Indian and Alaska Native population according to potential response to intervention and days lost due to premature death, hospitalization, and clinic visits.[10] The computed Q values

correlated closely with professional judgments of Indian Health Service administrators on the impact of disease on the target population. Miller concluded that the Q index was a valuable tool for the political process of determining program priorities. The Q index was one of the first single indexes to use existing data, a significant advantage in index construction. Although the index excluded a wide number of concepts and the method for combining the different concepts into a single score was atheoretic, this index was among the first to demonstrate the possibilities that indexes have for assisting in the selection of health programs and for program budgeting. Chen constructed a number of similar indexes to quantify unnecessary disability and death.[11, 12]

Research sponsored by the NCHS contributed to the creation of a national health index. In 1965, Chiang proposed a technique for combining mortality and morbidity rates into a single index based on mathematical models of illness frequency, illness duration, and mortality.[13] This single-index technique made an important contribution to the development of composite measures: It weighted the rapidity with which individuals in various health states returned to a state of perfect health. These weights could be obtained from incidence rates calculated from national population surveys.

Several groups of researchers working largely independently began to pick up on the idea of using summary measures in population health assessment or in applications of decision making, including Card and colleagues,[14] Rosser and Watts,[15] Fanshel and Bush,[16] Torrance,[17] Williams and colleagues,[18] and Weinstein and Stason.[19] It was Weinstein and Stason's paper in the *New England Journal of Medicine* in 1977 that piqued people's interest in what a summary measure would actually do in economic evaluation. All of these efforts preceded the current attention to the World Health Organization (WHO) proposal to measure the burden of disease and create a health status index combining mortality and morbidity for use worldwide.[20]

Summary measures of health make better progress when a population health model and a decision strategy are used to help guide the use of the health status measures and the summary measures. The March 2003 issue of the *American Journal of Public Health* has a very nice display of population health models, creating a context for what it is that we need to do in order to improve population health. I also feel that summary measures cannot be interpreted unless we

have the determinants of health to help interpret them. Thus, when we get to the questions of decision making and interpretation, we are going to have to figure out the data that will help us interpret those issues. It is not just economic trends; as we know, ethnicity, income, and socioeconomic status are also important determinants of health, in addition to what we might put into the health care system.

We need to realize that the development of summary measures is not just a technical challenge. Interpretation of results in relation to what can or should be done to improve health is difficult. That is, what do we make of a summary measure going up or down, and do we interpret the change as good or bad in relation to its determinants? Economic determinants are well known to influence population health levels as are a myriad of social determinants.[21] How we relate the determinants of population health to policy requires demonstration, research, and application.

Another important question is, How do we get acceptance and encourage use of population health measures? We have involved as many end users of measures in this meeting as possible. Decision makers and policy analysts who make policy proposals must view population health status measures as useful tools. Without this acceptance and use, we shall not be able to move forward in an agenda to collect the data.

This conference has promise to push the field ahead in many ways. First, it can make transparent some of the measurement issues involved in creating health status indexes. Population health status indexes require a definition of health states, weights for these health states, and a model of transition of health over time. These required elements all involve assumptions that can be challenged by theory and by empirical data. Second, it can bring together the technical expertise required to address these measurement issues and the users of the data. Finally, it can help set an agenda for the next decade in building on the considerable progress made over the last five decades in describing and monitoring population health.

It is my hope that our discussions at the meeting illuminated the problems and identified the means by which we can address the problems.

References

1. McKeown T. *The Role of Medicine: Dream, Mirage, or Nemesis?* Princeton, NJ: Princeton University Press; 1980.
2. Gruenberg EM. The failures of success. 1977. *Milbank Q.* 2005;83:779-800.
3. Fries JF. The compression of morbidity: near or far? *Milbank Q.* 1989;67:208-232.
4. Stouffer SA, Guttman L, Suchman EA, et al. *The American Soldier: Studies in Social Psychology in World War II. Vol 4. Measurement and Prediction.* Princeton, NJ: Princeton University Press; 1950.
5. Sullivan DF. *Conceptual Problems in Developing an Index of Health.* Washington, DC: US Dept of Health, Education, and Welfare; 1966. Publication HRA 74-1017. Vital and Health Statistics Series 2, No. 17.
6. Sanders BS. Measuring community health levels. *Am J Public Health.* 1964;54:1063-1070.
7. Linder FE. The health of the American people. *Sci Am.* 1966;214(6):21-29.
8. Sullivan DF. A single index of mortality and morbidity. *HSMHA Health Rep.* 1971;86:347-355.
9. Moriyama IM. Problems in the measurement of health status. In: Sheldon EB, Moore WE, eds. *Indicators of Social Change.* New York, NY: Russell Sage Foundation; 1968:573-600.
10. Miller J. An indicator to aid management in assigning program priorities. *Public Health Rep.* 1970;85:725-731.
11. Chen MK. The G index for program priority. In: Berg RL, ed. *Health Status Indexes: Proceedings of the Conference on a Health Status Index.* Chicago, Ill: Chicago Hospital Research and Educational Trust; 1973:28-39.
12. Chen MK. The K Index: a proxy measure of health care quality. *Health Serv Res*. 1976;11:452-463.
13. Chiang CL. *An Index of Health: Mathematical Models*. Washington, DC: National Center for Health Statistics; 1965. Vital and Health Statistics Series 2, No. 5.
14. Card WI, Rusinkiewicz M, Phillips CI. Utility estimation of a set of states of health. *Methods Inf Med.* 1977;16:168-175.

15. Rosser RM, Watts VC. The measurement of hospital output. *Int J Epidemiol.* 1972;1:361-368.
16. Fanshel S, Bush JW. A health-status index and its application to health services outcomes. *Oper Res.* 1970;18:1021-1066.
17. Torrance GW. Health status index models: a unified mathematical view. *Manage Sci.* 1976;22:990-1001.
18. Williams AH. Applications in management. In: Teeling Smith G, ed. *Measuring Health: A Practical Approach*. New York, NY: Wiley; 1988:225-243.
19. Weinstein MC, Stason WB. Foundations of cost-effectiveness analysis for health and medical practices. *New Engl J Med.* 1977;296:716-721.
20. Murray CJ, Lopez AD. Evidence-based health policy—lessons from the Global Burden of Disease Study. *Science.* 1996;274(5288):740-743.
21. Marmot M, Wilkinson RG, eds. *Social Determinants of Health.* Oxford, England: Oxford University Press; 1999.

Contact Information:

Donald L. Patrick, Ph.D., M.S.P.H., Professor, Health Services, University of Washington
donald@u.washington.edu.

Health Status and Summary Measures of Population Health: Recommendations Past and Future

Marthe Gold, M.D.

Introduction

Over the past decade, numerous conferences and reports have discussed and debated the measurement of health status and health-related quality of life in the context of program evaluation (in public health and medical care) and economic analyses. To begin this report on the workshop, *To Develop a Research Agenda and Research Resources for Health Status Assessment and Summary Health Measures,* I review findings and recommendations from pertinent meetings and reports regarding the classification of health status and summary measures of population health. My intent is twofold: to remind us where we have already been, and to highlight recurring dilemmas and debates we will wish to take on. I will end with some thoughts about end users our efforts are intended to influence.

Because the findings and recommendations I am presenting have been harvested from similarly distinguished and thoughtful enterprises, they all have both breadth and depth. In the interest of parsimony, and given the focus of the workshop, I have engaged in some organizing and streamlining. First, findings and recommendations have been grouped broadly by the following themes: (a) the value of health status and health-related quality-of-life measures; (b) whether an "all purpose" measure is possible or desired; and (c) how measures relate or should relate to a decision-making framework. Following that, previously identified conundrums are presented, and recommendations to address them through research are summarized. Although verbatim language from the parent reports is used in presenting the overarching themes, when similar problems have been identified by more than one entity, language is melded. Finally, the perspective in this presentation is national rather than international. This is not to gainsay the needs of the field in the context of a global community but rather to emphasize the importance of

pressing the health services research and health policy communities in the United States to better coordinate efforts in this field.

Reports Reviewed

Differences in function or purpose and in the constituencies for each of the conferences and reports to be addressed led to different presentations and emphases in their findings and recommendations. Reports reviewed for this paper, together with their broad objectives, are as follows:

The Third Conference on Advances in Health Status Assessment—Institute of Medicine (IOM), 1992.[1] This conference was convened to "examine the use and usefulness of health status and health-related quality of life measures in clinical practice, and secondarily, in clinical research that will directly facilitate the application of these measures by practitioners."

Summarizing Population Health: Directions for the Development and Application of Population Metrics—IOM, 1998.[2] This report, arising from a 1997 IOM meeting, was intended to "encourage methodologists, ethicists, and policy makers to learn from each other and to work together to identify the strengths, limitations, and appropriate uses of summary measures."

Summary Report of Workshop: Identifying Summary Measures for Healthy People 2010—National Center for Health Statistics (NCHS), 1998.[3] This workshop was convened to identify a set of "summary measures" that might be used by federal and state governments to determine trends in health among subpopulations as well as differences between these groups so as to evaluate progress toward the two overarching goals of *Healthy People 2010*[4]: "Increase quality and years of healthy life" and "Eliminate health disparities."

Health Outcomes in OECD Countries: A Framework of Health Indicators for Outcome-Oriented Policy-Making—Organization for Economic Co-operation and Development (OECD), 1999.[5] This report summarized the state of the art in health outcome indicators for monitoring the health status of populations and for evaluating the performance and effectiveness of various health

policies and medical care interventions, with the intent to contribute to the future development of a common set of international health indicators.

The National Institutes of Health (NIH) Meeting on Burden of Illness (BOI) Summary Report and Recommendations—NIH, 1999.[6] This meeting and report were designed to respond to an IOM study on setting priorities at the NIH that asked the Institutes to "strengthen their analyses and use of health data such as burdens and costs of diseases, and of data on the impact of research on the health of the public."

Marrakech Conference, World Health Organization (WHO)—WHO, 1999.[7] In the context of the proliferation of work on summary measures and increasing debate about their application in public health, this conference was convened to provide a forum for discussion and debate over the scientific, ethical, and policy issues that surround summary measures of public health.

Report of the October 2000 Joint Economic Commission for Europe (ECE)/WHO Preparatory Meeting on Measuring Health Status—ECE/WHO, 2000.[8] Following the recommendations of the 1998 Joint ECE/WHO Meeting on Health Statistics, this meeting, hosted by Statistics Canada, was convened to make progress toward a common framework for measuring population health.

Themes

1. The Value of Summary Measures of Population Health (SMPH) and Health Status Instruments

"The value of formal health status questionnaires, for clinical settings, may lie in part in the documentation they provide for what was done to and for the patient, and for the expected and observed results of those services."[1]

"Summary measures of population health integrate mortality and morbidity and are increasingly relevant to both public health and medical decision makers in order to:

1. Describe differences and trends in the health of populations.
2. Inform decisions about alternative uses of health care dollars.
3. Assess cost-effectiveness of alternative personal health services and technologies."[2]

"These uses range from comparisons of the health of populations (or of the same population over time), quantification of health inequalities and priority-setting for health services delivery and planning, to guiding research and development of the health sector, improving professional training, and analyzing the benefits of health interventions in cost-effectiveness studies."[7]

Relevant and comparable information on health outcomes serves two purposes:

"1. Monitor current trends and forecast future needs in population health both within Member countries and across Member countries; and
2. Measure and evaluate the performance and effectiveness of various health policies and medical-care interventions."[5]

"Because they combine all dimensions of health, summary measures are valuable for comparing health of population groups over time, across disease categories, between countries, and distinguished by socioeconomic status, environment, or access to health care. They have been included as endpoints in clinical trials, and they facilitate cost-effectiveness analysis and comparisons of interventions with outcomes measured in different physical units."[6]

2. The Possibility of an "Ideal" Measure

"Although methodological innovation in population metrics has strengthened the analytical base for health decisions, the lack of accepted standard measures can create confusion and caution among potential users."[2]

"No single measure can adequately incorporate all aspects of health and mortality."[3]

"There is a significant lack of international consensus on the concepts of health and morbidity to measures, as well as the methodology and administration of these surveys, making international comparisons next to impossible."[5]

"While a universal composite health measure is an attractive goal for many, an indicator which can address all types of health problems for any population group is not feasible. Only an integrated set of international indicators will serve to underpin outcome-oriented policy making."[5]

"It would be desirable to develop a consensus summary measure of health. If the consensus measure were used on all population-based studies of health, it would facilitate comparisons over time, space and subpopulations. The ideal measure would be sensitive to small changes in every dimension or domain of health."[6]

"Current approaches to cross-population comparability need strengthening and warrant further research and development."[8]

3. The Relationship of Measures to a Decision-Making Framework

"There would be a clear understanding of the relationship between medical care intervention and health status; indicators would clearly relate to areas involving substantial resources or burden of disease."[5]

"Health status measurement is insufficient by itself; it needs to be embedded in a framework that allows health status assessment to be meaningfully connected to health policy."[8]

"Better measures that communicate investment in health, other than the most widely used health statistic, the 'percentage of GDP (gross domestic product) spent on health care,' are needed for enriched policy discussions."[8]

4. Problems Identified

i. There is no consensus as to which domains/dimensions to include in health status/health-related quality-of-life measures.

It is generally accepted that the breadth of domains can be more circumscribed for populations than for clinical uses. At the Ottawa conference, for example, in response to information developed by WHO and Statistics Canada, a sub sample of attendees ranked 21 possible priority domains for a shorter profile to be used as the basis for SMPH. When measures are considered with respect to their ability to discern clinical changes that can be linked to interventions, the requirement for more domains/better discrimination is habitually noted.[5,6]

Some domains are likely to be more relevant to different populations/subpopulations; excluding these domains means these populations will not be taken into account. Both at the national and international levels, appropriate domains might be related to underlying population characteristics, such as culture, religion, education, occupation, and income.[2,6]

Recommendations for Research

- Develop a clearer understanding of the ways in which partial and summary measures behave in depicting health and the extent to which users of the measures appreciate their characteristics or limitations; identify what additional information they need.[2]
- Perform more comparative analyses of the results of applying different instruments to the same population and across populations and over time. Clarify the merits and disadvantages of each measure in satisfying needs for decision making and monitoring.[5]
- Identify how diseases affect people's functioning and their needs for survival.[6]
- Gather information on multiple summary measures. Doing so would allow investigators to test the hypothesis that each of the measures results in essentially the same conclusions about health priorities. Comparisons of multiple measures on the same data set could advance understanding of the cross-sectional and longitudinal behavior of summary measures and facilitate development of a crosswalk between

measures. The experience with data sets using multiple measures might also encourage consensus for accepting a new measure that combines the best aspects of available measures.[6]

- Perform side-by-side comparisons of different measures, comparisons that are essential for adopting a standard, comparable set of measures.[2]
- Explore both the feasibility and value of developing models for specific diseases.[6]
- Obtain disease-specific data to measure changes in health status and emerging health needs.[5]
- Clarify which aspects of health would be suitable for a summary measure.[3]
- Consider the value of developing additional or alternative measures to use in the future.[3]

ii There is no consensus on how to assign meaningful values to health states.

Different valuation strategies, for example, standard gamble, time trade-off, visual analogue scales, and person trade-offs, will give rise to different weights for a particular health state or disease. Which of these is "best," and by what standard do we judge?[2,6]

The category of respondent (for example, patient, community member, health professional) has been found to influence the valuation of health states. Arguments have been made to use (a) patient responses to promote greatest accuracy, (b) community weights to best represent the preferences of a society, or (c) weights determined by professionals because of their ability to best understand and reflect upon the nuances of the different health states and the tasks involved in trading off.[2,6]

"Ethical difficulties have been noted with assuming that respondents' understanding of the weights they provide is consistent with how the weight is used in creating a health-adjusted life year (HALY), where survival and morbidity are traded off."[2]

"All measures of population health involve choices and value judgments in both their construction and their application. If these choices and judgments—and their policy

implications—are not understood and acknowledged, the result can be distrust and disregard of the measures."[2]

Recommendations for Research

- Examine the ways in which choices of values enter into both constructing and applying different measures of population health. Clarify the ethical assumptions and value judgments embedded in these different measures.[2]
- Compare different summary measures to determine how much variation in summary measures is caused by assigning different weights to health states.[2]
- Develop preferences for health states with a deliberative process rather than a simple population survey.[6]
- To ensure the ideal study of valuation of health states, survey people with and without any condition. Surveys of people with the condition should be recurrent to capture how they adjust.[6]
- Examine the extent to which weights vary across cultures and by age and sex.[6]

iii. The utilitarian notion of maximizing QALYs (quality-adjusted life years) no matter who benefits may not be consistent with dominant or ethical views held by society. Other issues of known importance to the public, such as helping those most in need or providing life-saving treatments, should also be built into the metric.[2,6]

Recommendations for Research

- Develop more sensitive, sophisticated techniques for examining the public's attitudes and reasoning related to the allocation of resources.[2]
- Deploy and test several measures to develop a body of empirical work on the distributive implications of these measures.[2]
- Conduct further philosophical work on issues of distribution.[2]

iv. The people whom health status and SMPH were designed to inform are not always clear as to what they represent or what function they can serve.

"The value and utility of these measures in guiding patient care and in improving patient outcomes and well-being must be demonstrable—if not fully demonstrated—before they will be widely adopted in clinical practice."[1]

"Policy makers have little understanding of the technical underpinnings of SMPH and have been wary about using them for decision making."[2]

"One of the major criticisms of aggregate summary measures is the lack of ability to identify why a change has come about. To fully inform policy decisions, the complete pyramidal structure of interlinking data and measures is needed... (allowing) an analyst to "drill down" to a lower level of data to explore the reasons for the observed change. With the information... routinely collected in the United States... such explorations are not feasible."[6]

Recommendations for Research

- Study how the use of summary measures and the differences between them can shape, improve, or distort policy decisions.[2]
- Examine the similarities and differences between SMPH as part of a strategy for assessing how well particular measures and strategies for measurement may serve different purposes at the local, national, and international levels.[2]
- Link information on the health of populations to information about risk factors to provide important epidemiologic insights for planning public health strategies and shaping programs for delivering health services.[2]
- Assess the critical ethical and policy implications of differing designs, approaches to implementation, and uses of measures.[2]
- Compare health outcomes flowing from different interventions and programs on a national and international level to obtain important information for policy makers about how health care resources should be allocated.[5]

Prioritizing Research: Some Challenges

These findings and recommendations demonstrate the large research agenda that remains. It therefore seems useful to do some clear thinking here about how we should prioritize directions for research so that more significant gains can be made in advancing these measures in decision-making environments in the United States. There are numerous problems in advancing the use of summary measures, and I highlight a few here in the hope of provoking some debate and discussion.

First, we often seem to be working with the right kinds of tools but at the wrong time in the evolution of U.S. health care policy. It is ironic that so much diligent work has been done to develop methods that are intended to help rationalize health care delivery in a country that seems so reluctant to rationally deploy its health care resources. Many of us who are primarily on the public health side of things, where budgets are leaner and where it is more difficult to say "that person is not my problem," continue to believe that at some point the tools we have been developing will be seen as helpful not only at the population level, as descriptors, but across the full health care system, in helping us to design more effective and efficient systems. But until that interest becomes more explicit, we will probably have to make progress on the basis of relatively small investments in these methods. And that means it could take awhile to gain the experience we require for substantial innovation.

Second, although interest is frequently recounted in developing a family of measures that can be all things to all people—allowing us a way to navigate between public health and medical interventions and across populations, link to specific diseases and risk factors, and be sensitive and responsive in documenting the effects of environmental or clinical intervention—we are still far from there. The 1992 IOM conference and the 1999 NIH meeting provide a forceful message that for measuring clinical outcomes, that is, evaluating interventions and assessments of quality of care, we need measures with a good deal of sensitivity and responsiveness. Drugs and other therapies often create changes in highly specific dimensions that do not make it to the more constrained list of domains that may be appropriate to population health settings. Although many of the affected dimensions may load onto broader domains, it is likely that interventions directed at specific clinical symptoms and conditions would be better serviced by disease/condition-

specific measures. Certainly we still have far to travel in understanding the relationship of some of these lesser symptoms and dimensions to the broader ones. And it is possible that this is not a first-order activity for which to strive, given the differing requirements for measurement in medicine and public health. (Please note that I say this with considerable reluctance.)

Third, the "weighting problem" has many layers to it, and we may wish to prioritize them for investigation, based on the needs of users. Much has been made of the variation in weights that arise when different measurement strategies (e.g., person trade-offs, time trade-offs, standard gamble) are used to make HALYs. Accordingly, we may want to figure out how important those differences are in the scheme of things by seeing how the ordering of burden changes with the different strategies. A lot has also been written about variations in value structures in different populations—again, we could test empirically what difference variation in valuation makes in the overall magnitude of burden. Also a source of contention is whether distributional values should be loaded on to HALYs by finding ways to incorporate and weight societal concerns for fairness and equity within them. This is an area where empirical data are relatively sparse.

But behind all of these weighting problems lays a "meta-question," which is, to what extent does the public and its decision makers buy into using these measures to inform and resolve questions of equity and allocation? It may be that Americans are fundamentally market driven and wish to maximize efficiency. Or it may be that our vast stores of commitment to fairness have yet to be uncovered. It seems to me that at the heart of the research priorities we identify here should lie the question of whether the schema we develop have saliency for how people wish to conduct business and make policy.

Mining the Policy Environment

Numerous opportunities are available to explore the use of measures in policy-making contexts. Efforts could be made to better understand how measures can service real-world decision-making needs. A few examples:

The Office of Management and Budget (OMB) has proposed new guidance to federal agencies for conducting regulatory analyses, which are used by the agencies to anticipate and evaluate the likely consequences of their actions.[*] OMB is recommending that cost-effectiveness analyses be prepared "for all major rulemakings for which the primary benefits are improved public health and safety."[†] A recent meeting sponsored by Resources for the Future joined methodologists who work on cost-benefit analyses for areas such as environment, transportation, and food safety with methodologists who work in valuation of outcomes in the realm of HALYs to begin to explore how regulatory cost-effectiveness analyses might best account for outcome evaluation.

A report from IOM[‡] regarding extension of Medicare coverage to several services argued more generally for furnishing cost-effectiveness analyses to evaluate additions to Medicare coverage. The report suggested that the current procedure for estimating the costs to Medicare of covering a new service—although necessary for understanding budgetary implications—provides an incomplete picture of the value for money of such actions. Last summer (2002), at a hearing of the House of Representatives' Energy and Commerce Committee, the leadership signaled its interest in pursuing broader types of analyses in considering services for insurance under Medicare. Systematically exploring how measurement of disease burden and cost-effectiveness analysis would be considered by congressional personnel seems highly relevant to advancing the research agenda.

State government provides another venue for exploring the use of measures. Large budget deficits have caused states to search for ways to curtail Medicaid spending. During its initial development, the Oregon Health Plan, an early innovator in efforts to broaden the coverage of health insurance, solicited public values to inform its policies. No such efforts are going on now in Oregon, as large limitations are being imposed on coverage for the state's Medicaid and

---

[*]Regulatory analyses are intended to determine whether the benefits of an action are likely to justify the costs and to discover what alternative actions are likely to be most cost-effective.

[†]OMB is seeking public comment on this regulation by April 3, 2003. The regulation can be accessed at: www.omb.gov.

[‡]Field MJ, Lawrence RL, Zwanziger L, eds. *Extending Medicare Coverage for Preventive and Other Services.* Washington, DC: National Academy Press; 2000.

uninsured populations. Research here with respect to public preferences in how decisions are made would be useful to public discourse as well as to building this field.

The managed care setting is another venue to explore how medical personnel and consumers understand the measurement of quality and the prioritization of resources. Both of these areas are center stage considerations for all managed care organizations. By describing to caregivers and consumers what it is we seek to do, and by gaining guidance in how we can capture their health and priorities in plausible ways, we will build measures in which people have more confidence.

Conclusion

In closing, let me suggest that we are to some degree stymied by a lack of communication with our end users and a failure to understand them. Rarely are inventions and innovations developed at the bidding of their ultimate users, who ideally would have come forward to sagely direct their progress, yet in the world of public policy, the understanding of users and their feedback and endorsement are mandatory for uptake of these products. The report of the IOM Committee on Summary Measures of Population Health[2] was heavily influenced by the presence of decision makers from medicine and public health at its preceding conference. Many of these persons noted that they were unfamiliar with these methods or skeptical toward them. A primary recommendation of the Committee was that public health and medical professionals be educated and trained to promote their understanding of the appropriate uses of SMPH.

But more than understanding is required. We need active engagement of the public and the decision-making community to tell us what will and will not fly. And to get that, we might wish to give some priority to broader research questions that help us understand how people want decisions on allocating resources to be made. Not only what technique we should use, but what the process should be. If we ask those sorts of questions, we will not only educate potential consumers but, more importantly, will also craft more credible and useful methods that will attract investment from sources of funding because these methods are central to answering questions of great societal need.

References

1. Lohr KN. Applications of health status assessment measures in clinical practice. Overview of the third conference on advances in health status assessment. *Med Care.* 1992;30(5 Suppl):MS1-14.
2. Field MJ, Gold MR, eds. *Summarizing Population Health: Directions for the Development and Application of Population Metrics.* Washington, DC: National Academy Press; 1998.
3. Crimmins EM. Summary Report of Workshop: Identifying Summary Measures for Healthy People 2010 (workshop organized by National Center for Health Statistics). September 17-18, 1998; University of Maryland, College Park, MD. 1998.
4. U.S. Department of Health and Human Services. *Healthy People 2010.* 2nd ed. Vols. I and II. Washington, DC: U.S. Government Printing Office; 2000.
5. Jee M, Zeynep O. Health Outcomes in OECD (Organisation for Economic Co-operation and Development) Countries: A Framework of Health Indicators for Outcome-Oriented Policy-Making. *OECD Labour Market and Social Policy Occasional Papers,* No. 36. Paris: OECD Publishing; 1999.
6. National Institutes of Health Meeting on Burden of Illness: Summary Report and Recommendations. Bethesda, MD: Office of Science Policy, Office of the Director, National Institutes of Health; 1999.
7. Murray CJL, Salomon JA, Mathers CD, et al. *Summary Measures of Population Health: Concepts, Ethics, Measurement and Applications.* Geneva: World Health Organization; 2002.
8. Report of the October 2000 Joint ECE/WHO Preparatory Meeting on Measuring Health Status.

Contact Information:

Marthe Gold, M.D., Medical Professor and Chair, Department of Community Health and Social Medicine, The City University of New York, goldmr@med.cuny.edu.

Commentary

Article: Health Status and Summary Measures of Population Health: Recommendations Past and Future, by Marthe Gold, M.D.

Dean Jamison, Ph.D.

In her article, "Health Status and Summary Measures of Population Health: Recommendations Past and Future," Marthe Gold, M.D., has given a very detailed review of the findings and recommendations from various meetings and reports concerning the classification of health status and summary measures of population health. In this commentary, I will slightly extend her discussion using an analogy involving the National Income Product Accounts.

If you go through the sets of common themes that Dr. Gold discussed, almost every one is an issue that appears in the whole construction of the National Income Product Accounts. However, reports on health status and summary measures focusing on several important areas of economic activity that arguably are essential for inclusion in the National Income Product Accounts are simply nonexistent. That is, the value of women's work or any other service that is performed at home or not bought and sold in markets is an important thing that is simply left out. It is not that people are not aware of these issues; they are constantly argued and discussed. Rather, it is that the whole valuation mechanism is clearly weighted into an "ability to pay" approach that puts prices on things through weights that are importantly influenced by the income of the people holding the preferences. This is not to say that this is an altogether bad approach or that the

shortcomings of the approach are not well understood. However, it does show how familiar many of the problems are that we have in trying to construct summary measures of population health through a combination of industry-specific or condition-specific burden elements into more aggregate measures.

The National Income Product Accounts was first developed about 60 years ago by the British. They were trying to create a systematic set of accountabilities for political leaders around reasonably well-measured aggregate total gross domestic product (GDP) indices, unemployment rates, and inflation rates for four, five, or six summary measures of the economic health of a country. Their objective—political accountability through good measurement—is one of the principle things that we are trying to accomplish.

Their other objective in constructing good measures lay in the essential nature of good measures for doing policy research that goes beyond good anecdotal and case study type policy research. If one is trying to examine the relationship of, for example, health care finance policies or broad-based population interventions—if one is to understand the consequences of policy—one clearly has to have these measures. Thus, it is for policy research and political accountability that we are trying to establish the broad population measures, and I think that we are in fairly good shape. However, there really are clearly a lot of major issues still in front of us. As I look at my own personal experience 10 years ago with the World Bank, with the development of burden-of-disease measures based on disability-adjusted life years, I note that those measures had the virtue of being comprehensive. There were many well-founded and well-justified criticisms of specific methods and certainly of the weaknesses of the data, which is not something that we were in a position to do much about. However, the resistance in the international community to using

measures of disease burden by regions for specific countries that are still based on disability-adjusted life years—and the persistence of the World Health Organization (WHO) in publishing those kinds of measures despite well-founded and well-documented criticisms—I think reflects a failure of the critical community to be in the business of developing alternatives. To develop an alternative set of measures of the burden of disease, state by state for the United States or country by country for the Americas, is a big job. Part of it is interesting research, but much of it is just a long, hard slog. However, only in that developmental exercise will we provide feedback into the research endeavors that tells us how well we are addressing the conceptual problems, the practical problems that are quite reasonably put forward usually by the academic critics.

I think that the theoretical basis, particularly for the preferences-based measurement domain, is really in extraordinary disarray. We need to tighten up the theoretical underpinnings, which will play an important role in some of the specific technical advances that will improve the actual measurement processes.

Lastly, it is important to have as part of the government's research agenda or some of the academic research agendas a constant effort of trying to develop complete products that are competitors with whatever the existing products might be. We should focus our discussions not on the questions of what is wrong with a measure or what can be made a little bit better in it but rather on the question of whether one measure is better than another. Until there are better alternatives, criticisms of any particular measure are of little value.

Contact Information:

Dean Jamison, Ph.D., Senior Fellow, Fogarty International Center, National Institutes of Health, jamisond@mail.nih.gov.

Understanding and Comparing Existing Summary Measures of Health and Health-Related Quality of Life: The State of the Art

Dennis G. Fryback, Ph.D.

Introduction

The purpose of this brief paper is to discuss the state of the art in summary measures of health.

I will not attempt a systematic and comprehensive review of summary measures, for two reasons: First, there are far too many candidate instruments for a work of this length. McDowell and Newell's encyclopedic book on health measures lists 21 instruments in its chapter on "General status and quality of life"[1]; as of February 2003 the MAPI Institute's (Lyon, France) Quality of Life Instruments Database (QOLID) Web page listed 73 instruments under the rubric "Generic Instruments,"[2] and no doubt further instruments will be added. The two lists contain only 11 instruments in common, however. Thus, the second reason this overview is not comprehensive: it is doubtful we could agree upon the set of summary measures that would be deemed "comprehensive."

What is meant by "summary measure of health"? For purposes here I mean a measure that represents the overall, generic health of a person as a single number (or at most a few numbers that each summarizes a domain of health). Michael Wolfson, of Statistics Canada, speaks eloquently about the need for a system of health indicators and measures forming a pyramid for measuring population health.[§] He envisions a broad base of detailed health information, with successively higher layers in the pyramid being increasingly aggregated summary measures. At the apex of the data pyramid is a scalar summarizing all the health information below. A

"summary measure," which sits at or very near the apex of Wolfson's data pyramid, does not refer to any specific disease or condition but instead summarizes the joint impacts of all aspects of a person's health. McDowell and Newell[1] define this apex as the realm of "health profiles" and "health indices." Advocates of health profiles contend that the tip of the pyramid is flat with no further aggregation possible or needed, and advocates of indices contend that it must end at a point with a single final summary score representing overall health:

> The profiles emphasize the diverse aspects of health or quality of life; proponents of this measurement school hold that the dimensions of health should be kept separate and that measurement is only meaningful within each domain. Supporters of the health index school agree that health has several dimensions but argue that real-life decisions demand that we combine the impressions from each dimension into an overall score.[1]

As fair disclosure, my personal inclinations place me in the second group. Finally, I will focus on summary measures of *health,* not summary measures of *population health.* Summary measures of health are generally conceived as summarizing morbidity, its impact on well-being, and possibly its impact on the social and role functions of an individual person. Summary measures of population health combine summary measures of the health of individual persons within a population with population statistics relating to mortality into one summary statistic, such as health-adjusted or disability-adjusted life expectancy.[3,4]

The requirement of general applicability: Generic summary measures.

The summary measures of interest here are generic measures, i.e., measures not designed to relate to any particular health condition but rather intended to measure health as a holistic, generic quality affected by any disease or combination of diseases. Generic instruments are intended to be used without modification across all diseases and conditions as well as across all medical interventions. A generic instrument should also apply across differing populations.[5]

[§] Wolfson M. An overview of issues and objectives regarding comparable population health status information. Presentation at a conference on Measuring Health Status, Joint U.N./E.C.E. and W.H.O. Expert Meeting, October 23–26, 2000. Ottawa, Canada.

Why demand such sweeping generality from a summary measure of health? After all, measures developed specifically to measure effects of a given disease will surely be more sensitive to specific impacts of that disease (e.g., for the case of dementia, see Silberfeld and colleagues[6]). And measures developed for use in special populations will certainly include aspects of health important to those populations that may not be so important in others. So why not use a collection of specialized measures? The simple answer is that a measure needs to be generic for comparisons across diseases, across persons with multiple conditions, and across heterogeneous subpopulations if it is to be used to summarize population health in locations with the diversity of North America.

A collection of specific measures does not summarize health. Were there separate measures for each disease or population, we would still have to combine them into a single summary. We would also have administrative problems as to which instrument should be used for a person who has multiple conditions or who belongs to several different subpopulations. In any case, a post hoc combining of specialized measures amounts to making a generic measure at the apex of the data pyramid; it seems prudent to conduct this task with forethought instead of afterthought. Accordingly, only generic measures will be assessed here.

Examples of Measures

Any list of summary measures put forth as "the state of the art" will be contentious. At the risk of contention, I suggest as exemplars the following instruments:

Health profiles using summated rating scales:

1. The Medical Outcomes Study 36-Item Short Form Health Survey (SF-36)
2. The World Health Organization WHOQOL-BREF Quality of Life Assessment

Health indexes:

3. The Quality of Well-Being scale (self-administered form) (QWB-SA)
4. The EuroQol EQ-5D (EQ-5D)

5. The Health Utilities Index Mark 2 and Mark 3 (HUI2/3)
6. SF-6D, a preference-based measure derived from the SF-36.

These lists could be expanded a great deal. The first category could well include the Nottingham Health Profile (NHP)[7] or the Sickness Impact Profile (SIP),[8] which the SF-36 has more or less replaced in general use. The second category might also include the Assessment of Quality of Life (AQoL),[9] a more recently proposed index, or the Health and Activities Limitation Index (HALex),[10] an *ad hoc* index created to approximate the QWB and HUI using data from the National Health Interview Survey. The list could also be augmented by other, derivative summary measures using SF-36 profiles as inputs to a regression equation to predict QWB scores[11] or HUI2 scores.[12] Instead of adding a "derivative indexes" category, I have put the SF-6D under the health index rubric and only note the others here. All of these will be mentioned but not discussed at length in what is to follow.

A brief history of these measures.

SF-36.

The SF-36 is one of the most widely used health profile questionnaires. A Medline search for "SF-36" in the title or abstract yields over 2,200 references. The SF-36 consists of 36 questions ("items") in a self-administered questionnaire asking the respondent about various aspects of health "over the past 4 weeks." Items were selected from a larger pool—some 250—used as a general instrument for assessing health in the RAND Health Insurance Experiment.[13] SF-36 Version 1.0 (SF-36v1) came into wide use in the early 1980s among both researchers and managed care organizations to studying patient outcomes. The original, larger item pool was constructed through a process beginning with a conceptualization of general health status, the postulating of hundreds of statements bearing on domains in this conceptual structure, and psychometric reduction of the pool, where items were discarded that were unreliable or that did not add information to the basic factor structure. The final 36 items were selected by further paring of the remaining pool to retain a desired level of precision on each subscale while maximizing efficiency.

This first version of the SF-36 was distributed by the RAND Corporation in the public domain as the RAND-36, with items corresponding exactly to the SF-36v1; scoring of the RAND-36 differs slightly but largely inconsequentially from scoring of the SF-36v1.[5, 14, 15]

The SF-36 does not yield a single summary score. Instead, it divides into a profile of eight scores (vitality, general health perceptions, mental health, bodily pain, social function, role function as limited by emotional problems, role function as limited by physical problems, and physical function), which in turn are aggregated into two summary scores developed using factor analysis, the mental health component score (MCS), and the physical health component score (PCS).[16] A similar two-component scoring algorithm exists for RAND-36.[15] SF-36 scoring is not based on preference judgments; instead, its scales are psychometric summated rating scales.[17]

U.S. norms for SF-36v1 were established in a 1992 national survey of non-institutionalized adults.[18] In 1996, Version 2 (SF-36v2) was released, and it appears to be the standard today.[19] A 1998 mail-out survey using a list of households maintained by National Family Opinion Research as representative of the noninstitutionalized U.S. adult population established age- and sex-specific norms for the SF-36v2 based on some 6,700 persons (67.8% response rate) with an age range of 18–96 years, of whom 84% were white and 80% had completed high school. SF-36v2 clarified some of the wording of questions in Version 1 (e.g., "full of pep" became "full of life") and changed from response scales with six categories on some items to five categories, eliminating a category, "a good bit of the time," deemed ambiguous. SF-36v2 also represents a change to norm-based scoring and interpretation; rescoring SF36-v1 using similar techniques allows backward comparability. Extensive tables from both surveys are available as privately published, limited-use paper documents at the SF-36 Web site (http://www.sf36.com). SF-36v1 has generally been made available without charge to researchers. Significantly, since early 2002, a fee has been charged even for research purposes to license use of SF-36v2, to obtain the population norms using this version, and to use the norm-based SF-36v2 scoring algorithms, as these are not published in the open literature.** This charge has led some researchers and

** Personal communication by e-mail July 22, 2002, Bonnie Denis, Client Services–Licensing Department, QualityMetric Inc., www.qualitymetric.com, who quoted $2,000 annual fee for up to 5,000 survey administrations in U.S. English.

government users to continue using SF-36v1. SF-36v2 is available in over 20 different translations. A shortened form of the SF-36, using a subset of 12 items, is available as the SF-12.

WHOQOL-BREF.

In the early 1990s the World Health Organization formed the Quality of Life Group, or WHOQOL group, to develop an instrument intended from the start to be applicable cross-culturally. This collaboration, spanning 15 field centers around the globe and across many cultures, defined quality of life broadly:

> [Quality of life is] individuals' perception of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards and concerns. It is a broad-ranging concept affected in a complex way by the persons' physical health, psychological state, and level of independence, social relationships and their relationship to salient features of their environment.[20]

A lengthy process of discussion, worldwide focus groups, and reduction through factor analysis resulted in the WHOQOL-100, a 100-item instrument encompassing 24 facets of quality of life, each measured by four items using 5-point Likert response scales, and four general items reflecting overall quality of life and perceptions of health. The realization that a 100-item instrument is unwieldy for use with populations led to further psychometric reductions, resulting in the WHOQOL-BREF, a 26-item instrument. Twenty-four items were selected from the 100 to cover the 24 facets of the WHOQOL-100, and two general items were added.[20] Because there is only one item per facet, individual facets are not scored. Responses to the facets can be collected into the four domain scores based on factor loadings from a confirmatory factor analysis of the 100 items, which apportions facets among four constructs representing physical and psychological aspects of quality of life, a construct for social relationships, and a construct for environmental impacts on quality of life.

QWB.

The Quality of Well-Being Scale (QWB) evolved from the Index of Well-Being, the earliest summary health-related quality-of-life index. The QWB also was the first index to arise from a

conjoining of multi-attribute utility theory with psychometric methods for scale construction.[21–23] Four dimensions are summed into the aggregated measure: physical function, mobility, social function, and a dimension that represents symptoms/problems on a given day. The original QWB summary averaged the daily well-being scores for the 6 days immediately prior to the time of administration and required direct administration by a trained interviewer.

The developers of the QWB set out to define a descriptive framework covering all health states between optimum function and death that "might serve as a classification matrix and sample frame."[23] Using an extensive review of medical reference works, they attempted to list all the ways that diseases and injuries can affect a person's behavior and role performance without respect to etiology. From this list, they created a framework of dimensions for the index, distinguishing levels for each dimension. Based on health states observed in large surveys, they reduced the framework to define 43 distinct states. A similar comprehensive review of medical texts resulted in a comprehensive list of symptoms and health problems, which they grouped into what was termed "symptom/problem complexes" and which designated health impacts beyond those affecting the functional dimensions. The symptom/problem complexes, applied in combination with health states defined by combinations of three other dimensions (physical function, mobility, social function), resulted in some 1,548 composite health states, of which only 1,100 were deemed feasible combinations. Varying sets of 42 hypothetical persons were constructed by assigning each person to one of the 1,100 health states; the collection of sets covered all 1,100 states but the sets were intentionally not mutually exclusive. In 1974–75, a panel of 690 adults from a probability sample of households in San Diego rated every person in one of the sets of hypothetical persons employing a category estimation procedure that was psychometrically equivalent to estimating magnitude of preference for the health state, with zero representing death. These preference ratings were pooled, and regression analysis yielded a scoring system using the three scales and the symptom/problem complex list to assign an additive score between 0 (dead) and 1 (perfectly well) to each possible health state. In application, this score is computed for each of the previous 6 consecutive days, and the average is the final QWB scale score. The QWB requires administration by a trained interviewer.

Recently, a self-administered form, the QWB-SA, has been developed and tested.[24] The QWB-SA asks about the past 3 days instead of 6. It has been validated and compared to SF-36 in a population of older adults.[25] The discussion below focuses on the QWB-SA, as I expect it largely to supplant the QWB because of the reduced costs of administration, bringing it more into line with other self-administered instruments.

EuroQol EQ-5D.

The EuroQol EQ-5D (EQ-5D) was developed by the EuroQol consortium of European health researchers, which was formed in 1987 out of a shared desire to have a standardized, simple, self-administered instrument that was not disease-specific to describe and value health-related quality of life. Developed concurrently in five languages,[26] the EQ-5D is intended to complement other quality-of-life measures, and it has been purposefully developed to generate an interval scale index of health, thus giving it potential for use in economic evaluation.[27]

The EQ-5D consists of two components. One is a self-rating of health on a 20-cm visual analog scale anchored by "best imaginable health state" and "worst imaginable health state." The second component, on which I focus here, is a descriptive system with five dimensions: mobility, self-care, usual activity, pain/discomfort, and anxiety/depression. Each dimension has three levels designated simply as no problem, some problem, and extreme problem. Respondents rating their health are asked to check the level of each dimension most descriptive of "your health today." The most commonly used scoring employs a system of weights ("tariffs" in the British EuroQol members' jargon) which was derived from a community sample in the United Kingdom.[28,29] The resulting health state valuations have a high of 1.0 (the health state with no problems), assign death to 0.0, and allow valuations less than zero (states worse than death). A project to develop United States scoring weights was funded by the Agency for Healthcare Research and Quality and is headed by Stephen Coons of the University of Arizona. A report of the United States weights is anticipated by mid-decade.

The EQ-5D has been used for a mailed population survey in the United Kingdom.[27, 30] and for a computer-assisted population survey by Statistics Canada (where it was paired with the HUI3[31]); it has been employed in a number of European countries. Its low response burden and absence of

a licensing fee have made it an increasingly popular choice in quality-of-life surveys. In addition, it was a component of the most recent wave of the Medical Expenditure Panel Survey fielded by the Agency for Healthcare Research and Quality.

Health Utilities Index, Mark 2 and Mark 3 (HUI2/3).

Developed in the early 1970s, the Health Utilities Index is the second oldest of the indexes described in this paper. Torrance and colleagues conceived the index using multi-attribute utility theory.[32-35] In this approach, data from focus groups are used to devise a set of dimensions that collectively are comprehensive in covering important aspects of health. The investigators then set forth verbal descriptions of successively worse levels of function for each of these dimensions to serve as a category scale for each. Finally, systematic elicitation methods are used to assess scale weights for each level of each dimension and the scaling constants in a multiplicative utility model.[36] Torrance invented the time-tradeoff assessment method as an alternative elicitation technique to standard gambles for quality-of-life weights to scale the HUI1. Two major revisions, HUI2 and HUI3, are the indexes in current use.[1, 37–40] Scoring of HUI2 and HUI3 is based on standard gamble assessments carried out with probability samples from the community of Hamilton, Ontario.

The HUI3 consists of eight scales: vision, hearing, speech, ambulation, dexterity, emotion, cognition, and pain. The HUI2 combines vision, hearing, and ability to speak into one dimension of "sensation," then has additional scales for mobility, emotion, cognition, self-care, pain, and fertility. The two indexes are generally scored using separate mappings from the same questionnaire ("HUI2/3"), which is available in self-administered form (15 questions) and computer-assisted telephone interview (CATI) form (40 questions). The HUI2/3 is available in 16 different copyrighted questionnaires, if one considers variation in recall intervals and mode of administration. Statistics Canada uses an abbreviated version of the CATI questionnaire to collect only the HUI3 for the Canadian National Population Health Survey. Health Utilities Group, Inc., is the exclusive distributor of the HUI2/3 and licenses permission to use its proprietary materials (questionnaires, coding algorithms, and procedure manuals) one study at a time. The typical fee to use one administration mode and associated manuals in one research

study is $CAN 4,000.[41] The HUI2/3 questionnaire is available in seven languages, with others in preparation.

SF-6D.

Several researchers have tried to collapse the SF-36 into a single summary score equivalent to the preference-based health indexes. Fryback used data from the Beaver Dam Health Outcomes Study, which collected both SF-36 and QWB data, to develop a predictive equation for QWB scores using the eight SF-36 scales as independent variables.[11] Nichol used another large survey to estimate a similar equation to predict HUI2 scores from SF-36.[12]

Brazier and colleagues set out to develop a preference-based scoring system for the SF-36 from first principles. They used a subset of SF-36 items to represent statements about health in six domains—physical functioning, role limitations due to health, social functioning, pain, mental health, and vitality. They specifically excluded the items on perceptions of general health, which they considered to be redundant constituents of a generic health index. The resulting classification system is termed the SF-6D (the "6D" referring to the six dimensions covered), and responses on a completed SF-36 questionnaire can be used to compute a corresponding SF-6D.[42] Some 18,000 unique health states can be defined by the SF-6D. A subset of 249 of these states, carefully selected to allow identification of interactions among SF-6D dimensions, was identified, and sets of six each were rated by a sample of 836 members of the general public (United Kingdom) using the standard gamble technique. Econometric modeling was used to develop equations based on these ratings, extending the valuations to the entire space covered by the SF-6D.[43]

The six instruments are summarized in Table 1 with respect to nominal number of dimensions of health, number of questions, and number of alternative formulations.

Table 1. Characteristics of the instruments.

Item	**SF-36v2**	**WHOQOL-BREF**	**QWB-SA**	**HUI2/3**	**EQ-5D**	**SF-6D**
National origin	USA	International	USA	Canada	Europe/UK	USA
Individual questions in the instrument	36	26	73	15 or 40*	5	11
Alternative dimensional structures	2	1	1	2	1	1
Structure 1:						
Dimensions to represent health state	(summated scales) 8	4	4	(HUI2) 7	5	6
Structure 2:						
Dimensions to represent health state	(factor scores) 2	—	—	(HUI3) 8	—	—
Single summary scalar score to represent overall health?	no	no	yes	yes	yes	yes**

*Length depends on mode of administration: 15-item self-administered questionnaire or 40-item branching, interviewer-administered questionnaire.

**In theory, this will yield a single score, but as yet there is no final score among several alternative functions that have been derived.

All generic instruments at the current front of the science of summarizing health have strengths and weaknesses. Every author reviewing them has noted that there is no one clear choice for all purposes (e.g., Coons and colleagues[5] and Hawthorne and coworkers.[44])

Where Is the State of the Art?

In this section, I address what appear to be the issues concerning the "state of the art." These include content of the measures, ability to discriminate among health conditions, problems associated with mode of administration in a linguistically and culturally pluralistic population, and the problem of weights (in the case of indexes).

Content of Summary Health Measures.

The question is, are the state-of-the-art measures leaving out something critical to measuring health? In other words, is a new measure likely to be developed that will differ radically from those now existing with respect to content? I think probably not. But the leading measures differ in nominal content and each represents a particular—and debatable—set of choices about inclusions and exclusions of health concepts and domains.

The constructs generally listed as domains of health status include physical, mental, and social health, each of which is usually measured by indicators of function on varying aspects within them ("facets" in the language of the WHOQOL). For example, within physical health are indicators of mobility, energy/vitality, dexterity, pain, and so forth. Anxiety, depression, cognitive abilities, emotion/mood, and memory are often listed as the major aspects of mental health. Social health refers to interaction with others, social support, intimacy, and other such phenomena. Many indicators overlap domains—ability to perform in one's major life role, be this work, study, or leisure activities—can be affected by a deficit in any or all of these health domains.

Every summary measure represents a different trade-off between the desire to assess every aspect of all possible domains with great precision and the preference for a brief instrument that can be administered in a reasonable time. Because generic summary measures are usually part of

a larger battery of items in a research or survey project, brevity is not a luxury but a necessity. On the other hand, every measure has been criticized for leaving something out. The EQ-5D, being the most brief, receives the most criticism, with the most recent complaints being that it does not cover hearing problems or mental retardation[45] or problems of cognition.[6] EQ-5D advocates may agree that these are deficits but also point out that the "usual activities" dimension may incorporate some effects of these problems.

No generic measure is immune to such critique. Every one of them has been criticized by researchers interested in some specific disease entity—arthritis, diabetes, dementia, sleep disorders, cancer, stroke, and many others—as leaving out details critical to measuring important effects of that disease. But the pressure for brevity has led to broad-brush characterizations of health rather than use of many items or dimensions each specific to only one or a few diseases. On the other hand, I do not expect that any new summary index will be devised that includes a startlingly new dimension of health that would cause us all to slap our heads and exclaim that this is an important aspect of health that we all missed before. The issues now largely concern operationalization, not a concern about missing a big piece of health.

The nominal content of the health domains for the measures used as examples in this essay is presented in Table 2. The listing of health domains in the table takes the measures' content at face value as labeled by the developers. It is difficult to match content on the basis of labels, however, as they may be imprecisely labeled regarding actual content as operational zed. The psychometric content of individual items and attributes is probably best compared across two or more instruments if the instruments are administered simultaneously to a large sample of persons with varying health conditions and the suite of responses at levels of a single item or the domain score is examined using factor analysis.

Even at face value, however, there are apparent similarities and differences. The two profiles (SF-36 and WHOQOL-BREF) ask for self-evaluated overall health; the indexes do not. HUI2 and HUI3 stay "within the skin," as they do not address social interactions,[46] so does the EQ-5D, although it includes one item asking about "usual activities," which some respondents might understand to include social interactions. The SF-36, SF-6D, and QWB all have questions about

social activities. The WHOQOL-BREF goes the furthest “beyond the skin,” with questions about four facets of the environment around the person as well as social support.

Table 2. Nominal content of the six instruments.

SF-36v2	WHOQOL-BREF	QWB-SA	HUI2	HUI3	EQ-5D	SF-6D
Physical Health • Physical functioning • Role limitation due to physical functioning • Bodily pain • General health **Mental Health** • Vitality • Social functioning • Role limitations due to emotional functioning • Mental health	**Physical Health** • Pain and discomfort • Sleep and rest • Energy and fatigue • Mobility • Activities of daily living • Dependence on medicinal and medical aids • Work capacity **Psychological** • Positive feelings • Thinking, learning, memory, and concentration • Self-esteem • Bodily image • Negative feelings • Spirituality **Social Relationships** • Personal relationships • Social support • Sexual activity **Environment** • Freedom, physical safety • Home environment • Financial resources • Access to health and social care • Opportunities to acquire new information and skills • Participation & opportunities for recreation and leisure • Physical environment (pollution, noise, traffic, climate) • Transport **Overall Health assessment**	**Mobility** **Physical Activity** **Social Activity** **Symptoms/Problems**	**Sensation** • Vision • Hearing • Speech **Mobility** • Ambulation • Dexterity **Emotion** **Cognition** **Pain** **Self-Care** **Fertility**	**Vision** **Hearing** **Speech** **Ambulation** **Dexterity** **Emotion** **Cognition** **Pain**	**Mobility** **Self-Care** **Usual Activities** **Pain/Discomfort** **Anxiety/Depression**	**Physical Functioning** **Role Limitations** **Social Functioning** **Pain** **Mental Health** **Vitality**

The discrepancies highlight two continuing debates regarding the content of summary measures:

1. Should self-rated health be included in a measure? The HALex includes only two dimensions, self-evaluated health and limitations in physical activity,[10] and assigns them equal importance. But in general, the indexes do not include self-rated health, partly on the theory that it is redundant in a summary measure meant to represent overall health.[43] Others believe that self-rated health will vary with a person's expectations and accommodation to limitations in abilities (termed a "response shift" in health ratings), and therefore, an index describing only limitations is missing critical information about how people in the population feel about their own health states.[47, 48]
2. Should the measure contain elements beyond the person being evaluated, i.e., "beyond the skin"? Ware and colleagues originally argued that the definition of personal health should "end at the skin," including only physical and mental health components and leaving out social interactions, which include factors beyond the person. They suggested that these data should be collected as useful explanatory variables but not as direct components of personal health.[49] A compromise is to collect limitations in social functioning with the attribution "due to your health."[50] Both profiles collect these data. The indexes are somewhat mixed as discussed above. The newest summary index, the AQoL, includes beyond-the-skin attributes as important inputs.[51]

Summary measures developed before the 1990s generally conceive of health as a deficit from an ideal level of health. None of the SF-36, SF-6D, QWB, HUI2/3, or EQ-5D measures is sensitive to positive aspects of health. The WHOQOL attempts to capture positive affect as well as spirituality and feelings of purpose. These may reflect an alternative emphasis on psychological well-being as a foundation for assessing quality of life in health and aging,[52–54] whereas the measures constructed earlier were based more on measuring deficits in health and function. So in addition to the two controversies enumerated above, the third controversy at the state of the art regarding content is how much of a general construct of well-being—and especially, in my opinion, the component of "spirituality"—should be included in a summary measure of health. Although this might be subsumed in part under the question whether a self-rating of health should be included, it seems to me that there is sufficiently different content to this positive

aspect that it constitutes a separate question. It is not clear to me that spirituality is a part of *health-related* quality of life, although it may be a part of overall well-being for some.

A final aspect of content relates to a special subpopulation: children. The generic instruments do not have content specifically identified as relating to the special concerns of children. The HUI was originally developed in the context of evaluating health outcomes in low-birth-weight babies, and the developers note that HUI2 and HUI3 may be used to measure health in children as well as adults. But an important research question is whether there should be special content for children.[55–58] I would add that the value placed on an individual's ability for self-care or his/her independence from others may well differ depending on whether the individual is a child or adult, but the indexes extended to evaluation of children do not differentiate valuation depending on age.

Are summary measures sufficiently sensitive to differences in health?

All state-of-the-art measures discriminate frankly ill from well persons. Even the EQ-5D, a measure often criticized as insensitive, can distinguish persons with differing levels of disease activity in the many serious diseases for which it has been tested. Because the EQ-5D does not elaborate beyond the three levels of each of the five dimensions, it yields less information to explain or describe differences. But in sufficiently large samples, the composite score has proved a useful discriminator. In smaller samples, the lack of more levels within the dimensions is an apparent disadvantage. The WHOQOL-BREF and SF-6D, both being more recently proposed, have not received nearly as much testing as the other measures, but there is no reason to believe they will be drastically different. If a summary measure is used to stand upon a data pyramid, allowing users to "drill down" to more detail associated data below, the explanatory power of the measure itself is less of a concern, as its job is to summarize.

Whereas all of the state-of-the-art measures are sensitive to changing levels of disease or impairment for most major health conditions, they are not equally sensitive to small changes in very good or very poor health states—i.e., they have ceiling and floor effects. Especially in general population surveys, the EQ-5D is affected by a low ceiling, with a high percentage of respondents placed in the top health state.[59,60] The SF-12, a subset of the SF-36, shows more

discrimination among respondents in general population surveys than the EQ-5D, although both the SF-12 and SF-36 also have ceiling effects. The QWB is often cited as not having a ceiling effect, as the distribution of QWB scores in a general population does not appear truncated at the topmost score, as do the distributions of most of the other measures. The HUI3 appears to have continuous scores right up to the maximum score in a general population survey[61] but seems to have a ceiling-limited distribution even in some clinical populations (e.g., see figures in Grootendorst and colleagues[62]).

Sensitivity at the lower end of the scales is a separate question. All except for WHOQOL-BREF have reportedly been used in seriously ill patient populations with apparent success. For longitudinal follow-up, however, a philosophical question arises: What to do with persons who die during follow-up? The indexes have been scored systematically so that the state "dead" is represented as 0, and thus, in principle, an index so scored will have no loss due to death at follow-up (those who die are given scores of 0). The profiles are scored only relative to the worst state described in the descriptive system, so we are left with a decision at data analysis about how to deal with interim deaths. None of the measures deal easily with the question of what is a meaningful change in score. All have been reported to have acceptable although not perfect test-retest reliability, but evaluation is subject to true changes in health over intervals sufficiently long to assure that memory artifact does not interfere with measures of reliability. In one sense, any change reliably above the level of measurement error should be meaningful. Describing clinically meaningful differences begs the question of what diseases and in whom we are talking about, however. At the level of changes in mean population scores, I do not believe the question of meaningful differences is entirely answerable. One collects as much data as one can afford in order to discriminate as small a difference as possible in meaningful subgroups, and the size of the subgroup samples more often than not determines the discriminating power.

Questions about administration.

All of the instruments discussed here can be self-administered or interviewer administered and completed by competent men and women. Most are available in many languages. Scores using different modes of administration can differ systematically.[63, 64] Because response rates and survey costs vary by mode of administration, any systematic differences in scores due to mode

should be known and accounted for. This can be especially important in reaching special populations.[65] Generally, while most researchers expect differences due to mode in all the measures, specific evidence is relatively sparse. McHorney discusses mode differences with respect to SF-36 administration[18]; for HUI2/3, see Grootendorst et al.[62] Administration face-to-face by interviewers tends to produce more socially desirable answers and answers reflecting somewhat better health; this effect seems to be related in complex ways to the age and socioeconomic status of respondents (and possibly of interviewers).

The newest trend in questionnaire-based health profiles is adaptive computer-based administration. A large item pool is maintained on the computer, and logic and computation determines the (n+1)st question to be asked conditioned on the first n responses. The object is to minimize the number of questions presented while maximizing discrimination of the respondent's location in the important health domains. The SF-36 has been transformed to adaptive testing in this fashion under the name Dynamic SF-36®, reportedly taking fewer questions and much less time to complete (http://www.amihealthy.com/static/DynamicSF36Info.asp).[66] Although adaptive testing can be used with questionnaire-based health profiles, this method of administration is not suitable for the index measures, as they require a fixed format for scoring.

A further issue in administration concerns language. There may be systematic effects due to match or mismatch of home language/culture and the language of administration.[67–69] In the United States, Hispanics are fast becoming the predominant minority and are even predicted to be soon the largest cultural/ethnic group in California. The growth in the Spanish-speaking population brings importance to the question of whether the various indexes are invariant to language of administration versus the home language, i.e., are responses similar when a person whose home language is Spanish is interviewed in English versus in Spanish, and is administration in Spanish equivalent to administration in English for comparison across persons whose home languages are one or the other? This will be equally true in the future for different Hispanic subcultures (e.g., those who derive from Puerto Rico versus those from Mexico) as well as for many other languages. As the United States becomes more linguistically and culturally

plural it is important to know performance characteristics for state-of-the-art summary measures administered in English and in other languages.

The method now used to translate instruments across languages is forward-and-backward translation along with a global judgment of linguistic equivalence made by bilingual persons. Analytical demonstration of cultural equivalence is still pending development of suitable methods. The World Health Organization is exploring use of "vignettes" to address this issue. Standard adjectives are used to describe degree of function or pain in short vignettes about hypothetical persons ("John is experiencing moderate pain"). Respondents rate the vignettes as well as their own health. Transformations for scale values associated with the ratings of one's own health are then derived under the assumption that the adjectives are invariant across respondents.[70] Although there are many complexities to implementing such an approach, it is an interesting direction in questionnaire-based summary measures.

The final problem of mode of administration concerns proxies. With cognitively incompetent adults and very young children, we cannot expect self-completion of any of the standard instruments, and researchers have turned to proxies—parents, spouses, physicians—to respond for the person who is unable. No study able to compare self- and proxy-completed measures has deemed the proxy to be completely adequate.

The problem of weights.

Every summary index depends on some weighting scheme. Health profiles, effectively by default, weight individual items within one domain as equal. The summary measures for quality of life go further, assigning relative importance in the form of scaling weights to the health domains, which are combined to a single summary score. In the case of QWB and the EQ-5D, as described in the introduction, these weights were effectively derived by regression analysis of holistic ratings of health states. Scaling constants for the domains in HUI2 and HUI3 are based on average standard gamble assessments of special health states.[38, 71]

Would collection of weights in different populations make a difference? There is not a great deal of evidence. As far as I am aware, the Canadian weights are used universally outside of Canada

for scoring the HUI2/3. Although the QWB was originally developed with weights from a sample of San Diego residents, different weights were derived as part of the Oregon Medicaid experiment.[72] With some notable differences, the majority of weights were quite similar between the San Diego and Oregon populations; the implied ranking from the two systems for common health states was very similar. As noted in the introduction, a large-scale project is under way to assess U.S. weights for the EQ-5D. It is an empirical question whether differences between the U.S. and the U.K. weights will lead to different policies based on EQ-5D data.

I personally regard the question of weights to be more important politically than empirically, in that each user nation or community wants to put its own imprimatur on the summary measure to legitimize its use in local policy decisions. I anticipate, however, that modest changes in weighting systems will make little difference in overall decisions, because most summary measure scoring systems are linear weighted averages and, as such, are relatively insensitive to modest levels of "noise" in the weights[73] This is particularly so in multi-attribute utility models, even the multiplicative form of the HUI2/3, where insensitivity of decisions to modest variation in weights is known as the "flat maximum" phenomenon.[74]

The existence of alternative scoring systems for the same health state classification system (e.g., two different sets of weights for EQ-5D health states) complicates the task of summarizing health, because the fact that they apply to the same classification system means the underlying health data are not changed—just their summary changes. Thus, in principle, we can compare data collected using the EQ-5D in the United States to data collected in the United Kingdom, using the same instrument employing either U.K. or U.S. weights (as long as we use only one set of weights for all data in the comparison); a Canadian data set using EQ-5D could be scored with either function as well. The ability to compare results is not lost because two scoring systems for the EQ-5D exist. If, however, the U.S. survey used the HALex, the U.K. survey the EQ-5D, and the Canadian survey the HUI3, our ability to compare results of the surveys across the populations is lost. This is the most critical aspect of the state of the art of summary health measures today: the multiplicity of existing measures and the prospect of promulgation of newer ones is potentially destructive to the fundamental purpose of measurement—the ability to compare results across studies or surveys.

The Issue of Comparison

This final issue is important enough to warrant its own section in this essay. The purpose of using a summary measure of health is to make comparisons. A research study or survey that uses a measure unique to that study or survey may have high internal validity for comparison of persons in the study to each other or to themselves over time. Such measures, however, cannot be used to compare that study's results externally to other studies or surveys.

A truly useful summary measure allows comparisons both internal and external to the particular study in which it is used. The state of the art in summary measurement of health is that we have a Babel of measures. In principle, the EQ-5D, the HUI2/3, the SF-6D, and the QWB should be measuring the same thing and should give the same results. They are all put forth as generic summary measures of health scaled to the anchors 1 = perfect health and 0 = dead, and in principle they should yield the same scores. This does not happen in practice, however.

In the pretest for Statistics Canada's National Population Health Survey, both HUI3 and EQ-5D were administered by telephone to a sample of about 1,500 adults from the general Canadian population. The correlation between summary scores on the two indexes was 0.69; correlations among single-attribute scores in approximately matching domains were on the same order (e.g., HUI3 pain and EQ-5D pain/discomfort correlated at 0.61; HUI3 ambulation and EQ-5D mobility correlated at 0.50).[31] The mean scores for subgroups were relatively close numerically; for example, the mean score for 200 persons aged 65–74 was 0.805 (S.E. 0.031) on HUI3 versus 0.786 (S.E. 0.032) on EQ-5D.[32] The corresponding EQ-5D score for adults aged 65–74 in a population sample in the United Kingdom was 0.78 (S.E. 0.012).[75] But studies using both measures in patient populations report substantially different scores for the same patients,[76, 77] even though the correlations are high.

The mean QWB score for 382 adults in the same age range as that reported above, but from a community population in the Beaver Dam Health Outcomes Study, was 0.72 (S.E. 0.011).[78] This score is statistically different from the NPHS HUI3 score and the United Kingdom EQ-5D score.

Is this difference a real reflection of health in the different populations or just differences in measurement systems? We cannot answer this question, although all three measures nominally measure the same thing. To answer, we need to have at least one common data set in which the QWB and the HUI3 and the EQ-5D are all administered in consistent fashion and then develop "cross-walk" equations among the measures. To date, there is no such master comparison data set.

The SF-36 health profile has been administered with QWB and with HUI2 in different studies, and "cross-walk" equations were developed to predict the single summary index scores using the SF-36 profile scales.[11, 12] These equations allow at least rough cross-comparison between studies reporting SF-36 profiles and other studies reporting QWB or HUI2 scores, but each equation accounts for only approximately 50% of the variance in the single summary index scores, and predictions are subject to regression effects. The unaccounted residual may be a function only of noise, given the reported reliabilities of the various scales, or may represent health content of one measure missing from the other, or both these sources.

It seems appropriate that a state-of-the-art summary measure be as widely comparable to other research studies and surveys as possible. The two ways to achieve this are (a) to dominate "the market," i.e., to be "the measure of choice" used in almost all studies (much like Microsoft Windows® has dominated the desktop computer operating system market), or (b) to have well-developed cross-walks between the measure and as many others as possible in broad population studies as well as patient populations. To date, method (a) seems to predominate. Perhaps it is time to enter the era of method (b).

The critical question at the frontier of the state of the art in summary measures of health is whether we need a new and better measure than those existing or whether we should make do with current measures. When a new measure is added to the pantheon of existing measures, data collected with that new measure are not immediately comparable to those collected in the past using the supplanted measures. To maintain contact with previously collected data so that we may compare current to past results, we must use the same measures as the previous research, or we must do the studies to develop the cross-walks among the new and old measures.

Conclusion

In this essay, I have discussed a number of issues concerning the "leading" summary health measures that characterize the state of the art today. I confess I am reluctant to endorse development of YASM—"yet another summary measure." Even if a new measure were to correct defects in existing measures, it would create a discontinuity with past data for purposes of comparison. The magnitude of gain in validity, decreased costs, precision, generality, or other benefits must offset the loss in ability to compare to previous research findings. Although all existing measures have defects, surely the trade-off between efficient administration and comprehensive coverage will leave any new instrument with defects as well. Furthermore, it will not be universally adopted, as existing measures now have large user bases.

Rather than embarking on developing a new instrument, I believe a different course of enabling research should be taken. First, investigators should undertake simultaneous development of cross-sectional data sets in populations of interest using multiple summary measures. Norm-generating surveys with simultaneously administered instruments provide standard background comparison data that tie the instruments together. These data sets can also be used to develop or refine "cross-walks" among instruments. Similarly, multiple instruments should be used to follow cohorts expected to change in one or more domains of health to explore and document differential responses among the instruments to changes in various aspects of health. And a research program should be initiated to document differential scores associated with modes of administration.

For summarizing population health, we will probably never have the perfect summary health measure. Instead, we will have multiple public use data sets systematically using multiple summary measures in long-term longitudinal population studies. These data sets will be a tool box for researchers who can use only one or two measures in their own studies to allow comparison of their results to a wide range of population and clinical data collected using other measures. It is time we built the tool box for those who make use of the tools.

References

1. McDowell I, Newell C. *Measuring Health: A Guide to Rating Scales and Questionnaires.* New York: Oxford University Press; 1996.
2. MAPI. QOLID version 1.8, Quality of Life Instruments Database, MAPI Research Institute.
   Available at: http://www.qolid.org/. Accessed March 5, 2003.
3. Murray CJ, Salomon JA, Mathers C. A critical examination of summary measures of population health. *Bull World Health Organ.* 2000;78:981-994.
4. Molla MT, Wagener DK, Madans JH. Summary measures of population health: methods for calculating healthy life expectancy. *Healthy People 2010 Stat Notes.* 2001(21):1-11.
5. Coons SJ, Rao S, Keininger DL, et al. A comparative review of generic quality-of-life instruments. *Pharmacoeconomics.* 2000;17:13-35.
6. Silberfeld M, Rueda S, Krahn M, et al. Content validity for dementia of three generic preference based health related quality of life instruments. *Qual Life Res.* 2002;11:71-79.
7. Hunt SM, McEwen J, McKenna SP. Measuring health status: a new tool for clinicians and epidemiologists. *J R Coll Gen Pract.* 1985;35:185-188.
8. Bergner M, Bobbitt RA, Carter WB, et al. The Sickness Impact Profile: development and final revision of a health status measure. *Med Care.* 1981;19:787-805
9. Hawthorne G, Richardson J, Osborne R, et al. The Assessment of Quality of Life (AQoL) Instrument: Construction, Initial Validation & Utility Scaling. West Heidelberg, Victoria, Australia: Centre for Health Program Evaluation, Monash University; 1997:24.
10. Erickson P. Evaluation of a population-based measure of quality of life: the Health and Activity Limitation Index (HALex). *Qual Life Res.* 1998;7:101-114.
11. Fryback DG, Lawrence WF, Martin PA, et al. Predicting Quality of Well-being scores from the SF-36: results from the Beaver Dam Health Outcomes Study. *Med Decis Making.* 1997;
    17:1-9.
12. Nichol MB, Sengupta N, Globe DR. Evaluating quality-adjusted life years: estimation of the health utility index (HUI2) from the SF-36. *Med Decis Making.* 2001;21:105-112.

13. Stewart AL, Ware JE Jr, eds. *Measuring Functioning and Well-Being. The Medical Outcomes Study Approach.* Durham, NC: Duke University Press; 1991.
14. Hays RD, Prince-Embury S, Chen H. *RAND-36 Health Status Inventory.* San Antonio, TX: The Psychological Corporation; 1988.
15. Hays RD, Sherbourne CD, Mazel RM. *The RAND 36-Item Health Survey 1.0. Health Econ.* 1993;2:217-227.
16. Ware JE Jr, Kosinski M, Bayliss MS, et al. Comparison of methods for the scoring and statistical analysis of SF-36 health profile and summary measures: summary of results from the Medical Outcomes Study. Med Care. 1995;33(4 Suppl):AS264-AS279.
17. McHorney CA, Ware JE Jr, Raczek AE. The MOS 36-Item Short-Form Health Survey (SF-36): II. A psychometric and clinical test of validity in measuring physical and mental health constructs. *Med Care.* 1993;31:247-263.
18. McHorney CA, Kosinski M, Ware JE Jr. Comparisons of the costs and quality of norms for the SF-36 health survey collected by mail versus telephone interview: results from a national survey. *Med Care.* 1994;32:551-567.
19. Ware JE Jr. SF-36 health survey update. *Spine.* 2000;25:3130-3139.
20. WHOQOL. The World Health Organization Quality of Life Assessment (WHOQOL): development and general psychometric properties. *Soc Sci Med.* 1998;46:1569-1585.
21. Fanshel S, Bush JW. A health-status index and its application to health-services outcomes. *Oper Res.* 1970;18:1021-1066.
22. Patrick DL, Bush JW, Chen MM. Methods for measuring levels of well-being for a health status index. *Health Serv Res.* 1973;8:228-245.
23. Kaplan RM, Bush JW, Berry CC. Health status: types of validity and the index of well-being. *Health Serv Res.* 1976;11:478-507.
24. Kaplan RM, Sieber WJ, Ganiats TG. The Quality of Well-Being Scale: comparison of the interview-administered version with a self-administered questionnaire. *Psychol Health.* 1997;12:783-791.
25. Andresen EM, Rothenberg BM, Kaplan RM. Performance of a self-administered mailed version of the Quality of Well-Being (QWB-SA) questionnaire among older adults. *Med Care.* 1998;36:1349-1360.

26. The EuroQol Group. EuroQol--a new facility for the measurement of health-related quality of life. *Health Policy.* 1990:16:199-208.
27. Brooks R. EuroQol: the current state of play. *Health Policy.* 1996;37:53-72.
28. Kind P, Dolan P. The effect of past and present illness experience on the valuations of health states. *Med Care.* 1995;33(4 Suppl):AS255-AS263.
29. Dolan P. Modeling valuations for EuroQol health states. *Med Care.* 1997;35:1095-1108.
30. Kind P. The EuroQol instrument: an index of health-related quality of life. In: Spilker B, ed. Quality of Life and Pharmacoeconomics in Clinical Trials. 2nd ed. Philadelphia: Lippincott-Raven Publishers; 1996:191-201.
31. Belanger A, Berthelot J-M, Guimond E, et al. A head-to-head comparison of two generic health status measures in the household population: McMaster Health Utility Index (Mark III) and the EQ-5D. Ottawa, Canada: Health Analysis and Modeling Group, Statistics Canada; 2000.
32. Torrance GW. A Generalized Cost-Effectiveness Model for the Evaluation of Health Programs. Buffalo, NY: Department of Industrial Engineering, State University of New York at Buffalo; 1971.
33. Torrance GW, Thomas WH, Sackett DL. A utility maximization model for evaluation of health care programs. *Health Serv Res.* 1972;7:118-133.
34. Torrance GW, Boyle MH, Horwood SP. Application of multi-attribute utility theory to measure social preferences for health states. *Oper Res.* 1982;30:1043-1069.
35. Torrance GW. Measurement of health state utilities for economic appraisal. *J Health Econ.* 1986;5:1-30.
36. Keeney RL, Raiffa H. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs.* New York: John Wiley & Sons; 1976
37. Feeny DH, Torrance GW, Furlong WJ. Health utilities index. In: Spilker B, ed. *Quality of Life and Pharmacoeconomics in Clinical Trials.* 2nd ed. Philadelphia: Lippincott-Raven Publishers; 1996:239-252.
38. Torrance GW, Feeny DH, Furlong WJ, et al. Multiattribute utility function for a comprehensive health status classification system: Health Utilities Index Mark 2. *Med Care.* 1996;34:702-722.

39. Torrance GW, Siegel JE, Luce BR. Framing and designing the cost-effectiveness analysis. In: Gold MR, Siegel JE, Russell LB, et al., eds. *Cost-Effectiveness in Health and Medicine.* New York: Oxford University Press; 1996:54-81.
40. Furlong W, Feeny D, Torrance GW, et al. *Multiplicative Multi-Attribute Utility Function for the Health Utilities Index Mark 3 (HUI3) System: A Technical Report.* Hamilton, Ontario: McMaster University Centre for Health Economics and Policy Analysis; 1998.
41. Health Utilities Group, Inc. Questionnaire Development, Translations and Support. Available at: http://www.healthutilities.com/HUI-frames.htm. Accessed February 28, 2003.
42. Brazier J, Usherwood T, Harper R, et al. Deriving a preference-based single index from the UK SF-36 Health Survey. *J Clin Epidemiol.* 1998;51:1115-1128.
43. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. *J Health Econ.* 2002;21:271-292.
44. Hawthorne G, Richardson J, Day NA. A comparison of the Assessment of Quality of Life (AQoL) with four other generic utility instruments. *Ann Med.* 2001;33:358-370.
45. Oostenbrink R, A Moll HA, Essink-Bot ML. The EQ-5D and the Health Utilities Index for permanent sequelae after meningitis: a head-to-head comparison. *J Clin Epidemiol.* 2002;55:791-799.
46. Feeny D. Health-status classification systems for summary measures of population health. In: Murray CJ, Salomon JA, Mathers C, et al., eds. *Summary Measures of Population Health: Concepts, Ethics, Measurement, and Applications*. Geneva: World Health Organization; 2002:329-341.
47. Carr AJ, Gibson B, Robinson PG. Measuring quality of life: is quality of life determined by expectations or experience? *BMJ.* 2001;322:1240-1243.
48. Ubel PA, Loewenstein G, Hershey J, et al. Do nonpatients underestimate the quality of life associated with chronic health conditions because of a focusing illusion? *Med Decis Making.* 2001;21:190-199.
49. Ware JE Jr, Brook RH, Davies AR, et al. Choosing measures of health status for individuals in general populations. *Am J Public Health.* 1981;71:620-625.

50. Stewart AL. The Medical Outcomes Study framework of health indicators. In: Stewart AL, Ware JE Jr, eds. *Measuring Functioning and Well-Being: The Medical Outcomes Study Approach.* Durham, NC: Duke University Press:1992:12-24.
51. Hawthorne G, Richardson J, Osborne R. The Assessment of Quality of Life (AQoL) instrument: a psychometric measure of health-related quality of life. *Qual Life Res.* 1999;8:209-224.
52. Ryff CD, Keyes CL. The structure of psychological well-being revisited. *J Pers Soc. Psychol.* 1995;69:719-727.
53. Ryff CD, Singer B. Psychological well-being: meaning, measurement, and implications for psychotherapy research. *Psychother Psychosom.* 1996;65:14-23.
54. Kahneman D, Diener E, Schwarz N, eds. *Well-being: The Foundations of Hedonic Psychology.* New York: Russell Sage Foundation; 1999.
55. Eiser C, Mohay H, Morse R. The measurement of quality of life in young children. Child *Care Health Dev.* 2000;26:401-414.
56. Eiser C, Morse R. The measurement of quality of life in children: past and future perspectives. *J Dev Behav Pediatr.* 2001;22:248-256.
57. Eiser C, Morse R. Quality-of-life measures in chronic diseases of childhood. *Health Technol.* 2001;5:1-157.
58. Eiser C, Morse R. A review of measures of quality of life for children with chronic illness. *Arch Dis Child.* 2001;84:205-211.
59. Johnson JA, Ohinmaa A, Murti B, et al. Comparison of Finnish and U.S.-based visual analog scale valuations of the EQ-5D measure. *Med Decis Making.* 2000;20:281-289.
60. Johnson JA, Pickard AS. Comparison of the EQ-5D and SF-12 health surveys in a general population survey in Alberta, Canada. *Med Care.* 2000;38:115-121.
61. Houle C, Berthelot J-M. A head-to-head comparison of the Health Utilities Index Mark 3 and the EQ-5D for the population living in private households in Canada. *QoL Newsl.* 2000;24:5-6.
62. Grootendorst P, Feeny D, Furlong W. Health Utilities Index Mark 3: evidence of construct validity for stroke and arthritis in a population health survey. *Med Care.* 2000;38:290-299.

63. Rockwood TH, Kane RL, Lowry A. Mode of administration considerations in the development of condition specific quality of life scales. In: Cynamon ML, Kulka RA, eds. *Seventh Conference on Health Survey Research Methods.* Hyattsville, Md: US Department of Health and Social Services; 2001.

64. Dillman DA, Sangster RL, Tarnai J, et al. Understanding differences in people's answers to telephone and mail surveys. In: Braverman MT, Slater JK, eds. *Advances in Survey Research.* San Francisco: Jossey-Bass; 1996:110. New Directions for Evaluation, No. 70.

65. Brown JA, Nederend SE, Hays RD, et al. Special issues in assessing care of Medicaid recipients. *Med Care.* 1999;37(3 Suppl):MS79-MS88.

65. Kaegi L. Medical Outcomes Trust Conference presents dramatic advances in patient-based outcomes assessment and potential applications in accreditation. *Jt Comm J Qual Improv.* 1999;25:207-218.

67. Stewart AL, Napoles-Springer A. Health-related quality-of-life assessments in diverse population groups in the United States. *Med Care.* 2000;38(9 Suppl):II102-II124.

68. Morales LS, Reise SP, Hays RD. Evaluating the equivalence of health care ratings by whites and Hispanics. *Med Care.* 2000;38:517-527.

69. Napoles-Springer AM, Stewart AL. Use of health-related quality of life measures in older and ethnically diverse U.S. populations. *J Ment Health Aging.* 2001;7:173-179.

70. King G, Murray CJ, Salomon JA, et al. Enhancing the validity and cross-cultural comparability of measurement in survey research. Am Political Sci Rev. 2004:98:191-207.

71. Torrance GW, Furlong W, Feeny D, et al. Multi-attribute preference functions. Health Utilities Index. *Pharmacoeconomics.* 1995;7:503-520.

72. U.S. Congress Office of Technology Assessment. *Evaluation of the Oregon Medicaid Medicaid Proposal.* Washington, DC: US Government Printing Office; 1992.

73. Wainer H. Estimating coefficients in linear models: it don't make no nevermind. *Psychol Bull.* 1976;83:213-217.

74. von Winterfeldt D, Edwards W. *Decision Analysis and Behavioral Research.* Cambridge: Cambridge University Press; 1986:436-444.

75. Kind P, Hardman G, Macran S. *UK Population Norms for EQ-5D.* York, United Kingdom: Centre for Health Economics, University of York; 1999.

76. Bosch JL, Hunink MG. Comparison of the Health Utilities Index Mark 3 (HUI3) and the EuroQol EQ-5D in patients treated for intermittent claudication. *Qual Life Res.* 2000;9:591-601.
77. Siderowf A, Ravina B, Glick HA. Preference-based quality-of-life in patients with Parkinson's disease. *Neurology.* 2002;59:103-108.
78. Fryback DG, Dasbach EJ, Klein R, et al. The Beaver Dam Health Outcomes Study: initial catalog of health-state quality factors. *Med Decis Making.* 1993;13:89-102.

Contact Information:
Dennis G. Fryback, Ph.D., Professor, Department of Population Health Sciences, University of Wisconsin–Madison, dfryback@wisc.edu.

Prepared for:
Workshop to Develop a Research Agenda and Research Resources for Health Status Assessment and Summary Health Measures, March 26–28, 2003, Alexandria, VA; Edward J. Sondik, Ph.D., and Donald L. Patrick, Ph.D., Cochairs. Sponsored by the Department of Health and Human Services Interagency Working Group for Summary Measures (IAWG).

Acknowledgment:
Much of the basis for this paper was drawn from a 4-year collaboration with Robert Kaplan and Ted Ganiats (University of California, San Diego), Ron Hays (University of California, Los Angeles, and RAND), David Feeny (University of Alberta), and Paul Kind (University of York, England) to produce a coordinated program of projects proposing to study many of the measures discussed here simultaneously. I have greatly benefited from our many discussions.

Commentary

Article: Understanding and Comparing Existing Summary Measures of Health and Health-Related Quality of Life: The State of the Art, By Dennis G. Fryback, Ph.D.

William Furlong, M.Sc.

In his article, "Understanding and Comparing Existing Summary Measures of Health and Health-Related Quality of Life: The State of the Art," Dr. Fryback provides a fine introduction to summary measures of health. He also presents convincing arguments for focusing the research agenda on using existing measures in head-to-head studies and in surveys to establish norms. The proposed agenda is laudable but does not provide much guidance about priorities within the major topics.

This discussion paper focuses on one of the classes of generic instruments: preference-based multi-attribute systems (MAS). It expands on Dr. Fryback's list of instrument design issues to inform decisions about selecting specific MAS for use in head-to-head "cross-walk" studies, population norm surveys, and new clinical studies.

MAS are important because they are practical and provide descriptive health profile and preference-based single summary scores of health-related quality of life. Preference-based summary scores are useful for a variety of health studies and required for cost-utility economic evaluations of health care programs.[1,2] There are five major preference-based MAS: Quality of Well-Being (QWB); Health Utilities Index Mark 2 (HUI2); Health Utilities Index Mark 3 (HUI3); EQ-5D; and Short Form 6D (SF-6D). Dr. Fryback described the differences and similarities among these instruments in terms of numerous factors such as number of questions, number of dimensions, type of dimensions, sensitivity to change, and questionnaire administration. These issues are important. Additional factors also deserve consideration:

- Integrated systems for describing and valuing health status;

- Maximum number of attributes;
- Structural independence of attributes;
- Identification of attributes;
- Meaning of index scores;
- Preference interactions among attributes;
- Utility versus value measurements;
- Combining VAS and SG preference measurement strategies;
- Quantifying preferences for states worse than dead; and
- Generalizability of index scores.

An efficient MAS is most likely to be the product of an initial overall design that included integration of the descriptive and valuation subsystems. The richness of health status description increases with the number of attributes and number of levels within each attribute defined by the classification system. For preference measurements, the maximum number of attributes is quite limited. We know this from research in psychology that has shown most people, because of limitations of immediate memory and span of attention, are able to process a maximum of nine chunks of information at a time.[3] Therefore, only attributes considered important by most people should be included. Furthermore, attributes should be structurally independent so that each attribute contributes unique information.

Index scores should be meaningful, such as representing mean community directly measured standard gamble scores. All major MAS scoring models are fitted to directly measured preference scores collected from members of the general population. However, there is little evidence for the validity of most models in terms of being able to predict directly measured scores. There is considerable evidence to suggest that the underlying structure of preferences for health states is not a linear additive function of preference scores for levels of individual attributes. There is evidence of quantitatively important and statistically significant preference interactions among attributes.[4]

If we accept that health status is intrinsically uncertain and that preference scores for health status should include the effects of uncertainty, then we should use utility scores representing

standard gamble utility measurements rather than value scores from other preference measurement techniques that do not include risk preferences.[5] The internal consistency of preference measurement survey data and the predictive validity of fitted preference scoring models are improved by combining ranking (e.g., feeling thermometer) and standard gamble measurements.[6]

The conventional preference scale for scoring health states is defined such that being dead equals 0.00 and perfect health (as defined by a MAS) equals 1.00. However, there is an increasing recognition that some health states are considered worse than dead by many people.[7,8,9] Furthermore, defining the utility scale to include negative scores, representing states considered worse than being dead, minimizes measurement floor effects and improves the ability of the measure to discriminate among subjects at the lower end of the scale. Thus, it is important for a MAS to quantify people's preferences for states considered worse than dead.

Preference scores for health states vary among individuals. It does not necessarily follow, however, that there are systematic differences among groups of individuals. In fact, there is little evidence across studies of consistent and systematic differences among groups defined by common demographic factors.[10] Differences in MAS index scores are much more likely to be a function of how health states are described, how preferences are measured, and how preference measurements are modeled. This is very good news because it implies that it is appropriate to work towards having one universally accepted scoring model for each instrument, rather than separate scoring models for various populations defined by demographic factors such as gender, race/ethnicity, and nationality.

Importantly, the above design factors vary across the major instruments. This variability is clearly evident when examining the measurement factors related to describing health status. Dr. Fryback notes that all systems have differing numbers of attributes, numbers of levels within attributes, and types of attributes. The combined effects of the first two of these factors are recognizably large when one considers that the number of unique health states is calculated as the factorial of the number of levels for each attribute. The number of unique health states varies by more than three orders of magnitude across the five major MAS measures. It may also be

surprising to note that few MAS measures include attributes for vision and hearing. Less obvious is that there is considerable variability within attributes having the same nominal label. Pain, mobility, and emotion are common across most of the systems, yet the underlying constructs vary. For example, HUI2 emotion focuses on frequency of fretful/irritable/anxious/depressed feelings, whereas HUI3 emotion is based on degree of happiness/unhappiness.[11]

In addition to the variability in how health states are described, there is also important variability among the preference scoring systems.[12] Preference measurements used to fit the scoring models may be utilities or values. Most of the measures use linear additive scoring models that do not include terms for effects of preference interactions between attributes. Two scoring models do not allow for quantification of states considered worse than dead.

Why should we care about these additional issues? These "details" are conceptually important. They can help explain empirical results and inform future research. Results from a head-to-head study by Statistics Canada of data from a sample of the Canadian general population are illustrative. One of the design criteria for the HUI3 health status classification system was that the attributes be structurally independent to maximize the amount of unique descriptive information and to facilitate fitting preference scoring models by ensuring that preference survey respondents could imagine all possible combinations of attribute levels. The Statistics Canada survey results confirm the structural independence by showing little linear correlation between HUI3 attributes (26 of 28 correlations were less than 0.25; maximum correlation was 0.33). On the other hand, the EQ-5D results showed a high degree of correlation between attributes (9 of 10 correlations were greater than or equal to 0.25 and the maximum correlation was 0.64). This indicates that each HUI3 attribute is associated with relatively unique health status information, whereas there is more redundancy among EQ-5D attributes.[13]

The importance of MAS richness, in terms of health status classification and associated scoring, was an explanatory factor in the cumulative frequency distributions of EQ-5D and HUI3 scores from the same Statistics Canada survey. The EQ-5D and HUI3 cumulative distribution curves are very different. A dominant feature of the EQ-5D curve is a very large step at the upper end of the distribution, representing a strong ceiling effect, due to the classification system defining no

health states with scores between 0.88 and 1.00. The HUI3 curve is smooth by comparison because it includes many health states with scores between 0.88 and 1.00.[13] More generally, health status classification systems need to be rich enough to span the full continuum, to describe morbidity for the very sick as well as those very close to perfect health. Floor effects are a threat to the responsiveness of a measure when applied to very ill subjects. Similarly, ceiling effects are a threat to responsiveness in subjects with mild problems.

In conclusion, the research agenda should focus on existing MAS measures. These systems vary substantially and are not equal. The devil is in the details! There are limited resources for including summary measures of health in major surveys, and only the most efficient measures should be proposed. The idea that crosswalk equations be used to convert results from one measure to another is enticing for use with existing data, but it is fraught with limitations. We should keep in mind that while it may be a sensible expectation to map information collected using a richly detailed system into a less rich system, the opposite is more problematic. A MAS should be able to describe and quantify health states considered worse than dead. Index scores should have a relatively simple interpretation, such as representing mean community preference scores for well-defined health states. There are good reasons why preferences for health should be measured under conditions of uncertainty using the standard gamble and scoring models should include effects of interactions among attributes. Measures for new applications should be selected on their individual merits. Replications, using exactly the same and not revised methods, of preference measurement surveys are required to assess the generalizability of scoring models for existing measures in the quest for a universal index of each MAS.

References

1. Feeny D. Preference-based measures: utility and quality-adjusted life years. In: Fayers P, Hays R, eds. *Assessing Quality of Life in Clinical Trials: Methods and Practice.* 2nd ed. New York: Oxford University Press; 2005:405-429.
2. Hawthorne G, Richardson J. Measuring the value of program outcomes: a review of multiattribute utility measures. *Expert Rev of Pharmacoeconomics Outcomes Res.* 2001;1(2):215-228.

3. Miller GA. The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychol Rev.* 1956;63:81-97.
4. Feeny D. The Health Utilities Index: a tool for assessing health benefits. *PRO Newsl.* 2005;34:2-6.
5. Torrance GW, Furlong W, Feeny D. Health utility estimation. *Expert Rev Pharmacoeconomics Outcomes Res.* 2002;2(2):99-108.
6. Torrance GW, Feeny D, Furlong W. Visual analog scales: do they have a role in the measurement of preferences for health states? *Med Decis Making.* 2001;21:329-334.
7. Patrick DL, Starks HE, Cain KC, et al. Measuring preferences for health states worse than death. *Med Decis Making.* 1994;14:9-18.
8. Feeny D, Furlong W, Torrance GW, et al. Multiattribute and single-attribute utility functions for the Health Utilities Index Mark 3 system. *Med Care.* 2002;40(2):113-128.
9. Dolan P. Modeling valuations for EuroQol health states. *Med Care.* 1997;35(11):1095-1108.
10. Furlong WJ. Variability of Utility Scores for Health States Among General Population Groups. Hamilton, Canada: McMaster University; 1996. 140 pp. M.Sc. thesis.
11. Feeny D, Furlong W, Boyle M, et al. Multi-attribute health status classification systems. Health Utilities Index. *Pharmacoeconomics.* 1995;7(6):490-502.
12. Furlong W, Barr RD, Feeny D, et al. Patient-focused measures of functional health status and health-related quality of life in pediatric orthopedics: a case study in measurement selection. *Health Qual Life Outcomes.* 2005;3:3.
13. Houle C, Berthelot JM. A head-to-head comparison of the Health Utilities Index Mark 3 and EQ-5D for the population living in private households in Canada. *Qual Life Newsl* 2000;24:5-6.

Contact Information:

William Furlong, M.Sc., Health Utilities, Inc., furlongb@mcmaster.ca.

Acknowledgment:

William Furlong is a Research Associate in the Department of Clinical Epidemiology and Biostatistics at McMaster University and a developer of the Health Utilities Index (HUI). He has a proprietary interest in Health Utilities Incorporated (HUInc). HUInc distributes copyrighted HUI materials and provides methodological advice on the use of HUI. Thanks to Dr. David Feeny and Dr. George Torrance for review comments on a draft of the manuscript.

The Ten Ds of Health Outcomes Measurement for the 21st Century

Colleen A. McHorney, Ph.D.

Introduction

The origins of health status assessment can be traced to the 1960s and the need at that time for a new armamentaria of health statistics to measure outcomes above and beyond mortality and morbidity. The state of health outcomes assessment in the 1960s has been characterized by the Five Ds: death, disease, disability, discomfort, and dissatisfaction.[1] In the United States, death registration was standardized in most states by 1930,[2] and disease surveys had been under way since the late 1880s.[3–6] Measurement of disability began in the 1930s[7–9] but earnestly gained momentum in the late 1950s.[10–12] The National Health Interview Survey, which is a major source of information on disease and disability, was instituted in 1957[13] and continues today. Measurement of discomfort (subjective and objective sickness impacts) began in the 1940s[14–15]; and continues to constitute a significant component of quality of life (QOL) surveys. Measurement of patient satisfaction commenced in the 1950s for mental health care[16–17] and the 1960s for general medical care.[18]

We have made great progress in measuring patient health outcomes since the Five Ds were first propounded. There are over 85 tools that measure basic and instrumental activities of daily living.[19] Myriad measures of depression exist.[20] Close to two dozens generic QOL instruments have been developed.[21] Hundreds of disease-specific instruments abound.[22–23] In cancer, over 75 different QOL measures exist.[24] The vast majority of these measures have been created under the umbrella of classical test theory (CTT). CTT is a set of assumptions and procedures that has been used to develop tests for much of the 20th century.[25-26]

There are a number of problems that arise in developing tests and in using test scores that CTT cannot overcome. First, the statistics used to describe item performance are dependent on the particular sample of respondents in which they are calculated.[26–27] Thus, items and scales will

have different statistics if the measured samples do not have similar distributions of ability. Second, CTT is *test*-driven rather than *item*-driven. Different sets of items will differ in difficulty and will provide different estimates of ability. Unless the tests have been equated, the scores of respondents taking one test cannot be compared to those of respondents taking another. To date, the bolus of health status measures have rarely been equated,[19, 28] and thus have not been placed on a common metric of ability.

In recent years, item response theory (IRT) has been increasingly used in health outcomes assessment.[21, 29–31] IRT is a theoretical framework and a collection of quantitative techniques for test construction, scaling, and equating, as well as for identifying item bias and supporting computerized adaptive testing (CAT). If its assumptions are met, IRT can overcome some of the limitations of CTT by providing item parameters that are theoretically invariant with respect to the sample of examinees and ability parameters that are theoretically invariant with respect to the set of items used.[27, 32, 33]

Given the simultaneous growth of patient-centered measures and the increasing use of IRT in health outcomes assessment, it is timely to take stock of our measurement progress. The purpose of this chapter is to suggest advances for the *process* of outcomes measurement in distinctive and informative manners. As a way of doing so, we propose and discuss the Ten Ds of health outcomes measurement for the 21st century.

1 Definitions of Quality of Life

Two seminal pieces of work have lamented the state of conceptual affairs in health outcomes assessment. Gill and Feinstein[34] argued that no unified approach has characterized QOL assessment and that there is little conceptual agreement on exactly what QOL means. They found researchers to be deficient in defining QOL and in justifying the selection of QOL instruments. The same criticism applies to instrument developers. Leplege and Hunt[35] have elegantly argued that:

> …a clear conceptual basis for quality-of-life measures is lacking, and the few attempts to develop models or operational definitions of quality of life have been

> woefully inadequate… It is difficult to progress in any field if there is no shared definition of the concept or phenomenon under study…assessment of quality of life measures too often has been based on arguments of authority rather than on rational debate.

Disease-specific measures often have biomedically driven measurement models based upon known or hypothesized manifestations of the underlying pathology. Conceptual frameworks for generic measures can generally be characterized as insubstantial, usually attributing conceptual models to the WHO[36] trinity of physical, social, and mental health. Yet, the WHO definition is just that—only a definition (not a conceptual framework) and one that is at the same time both vague and idealistic.[37] Unfortunately, adherence to the WHO definition has led researchers to implicitly ignore what they seek to measure—the patient point of view. Both Gill and Feinstein and Leplege and Hunt take issue with the fact that many QOL measures do not ask patients what is important to them. As Gill and Feinstein[34] argue:

> …quality of life can be suitably measured only by determining the preferences of patients and supplementing (or replacing) the authoritative opinions contained in statistically "approved" instruments.

Leplege and Hunt[35] resonate with this assessment by contending that:

> …there has been some confusion between questionnaires that are completed by patients and those that reflect the concerns of patients… most of the currently used questionnaires do no more than force patients to address themselves to the concerns of physicians and/or social scientists and statisticians.

We are in a transition phase from fixed-length, one-size-fits-all assessment to more tailored assessment. This transition will involve the construction and calibration of item banks and the use of IRT and CAT to assess QOL.[19, 21, 28] As we make this transition, we should not repeat the mistakes we have made over the past 30 years by having unbridled promulgation of myriad item banks, myriad redundant banks, and myriad redundant banks that, as Gill and Feinstein assert, possess psychometric elegance but are far removed from the subjects they purport to assess—the individual patients.

2 Discovery Methods

The thin conceptual cornerstone that has characterized many health status measures may have impelled instrument developers to use existing items rather than develop them *de novo*. Only two generic measures (the Sickness Impact Profile [SIP] and the Nottingham Health Profile) obtained their items from consumers themselves. Otherwise, items for generic measures have been recycled from the literature, often gleaned from clinically oriented tools.[38] Disease-specific tools derive their items from three sources: (1) existing generic tools[39–41]; (2) clinical expertise;[42] and (3) patient testimony about the impact of disease and treatment on health status.[43–45]

Use of existing items to construct QOL surveys has benefits and drawbacks. As to the former, one can theoretically select items with desirable psychometric properties. In reality, however, item characteristics are a combination of the item itself and the group in which it is tested.[21, 46, 47] As to the latter, many older items violate contemporary standards for item writing insofar as they often contain multiple attributions (e.g., do you have difficulty *bending, kneeling, or stooping*, or how much of the time have you been in firm control of your *behavior, thoughts, emotions, feelings*?). Cognitive interviewing has revealed the sources of invalidity that these practices can yield.[48, 49] Older, recycled items often have antiquated language. For example, in the SIP, there is an item "I get sudden frights," a phrase which is not common today.

We should be true to our intent—to measure patient-centered outcomes—and that means using patients and their caregivers as active participants in the item generation, selection, and pretesting phases of instrument development. Patients should be used to generate items, whether it be through semi-structured interviews,[50–53] ethnography,[54, 55] phenomenology,[56–58] existentialist methods,[59] or the more efficient focus group approach.[43–45, 60–62] Importance ratings provided by patients[63–67] should be used in item selection and reduction to complement psychometric criteria. Pretesting needs to be more patient-centered, with greater use of cognitive testing methods[68–72] in addition to standard psychometric analysis. There are two measurement applications for which discovery methods should prove critical in the years ahead.

First, the field has given scant attention to differential item functioning (DIF) (see D # 6). The purpose of DIF analyses is to identify items that exhibit dissimilar response patterns for persons having equal ability but different group membership (such as age, gender, race, ethnicity, etc). Once DIF is identified, one needs to discover the potential causes of the DIF. Psychometric methods do not easily lend themselves to such discovery research. Qualitative researchers, however, have a repertoire of methods that are better suited to understand the reasons *why* persons of equal ability answer questions differently. Thus, discovery methods should prove useful in understanding the myriad reasons for DIF. Once identified, DIF can be corrected at the item writing stage.

Second, item banking and CAT are on the measurement horizon.[21] CAT requires the calibration of a large bank of items. Items can be assembled through expert opinion or by cutting and pasting from the literature. However, many published items themselves were constructed by cutting and pasting from even earlier measures, thus perpetuating a long lineage of items that may have lost their relevance, salience, or discriminability over time. It would be appropriate, indeed essential, to inform the development of item banks by using discovery methods to obtain from patients and consumers new items to fill in known or hypothesized gaps in the functioning and well-being continuum as well as to use the patient point-of-view to help eliminate or cull out redundant items.

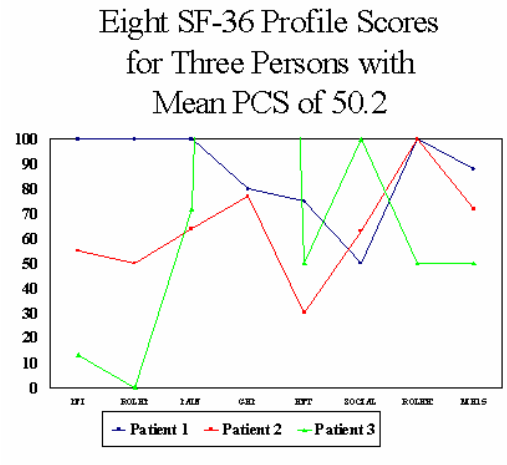
3 Dimensionality

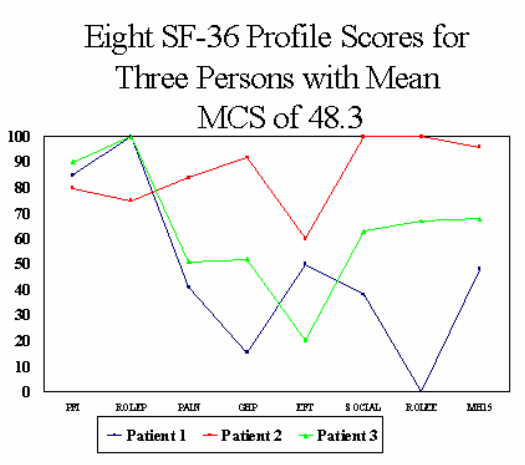
Unidimensionality is an important, but under-discussed, aspect of health status assessment. Unidimensionality concerns the extent to which the ability being measured is a single unitary trait or dimension. Unidimensionality is important for score interpretation. If a score is composed of more than one dimension, it is impossible to determine what is contributing to the score.

The development of superordinate summary scales, like those for the SIP,[73] the SF-36,[74] and other measures,[75–80] directly challenges the property of unidimensionality by factor scoring diverse profiles into a limited number of composite scores, often deriving orthogonal structures along the way. In this scoring scheme, a good "physical health" summary score is achieved by

having high scores on physical scales (e.g., good physical and role functioning and no pain) and low scores on mental scales (e.g., anxious and depressed). The same applies to the summary mental scale, whereby a high score is achieved by having good mental health status but poor physical health status. This orthogonalized scoring practice has been criticized on methodological and conceptual grounds.[81, 82]

Summary scales have been derived, in part, to address problems related to multiple comparisons in hypothesis testing.[29, 83] Despite the desire for parsimony in hypothesis testing, the use of superordinate summary scales can often complicate inferential testing more than it simplifies it. Numerous studies have reported that summary scores yield quite different substantive findings than their more unidimensional profile scales.[81, 82, 84–98] These discrepancies occur because different scoring weights (often derived from factor analytic procedures) are given to individual profile scores. When combined into a superordinate score, the effects of individual, substantively meaningful profiles can be blurred, obscured, underestimated, or overestimated because of the sign and direction of the scoring coefficients. As an example, the figures below show the eight SF-36 health profiles for three persons with the normative mean score of 48.3 for the PCS and 50.2 for the MCS.


Eight SF-36 Profile Scores for Three Persons with Mean PCS of 50.2
100
90
80
70
60
50
40
30
20
10
0
PFI
ROLEP
PAIN
GEN
EFT
SOCIAL
ROLEE
MHIS
Patient 1
Patient 2
Patient 3


Eight SF-36 Profile Scores for Three Persons with Mean MCS of 48.3
100
90
80
70
60
50
40
30
20
10
0
PFI
ROLEP
PAIN
GEN
EFT
SOCIAL
ROLEE
MHIS
Patient 1
Patient 2
Patient 3

As the Figures show, persons with the same exact score at the sample mean level have dramatically different individual health profiles. QOL measures should inform, rather than complicate or obfuscate, studies of treatment effectiveness. If we cannot attribute an observed treatment effect (i.e., whether it is due to deterioration in pain or improvement in anxiety), then we cannot confidently use QOL findings to support labeling and promotional claims for medical procedures or devices or pharmaceutical products. If we cannot attribute QOL effects in clinical trials, then those data will not be of measurable use to physicians when advising their patients on alternative therapies. If we cannot attribute QOL effects in treatment studies, then policy makers and payers will not consider such data in policy development.

The final issue about unidimensionality is that is it an underlying assumption for IRT analysis. Investigators utilizing IRT need to assess unidimensionality, preferably using more than one method. Unidimensionality is also a strong requisite for item banks and, thus, CAT.[99] Unidimensionality also plays a significant role in studies of DIF because items could be statistically flagged as being DIF if they are multidimensional. Thus, for both old and new applications of measures, unidimensionality assessment needs to become a more standardized aspect of instrument development and validation.

4 Disparities and Determinants

One of the national health goals in Healthy People 2010 is to eliminate health disparities.[100] NIH defines health disparities as "differences in the incidence, prevalence, mortality, and burden of diseases and other adverse health conditions that exist among specific population groups in the United States."[101] Health status and QOL fall squarely under the burden-of-disease umbrella.

Research over the past 40 years has consistently underscored the role of non-medical factors in determining individual and population health.[102, 103] It is now well-established that social, lifestyle, and psychological factors account for 50% of preventable morbidity and mortality, environmental factors and human biology account for another 20% each, and medical care

accounts for only 10%.[104] The unique signature of the outcomes movement is that it broadened the scope of our dependent variables to include functioning, well-being, and patient satisfaction in addition to more traditional indicators of mortality, morbidity, and costs. A shortcoming of the outcomes movement is that it has implicitly adopted the medical model's reductionistic view of health because it has focused on what works in *medicine*—thus, the independent variables are the same ones we have been studying for years. As a result, our knowledge base is rich in terms of the impact of disease, severity, comorbidity, symptoms, and treatment on health status and QOL, but less so in terms of the non-medical determinants. Further, clinically driven outcome research has tended to view health status in an episodic manner, with most studies being cross-sectional or with limited longitudinal designs. However, health status and QOL are dynamic phenomena that change in response to aging, illness adaptation, treatment, and natural history. Thus, future research needs to address the life-course character of health status above and beyond disease and treatment episodes.

Individuals and populations vary greatly in health status when they are free of pathological disease[105] as well as when they are matched on pathophysiologic disturbance. This occurs because social factors (e.g., socioeconomic status, stress, environment, etc.) exert an important influence on health status.[103] As Leplege and Hunt[35] argue:

> …the concept of health-related quality of life implies that people can analyze quality of life into its health and non-health related components. This view fails to acknowledge the interconnectedness of health status with other aspects of existence.

Since QOL is the illness-impact iceberg underlying disease, morbidity, and disability, future research needs to expand its explanatory potential by studying health determinants vis-à-vis QOL outcome measures. In part, this could be accomplished in the context of validity assessment. Generic measures are intended for use across population segments. Thus, one important validity test should be whether they exhibit the same patterns of social differentials (by age, gender, race, and socioeconomic status[106–113]) as have been observed with mortality and morbidity. Put differently, generic measures should reflect predominant social patterns of inequality in health, especially if they are to be used at the population level for planning and evaluation purposes. The

same case can be made for disease-specific measures. The co-modeling and co-presentation of social with clinical variables will help to situate the relative importance of both determinants. More profound analysis of social variables will contribute to a deeper and more meaningful understanding of health determinants.

5 Disadvantaged Populations, Disenfranchisement, and the Digital Divide

Relatively few health outcomes assessment tools have been assessed in disadvantaged or vulnerable populations.[114] That is, most reliability and validity studies have been conducted in white, middle-class populations. Few investigations have even superficially assessed the myriad conditions under which measures may become degraded in disadvantaged populations. Such psychometric ethnocentrism is regrettable because the United States is becoming a more diverse society. The population is aging, and the United States is becoming more ethnically diverse.[115] Because of the growing cultural pluralism, there is more than ever before a need for evidence that health outcome tools exhibit measurement equivalence across diverse population groups.[114]

For group comparisons to be meaningful, one must establish that the variable(s) measured in different groups are parallel enough to be considered as the same behavior, attitude, symptom, or feeling.[116] Tests of *psychometric equivalence* (determination of whether the derived scale provides equivalent measurement across groups) are often undertaken. However, they vary in the breadth and depth of analysis, ranging from simple group comparisons of means or Cronbach's alpha coefficients to structural equation modeling of factor structures to IRT-based analyses of DIF. Most investigators assess psychometric equivalence and then assert conceptual equivalence if no psychometric differences are found. This "absolutist approach"[117] begs the question of whether the construct under investigation is meaningful and relevant across different groups.

C*onceptual equivalence* involves assessing whether the construct under consideration has identical meaning, relevance, and significance across groups. It could be that the feelings or behaviors assessed are differentially salient across groups, that the selected items only partially represent the construct as defined by a group, or that some experiential aspects of the construct are omitted altogether.[114, 118] Studies of conceptual equivalence should assess the extent to which

the operationalization of the construct and the specific items used to represent the construct are portable across groups.[119, 120] Assessing conceptual equivalence ideally involves qualitative discovery research to gain knowledge of people's vocabularies and terminology and to understand the attributions or qualities they assign to feelings and behaviors.[121]

Disenfranchisement can occur through several mechanisms, such as failing to ensure conceptual equivalence and cultural appropriateness of an instrument, failing to adequately sample disadvantaged and vulnerable respondents, and failing to provide respondents with a user-friendly mode of administration. The reading level of instruments must be appropriate for poor readers. The field has been moving toward computerized assessment. For example, computerized surveys have been used for preoperative testing[122] and mental health assessment,[123–128] and computerized QOL assessment is increasingly used.[129–138] However, the nature and extent of the digital divide by age, race, and socioeconomic status is well documented.[139, 140] Accordingly, we need to ensure that disadvantaged and vulnerable patient groups are not disenfranchised from computerized outcomes assessment, including the anticipated use of CAT for health outcomes assessment.

6 Differential Item Functioning

Identification and correction of DIF items has a long history in achievement and educational testing. If items in an achievement test (or qualification, promotion, or certification tests) are answered differently by women versus men, or minorities versus majorities, when their underlying ability is the same, then the test scores would not be comparable and educational placement decisions would unfairly hurt one group and unfairly favor another. If items in such tests are biased, then inequitable treatment may likely result, thus materially affecting the lives of the test takers.[141] In educational and achievement testing and professional credentialing, item writers and instrument developers purify their items *a priori*.[142] Unfortunately, in health outcomes assessment, instrument developers have tended to ignore DIF or have identified it long after the measure has been in use.

DIF has been identified in a large number of health assessment tools, including measures of functional status[19, 143–150] cognitive status,[151–154] QOL,[155] satisfaction,[156, 157] and many mental health and personality measures.[118, 158–178] Across all of these studies, DIF has been identified by age, gender, race, ethnicity, socioeconomic status, language, and nationality. DIF has been large enough to cause meaningful shifts in group means or case rates when DIF items are removed from the scale.[143, 144, 159, 161, 165, 172, 173, 175, 177, 179, 180]

Identifying, understanding, and correcting DIF is fundamental to developing assessment instruments, to testing hypotheses, to theory building, to screening and diagnosing individuals, and to implementing and evaluating health service delivery programs. Culturally fair health outcomes assessments (with "culture" defined broadly as gender, age, racial, ethnic, socioeconomic, geographic, and language variations[181]) are crucial when individual decisions are in balance, such as with mental or physical health screening and diagnostic, placement, and referral decisions. If items in a health assessment instrument are biased, detection rates will be biased (overestimated or underestimated), leading to over- and under-detection and over- and under-treatment. Item bias in health outcomes assessment tools can have implications at the policy level (e.g., under- and over-utilization of health services, erroneous prevalence rates) and at the individual-patient level.

Research on DIF will not "throw the baby out with the bath water." DIF items have been identified in many health assessment tools and will surely be identified in others. Such identification in and of itself will not call for the mass abandonment of current assessment tools. Rather, advances in DIF identification and amelioration will help to polish current instruments, to iron out measurement kinks,[182] so that current and future assessment tools become more culturally applicable and fair across the board.

7 Item Difficulty

Item difficulty gets its name from the educational context in which IRT was developed. In such contexts, it is common to think of some items as "harder" or "easier" than other items. In scaling medical outcomes, however, the parameter name often does not communicate well. An

alternative is to think of the difficulty of an item as the item's *intensity*. On a headache pain scale, for example, the item, "my headaches kept me from being productive at work: yes/no" is more intense (harder to endorse) than, "my headaches caused me little interruption in my daily activities: yes/no." All IRT models estimate an item difficulty parameter.

As described elsewhere,[21] the field has been entrenched in a paradigm of psychometric efficiency over the past decade, with an emphasis on constructing measures with as few items as possible. Acceptable standards of reliability with few items can best be achieved by selecting items that are fairly homogeneous. Thus, items are often selected that are in the middle range of item difficulty and that are near alternate forms of one another. There is one major consequence of this measurement standard: the endpoints of the health continuum tend to be poorly defined, yielding substantial ceiling effects.[21] Score imprecision has two principal consequences. First, it is impossible to distinguish among individuals at the ceiling, even though they likely vary in the underlying construct. Thus, ceiling effects paint a more favorable image of population health than is true. For researchers, ceiling effects produce Type II errors in hypothesis testing. For clinicians, ceiling effects yield false-negative outcomes. Second, it is impossible to measure improvement in health over time for those at the ceiling. Thus, score distributions that are skewed at baseline will underestimate (or miss) the effects of effective treatment or natural history on health status.

The most common source of imprecision is the selection of items whose difficulty is incongruent with the *ability* of the population of interest.[21] Simply put, floor and ceiling effects derive from a poor marriage between the difficulty of an item and the ability of the targeted population. Ceiling effects occur when easy items are administered to high-ability populations, and floor effects happen when difficult items are administered to low-ability populations. Few positive well-being scales exhibit ceiling effects,[183–185] which is the result of two factors. First, they are often composed of items having multiple categorical rating points (five to seven response categories). Compared to dichotomous items, polytomous items provide information across a broader range of the measurement continuum and more precisely differentiate individuals on the underlying construct. Second, positive well-being measures often have balanced items (those tapping

negative and positive health states). Items that tap positive health states tend to have very low ceiling effects (i.e., few individuals report that they are a "happy person" all of the time).

Thus, problems with precision pertain largely to measures of physical, role, and social function. How can we "raise the bar" for the measurement of function? Recent work on calibrating basic and instrumental activities of daily living[19, 28, 147, 186, 187] has indicated obvious redundancies in measuring lower-level functioning and conspicuous gaps in measuring higher-order functioning. This is a clear beacon for future measurement development. Even for elderly populations, it is not necessary to oversample lower-level functioning because many items are redundant in terms of item difficulty and item discrimination.[19, 28, 147, 149, 186–192] The challenge is to more effectively sample and distribute the lower-level items while concurrently adding items to fill in known gaps at the difficult end of the continuum (e.g., higher-order functioning, productive activities, executive functioning, leisure exercise, and physical fitness). These are activities that will raise the ceiling while also being consistent with national health objectives and public health recommendations.[100]

The challenges for future advances in functional status assessment are both conceptual and methodological. Conceptually, qualitative methods should be used to glean from consumers themselves facets of contemporary functioning. For example, Porter[57] discovered numerous nuances about ADL performance in the context of qualitative research. The same applies to role and social functioning. The content of these concepts has been fairly narrow to date. A combination of focus group, diary, and time-use methods might yield valuable insights into what types of basic, intermediate, and advanced activities are performed on a regular basis, as well as what types of activities are abandoned and in what sequence. Further, qualitative methods would be useful to understand how people adapt to occult or incipient disability. A crucial area for future research is to more profoundly understand the physical, economic, and social compensatory strategies used by the elderly in maintaining independence. Future advances in functional assessment might benefit from developing rating scales that tap compensation rather than difficulty per se.

Methodologically, a better way of matching item difficulty with the ability of the targeted population is needed. If we want to know how well persons are with respect to function, the most efficient procedure is to ask them about activities that are close to their level of ability. What is required is some means of functionally relating performance on each test item to person ability. IRT is well designed for this purpose in the context of item banking and CAT. Thus, a logical extension for health status assessment is to move from pen-and-paper tools to computerized-adaptive assessment of health status.[21] Computerized health status assessment could (1) reduce the human capital involved in administering and scoring questionnaires; (2) challenge patients at their targeted level of ability instead of boring or discouraging them; (3) provide researchers with the exact amount of precision they require for each patient sample and each specific application; and (4) provide "real-time" scores to clinicians for use at the individual patient level in clinical practice.

Development of an adaptive framework would require four phases of methodological work.[21] The first task would be to assemble item banks on different health concepts (concept-specific banks). The second task would involve conducting cognitive interviews with a variety of patient groups to obtain in-depth information about respondent understanding and acceptance of the banked items. The cognitive interviews could also obtain input from patients/consumers on gaps in content coverage in each underlying continuum. The third task would be to employ IRT to calibrate items and to select a subset of items that comprehensively, and evenly, tap the underlying construct of interest. The final phase would be to develop and implement algorithms for adaptive testing (e.g., starting and stopping rules).

8 Item Discrimination

Item discrimination refers to an item's ability to distinguish among individuals who have different levels of the trait being measured. In the headache example above, we would expect the items to discriminate between those with severe headaches and those with mild headaches. A group of respondents with severe headaches should be more likely than those with mild headaches to endorse the item, "my headaches kept me from being productive at work: yes/no." If this is not the case, the item is a poorly (low) discriminating item. In polytomous IRT models,

the discrimination parameter is related to the item characteristic curves (ICC) for an item. The ICCs are the functions obtained by plotting the probability of scoring in a particular response category against the latent trait being measured. For very discriminating items, the probability function rises sharply; for low discriminating items, the function remains relatively flat across the measured continuum.

The discrimination parameter is related to another important construct in IRT, *information*. Information is defined as the reciprocal of the square root of the standard error of measurement. More highly discriminating items yield greater information and have smaller standard errors than lesser discriminating items. In the development of a scale, low discriminating items tend to be deleted from the pool of potential items since such items do not "cooperate" well in the measurement of the trait of interest. This is roughly analogous to deleting items that do not correlate with their hypothesized construct of interest (item convergent validity).

In Rasch models, items are assumed to have equal discriminations. If the data do not fit this assumption, estimates of the information functions and standard errors of measurement can be artificially inflated or deflated depending upon whether the value of the assumed discrimination is an overestimation or underestimation of the actual discrimination of the item. Items with unequal discriminations tend to be identified as "misfitting" items in a Rasch model and, therefore, would be dropped from the developmental item pool. Items with common item discrimination values would exhibit better model fit and, therefore, be more likely to be retained in the final Rasch-based scale.

The heterogeneity of the construct being measured impacts the value of the discrimination parameter. Items in a scale are a sample of the hypothetical population of all items that could be chosen to measure the construct. The selected items should adequately represent the domain being measured. The more narrow the domain, the more homogenous item discriminations will be and vice versa. It is often difficult to distinguish item heterogeneity due to the breadth of the domain sampled versus item heterogeneity due to multidimensionality. However, the distinction is important because the most widely used IRT models assume the measurement of a unidimensional construct. Researchers who favor the use of Rasch models argue that high

discrimination values result from multidimensionality and, therefore, indicate items inappropriate for scaling using unidimensional IRT models. It also is conceivable that the Rasch requirement that items equally discriminate could inadvertently restrict the content coverage of the set of items that compose a scale.

As discussed earlier, to be useful, an item bank must contain items that differ in difficulty. However, items with good discrimination are also desired to differentiate between persons close together on the ability distribution (i.e., to yield more information about persons of seemingly contiguous ability). Thus, a challenge for compilers of item banks is to write discriminating items. Ambiguity can degrade an item's potential discrimination.[28] In one of our studies,[28] highly discriminating functional status items (e.g., put underclothes on, move between rooms, take pants off, get into bed) were almost behavioral measures—they targeted daily activities that were specific, explicit, and unequivocal. In short, they were questions that respondents could understand (because they were simple and concrete) and evaluate with respect to their range of function because they were in the realm of daily experience. Improvements in item writing efforts may be facilitated by scrutiny of low and high discrimination items.[163, 193–195] Also, adherence to conventional item writing standards[196, 197] (such as write items that can only be interpreted in one way; use clear, simple, direct language; and avoid multiple attributions) may go far towards improving item discrimination.

9 Dispute and Divisiveness

A researcher who chooses to use IRT instead of CTT in the measurement of health outcomes has many IRT models from which to choose. For the outcomes researcher new to the use of IRT models, the heat of the debates regarding model selection can come as a surprise. At the extremes, there exist two "camps," one comprising those who favor the one-parameter (1-pl) Rasch model. In the other camp are those who favor the two-parameter (2-pl) models.

The debates in IRT model selection center on whether an item discrimination parameter is estimated or not. Those who favor Rasch models argue that theoretical considerations as well as empirical ones should govern the choice of IRT models.[198] They claim that the Rasch model

obeys "the rules of measurement."[199] An example of this "obedience" is the fact that, with Rasch models, persons who have higher raw scores also have higher calibrated scores. This is not necessarily the case with the 2-pl models. On a 5-item scale with 3 response categories for each item, a person may obtain a given raw score in many different ways. Response strings of "3, 2, 2, 1, 1" and of "3, 1, 2, 1, 2" both yield a raw score of "9." If the scale were calibrated using a 1-pl model, both response patterns would yield the same calibrated score. If the scale were calibrated using a 2-pl model, however, this would not necessarily be the case, because the discrimination of the items would be factored into the computation of the calibrated score.

In scale construction using the Rasch model, emphasis is placed upon finding data (items) that fit the model. For proponents of the 2-pl camp, emphasis is placed on finding a model to fit the data (items). Proponents of the 2-pl models note that, within medical outcomes, item discrimination can vary substantially.[28, 151, 158, 163, 166, 172, 174, 194, 195, 200–207] They argue that the Rasch approach is too "simplistic" to model the kinds of measures frequently encountered in outcomes research. From this perspective, the Rasch criteria for items is too selective, and, therefore, too many of the data (items) are "thrown away" because they do not fit the model.

It is in scale construction that fundamental differences between the 2-pl and 1-pl camps become particularly evident. In response to the question—How should a latent trait be measured?—the 2-pl camp's answer is statistical in its approach—*model the data; don't force the data to fit the model*. For the Rasch camp, the answer more closely follows the approach used in tool development and quality assurance—*build an instrument with the properties most desirable for measurement*.

Despite the heat of the arguments from both camps, the arguments of neither side are conceptually pure. Two illustrations suffice. With the partial-credit model,[208] reversals in calibrated step difficulties occur; that is, a person higher on the trait being measured can be more likely than a person lower on the trait to endorse an easy item category. Such a reversal would appear to be counter to "the rules of measurement." Among those in the 2-pl camp, indignation is sometimes offered regarding the Rasch approach of "throwing away data." No similar objection is made to discarding items that fail to load on the desired factor in a factor analysis.

For medical outcomes researchers trying to root through the arguments between measurement camps, it may helpful to recall that all models, by definition, are wrong. The questions of practical importance to the medical outcomes researcher are as follows: "How wrong are our models?" "What is the impact of specific kinds of 'wrongness'?" In other words, how robust are IRT models in health outcomes applications? These questions have yet to be addressed adequately.

The selection of an IRT model should be supported by careful consideration of the measurement application. We suggest two applications, one in which a 2-pl model would be the more appropriate and another in which a 1-pl would be more appropriate. A way in which IRT models have been applied in outcomes research is in the development of new measures. In the psychometric tradition, a large pool of items is developed and administered to a sample of respondents. The pool is refined based on factor analysis, measures of inter-item consistency, and estimates of item-to-total correlations. The scale developer could also choose to only select items that are homogenous with respect to discrimination, that is, items that fit a Rasch model. Fitting to the Rasch model may provide advantages. As discussed above, the Rasch model has some desirable measurement properties that are particular to it. Also, because there are fewer item parameters, stable parameter estimation can be achieved with smaller sample sizes. Before settling on a Rasch model for a scale's calibration, however, the scale developer should verify that the selection of homogenously discriminating items has not deleteriously affected the content coverage of the items. Content coverage affects the construct validity of the measure being developed and should be privileged over parsimony in the selection of an IRT model.

Another way in which IRT has been applied is in the evaluation and/or equating of well-known and often-used measures.[19, 28, 209, 210] For such applications, the pool of items has already been selected, and the onus is on the IRT model to adequately estimate the items properties. Within health outcomes assessment, items can vary substantially in discrimination.[28, 151, 158, 163, 166, 172, 174, 194, 195, 200–207] Therefore, except in the improbable case in which the set of pre-existing items happens to have equal discrimination, a 2-pl model would be the better choice. A 2-pl model might also be preferred in CAT application. The promise of CAT is to achieve maximum

information with as few items as possible, conditional on the desired precision of the obtained ability estimates. The efficiency of CAT increases as item informativeness increases,[211] and item information is directly related to item discrimination.[205, 212, 213]

10 Debate

The intellectual and technical infrastructure for item banking, CAT, and test equating under IRT is at hand. What is not clear is whether it is desirable for health outcomes assessment to move toward item banking. Health status and QOL assessment can be both praised and faulted for the number of tools that have been generated in the last 35 years. There have been both innovations and repetitions. The same, of course, could apply in the future to item banks, where investigators argue about the extent to which "my bank is better than yours." Applications of IRT in education have been led by world-renowned scholars in measurement at the Educational Testing Service (ETS), a not-for-profit enterprise. Because it is not-for-profit, conflict of interest due to profit motive is less salient. As movements begin in health outcomes assessment toward the development of item banks, linking studies, and CAT,[21, 30, 214–216] earnest thought will need to be given to whether profit motive will corrupt or enhance the measurement developments that are on the cusp.

A bank is a composite of the work of hundreds of individuals over time. For the most fair and productive use, it would be desirable to have health banks reside in the public domain, since their constituent parts were developed with public monies in one form or another. Health banks could reside with the National Center for Health Statistics or the Agency for Healthcare Research and Quality, or they could be operated by a non-profit organization, similar to the ETS. Regardless, item banks require regular attention in terms of retiring items that become outdated or obsolete or whose item parameters change over time. New items need to be added to the bank in response to natural history and would need to be linked into the bank and calibrated. The population invariance property of IRT makes it possible to update item parameters using different samples of examinees.[217]

In health status assessment, where there exist over two dozen generic measures[21] and hundreds of disease-specific measures,[22, 23] dialogue among measurement specialists has sometimes resembled childhood fisticuffs with claims of "my tool is better than yours." However, at least these dialogues have taken place within the context of peer-reviewed science. If measurement developments move from the halls of academia to the private sector, we may continue to hear the polemic of "my item bank is better than yours," but without the safeguard of peer review. As Shapiro[218] argues, privatization can be problematic in that it results in a loss of openness among scholars, a failure to completely disclose the methods and results of research, and a tendency to not publish at all or to only publish results that make the "product" look good.

Conclusion

Much has been accomplished in health assessment and QOL assessment in the last 40 years. Measurement specialists are at the cusp of a paradigm shift[21] away from sizable reliance on classical test methods to broader use of IRT methods. There is much to be both excited and cautious about as IRT methods are used for test construction, scaling, and score equating, as well as for identifying item bias and supporting functions such as CAT. It may be desirable to reach consensus among stakeholders—methodologists, users, policy makers, and funders—about the relative merits of any alternative course that outcomes measurement could assume in the years ahead before any one road is definitively taken. I offer the 10 Ds herein as a platform for informing and stimulating discussion about how and where measurement advances might proceed.

References

1. White KL. Improved medical care statistics and the health services system. *Public Health Rep.* 1967;82(10):847-854.
2. Trask J. Vital statistics. In: Rosenau M, ed. *Preventive Medicine and Hygiene.* New York, NY: D. Appleton-Century Company; 1935:1175-1220.

3. Cumming H. Chronic disease as a public health problem. *Milbank Mem Fund Q.* 1936;14:125-131.
4. Hailman D. Health status of adults in the productive ages. *Public Health Rep.* 1941;56:2071-2087.
5. Collins S. Sickness surveys. In: Emerson H, ed. *Administrative Medicine.* New York, NY: Thomas Nelson & Sons; 1949:511-535.
6. Logan W, Brooke E. The survey of sickness 1943 to 1952. *Stud Med Popul Subj.* 1953;12:43.
7. Sheldon MP. A physical achievement record for use with crippled children. *J Health Phys Educ.* 1935;60:30-31.
8. Deaver GG, Brown ME. *Physical Demands of Daily Life: An Objective Scale for Rating the Orthopedically Exceptional.* New York, NY: Institute for the Crippled and Disabled; 1945.
9. Bennett R, Stephens H. Functional testing and training. *Phys Ther Rev.* 1949;29(3):99-107.
10. Moskowitz E, McCann CB. Classification of disability in the chronically ill and aging. *J Chronic Dis.* 1957;5(3):342-346.
11. Mahoney FI, Wood OH, Barthel DW. Rehabilitation of chronically ill patients: the influence of complications on the final goal. *South Med J.* 1958;51(5):605-609.
12. The Staff of the Benjamin Rose Hospital. Multidisciplinary studies of illness in aged persons: II. A new classification of functional status in activities of daily living. *J Chronic Dis.* 1959;9(1):55-62.

13. US Department of Health, Education, and Welfare. Health survey procedure: Concepts, questionnaire development, and definitions in the health interview survey. *Vital Health Stat 1.* 1964;1(2):1-66.
14. Hoffer CR, Schuler EA. Measurement of health needs and health care. *Am Sociol Rev.* 1948;13:719-724.
15. Brodman K, Erdmann AJ, Wolff HG. The Cornell Medical Index: an adjunct to medical interview. *J Am Med Assoc.* 1949;140(6):530-534.
16. Souelem O. Mental patients' attitudes toward mental hospitals. *J Clin Psychol.* 1955;11:181-185.
17. Hillson JS, Klopfer WG, Wylie AA. Attitudes toward mental hospitals. *J Clin Psychol.* 1956;12:361-365.
18. Hecker J, Lewis CE. Factors determining attitudes towards medical care—study of a metropolitan area. *J Kans Med Soc.* 1965;66:123-128.
19. McHorney CA. Use of item response theory to link 3 modules of functional status items from the Asset and Health Dynamics Among the Oldest Old study. *Arch Phys Med Rehabil.* 2002;83(3):383-394.
20. Task Force for the Handbook of Psychiatric Measures. *Handbook of Psychiatric Measures.* Washington, DC: American Psychiatric Association; 2000.
21. McHorney C. Generic health measurement: past accomplishments and a measurement paradigm for the 21st century. *Ann Intern Med.* 1997;127:743-750.
22. McHorney C. Health status assessment methods for adults: Past accomplishments and future challenges. *Annu Rev Public Health.* 1999;20:309-335.

23. Bowling A. *Measuring Disease: A Review of Disease-Specific Quality of Life Measurement Scales.* 2nd ed. Buckingham: Open University Press; 2001.

24. McHorney CA. Prospects and problems associated with item banking and computerized adaptive testing in cancer clinical trials. National Cancer Institute, Cancer Outcomes Working Group Symposium, 2001.

25. Gulliksen H. *Theory of Mental Tests.* New York: John Wiley and Sons, Inc.; 1950.

26. Lord FM, Novick MR. *Statistical Theories of Mental Test Scores.* Reading, Mass: Addison-Wesley Publishing Company; 1968.

27. Hambleton R, Swaminathan H. *Item Response Theory: Principles and Applications.* Boston, Mass: Kluwer Nijoff Publishing; 1985.

28. McHorney CA, Cohen AS. Equating health status measures with item response theory: illustrations with functional status items. *Med Care.* 2000;38(9 Suppl):II43-II59.

29. Cleary PD. Future directions of quality of life research. In: Spilker B, ed. *Quality of Life and Pharmacoeconomics in Clinical Trials.* 2nd ed. Philadelphia, Pa: Lippincott-Raven Publishers; 1996:73-78.

30. Hambleton R. Emergence of item response modeling in instrument development and data analysis. *Med Care.* 2000;38(9 Suppl):II60-II65.

31. Teresi JA. Statistical methods for examination of differential item functioning (DIF) with applications to cross-cultural measurement of functional, physical and mental health. *J Men Health Aging*. 2001;7(1):31-40.

32. Lord FM. *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Lawrence Erlbaum Associates; 1980.

33. Baker FB. *Item Response Theory: Parameter Estimation Techniques.* New York, NY: Dekker; 1992.
34. Gill TM, Feinstein AR. A critical appraisal of the quality of quality-of-life measurements. *JAMA*. 1994;272(8):619-626.
35. Leplege A, Hunt S. The problem of quality of life in medicine. *JAMA* 1997;278(1):47-50.
36. World Health Organization. *Chron World Health Organ.* 1947;1(1-2).
37. Ahmed P, Kolker A. The role of indigenous medicine in WHO's definition of health. In: Ahmed P, Coelgo G, eds. *Toward A New Definition of Health: Psychosocial Dimensions.* New York, NY: Plenum Press; 1979:113-128.
38. Stewart AL, Ware JE. *Measuring Functioning and Well-Being: The Medical Outcomes Study Approach.* Durham, NC: Duke University Press; 1992.
39. Meenan RF, Gertman PM, Mason JH. Measuring health status in arthritis: the arthritis impact measurement scales. *Arthritis Rheum.* 1980;23(2):146-152.
40. Roland M, Morris R. A study of the natural history of back pain. Part I: development of a reliable and sensitive measure of disability in low-back pain. *Spine.* 1983;8(2):141-144.
41. Wu AW, Rubin HR, Mathews WC, et al. A health status questionnaire using 30 items from the Medical Outcomes Study. Preliminary validation in persons with early HIV infection. *Med Care.* 1991;29(8):786-798.
42. Ingersoll GM, Marrero DG. A modified quality-of-life measure for youths: psychometric properties. *Diabetes Educ.* 1991;17(2):114-118.
43. Marks GB, Dunn SM, Woolcock AJ. A scale for the measurement of quality of life in adults with asthma. *J Clin Epidemiol.* 1992;45(5):461-472.

44. Hyland M, Bott J, Singh S, et al. Domains, constructs and the development of the breathing problems questionnaire. *Qual Life Res.* 1994;3:245-256.

45. McHorney CA, Bricker DE, Kramer AE, et al. The SWAL-QOL outcomes tool for oropharyngeal dysphagia in adults: I. Conceptual foundation and item development. *Dysphagia.* 2000;15:115-121.

46. McHorney CA. Methodological and psychometric issues in health status assessment across populations and applications. In: Albrecht GL, Fitzpatrick R, eds. *Advances in Medical Sociology.* Vol. 5. Greenwich, Conn: JAI Press; 1994:281-304.

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

48. Jobe JB, Mingay DJ. Cognitive laboratory approach to designing questionnaires for surveys of the elderly. *Public Health Rep.* 1990;105(5):518-524.

49. Lessler JT. Choosing questions that people can understand and answer. *Med Care.* 1995;33(4 Suppl):AS203-AS208.

50. Weitzner MA, Meyers CA, Steinbruecker S, et al. Developing a care giver quality-of-life instrument. Preliminary steps. *Cancer Pract.* 1997;5(1):25-31.

51. Moore KN, Estey A. The early post-operative concerns of men after radical prostatectomy. *J Adv Nurs.* 1999;29(5):1121-1129.

52. Hobart J, Lamping D, Fitzpatrick R, et al. The Multiple Sclerosis Impact Scale (MSIS-29): a new patient-based outcome measure. *Brain.* 2001;124:962-973.

53. Kadam U, Croft P, McLeod M, et al. A qualitative study on patients' views on anxiety and depression. *Brit J Gen Pract.* 2001;51:375-380.

54. Ware NC, Tugenberg T, Dickey B, et al. An ethnographic study of the meaning of continuity of care in mental health services. *Psychiatr Serv.* 1999;50(3):395-400.

55. Angel R, Frisco ML. Self-assessments of health and functional capacity among older adults. *J Ment Health Aging.* 2001;7(1):119-135.

56. De Geest S, Abraham I, Gemoets H, et al. Development of the long-term medication behaviour self-efficacy scale: qualitative study for item development. *J Adv Nurs.* 1994;19:233-238.

57. Porter EJ. A phenomenological alternative to the "ADL Research Tradition." *J Aging Health.* 1995;7(1):24-45.

58. Krasner D. Painful venous ulcers: themes and stories about their impact on quality of life. *Ostomy Wound Manage.* 1998;44(9):38-42.

59. Koch T, Webb C, Williams AM. Listening to the voices of older patients: an existential-phenomenological approach to quality assurance. *J Clin Nurs.* 1995;4:185-193.

60. Spies JB, Coyne K, Guaou N, et al. The UFS-QOL, a new disease-specific symptom and health-related quality of life questionnaire for leiomyomata. *Obstet Gynecol.* 2002;99(2):290-300.

61. Lerner D, Amick BC 3rd, Rogers WH, et al. The Work Limitations Questionnaire. *Med Care.* 2001;39(1):72-85.

62. Wu AW, Fink NE, Cagney KA, et al. Developing a health-related quality-of-life measure for end-stage renal disease: The CHOICE Health Experience Questionnaire. *Am J Kidney Dis.* 2001;37(1):11-21.

63. Juniper EF, Guyatt GH, Epstein RS, et al. Evaluation of impairment of health related quality of life in asthma: development of a questionnaire for use in clinical trials. *Thorax.* 1992;47:76-83.

64. Wilde B, Larsson G, Larsson M, et al. Quality of care. Development of a patient-centred questionnaire based on a grounded theory model. *Scand J Caring Sci.* 1994;8:39-48.

65. Launois R, Reboul-Marty J, Henry B. Construction and validation of a quality of life questionnaire in chronic lower limb venous insufficiency (CIVIQ). *Qual Life Res.* 1996;5:539-554.

66. Rubin HR, Jenckes M, Fink NE, et al. Patient's view of dialysis care: development of a taxonomy and rating of importance of different aspects of care. CHOICE Study. Choices for Healthy Outcomes in Caring for ESRD. *Am J Kidney Dis.* 1997;30(6):793-801.

67. Cronin L, Guyatt G, Griffith L, et al. Development of a health-related quality-of-life questionnaire (PCOSQ) for women with polycystic ovary syndrome (PCOS). *J Clin Endocrinol Metab.* 1998;83(6):1976-1987.

68. Jabine T, Straf ML, Tanue JM, et al. *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines.* Washington, DC: National Academies Press; 1984.

69. Fienberg SE, Loftus EF, Tanur JM. Cognitive aspects of health survey methodology: an overview. *Milbank Mem Fund Q Health Soc.* 1985;63(3):547-564.

70. Lessler JT, Sirken MG. Laboratory-based research on the Cognitive Aspects of Survey Methodology: The goals and methods of the National Center for Health Statistics study. *Milbank Mem Fund Q Health Soc.* 1985;63(3):565-581.

71. Loftus E, Fienberg S, Tanur J. Cognitive psychology meets the national survey. *Am Psychol.* 1985;40(2):175-180.

72. Jobe JB, Mingay DJ. Cognitive research improves questionnaires. *Am J Public Health.* 1989;79(8):1053-1055.

73. Bergner M, Bobbitt RA, Pollard WE, et al. The sickness impact profile: validation of a health status measure. *Med Care.* 1976;14(1):57-67.

74. Ware JE Jr, Kosinski M, Bayliss MS, et al. Comparison of methods for the scoring and statistical analysis of SF-36 health profile and summary measures: summary of results from the Medical Outcomes Study. *Med Care.* 1995;33(4 Suppl):AS264-AS279.

75. Bozzette SA, Hays RD, Berry SH, et al. A Perceived Health Index for use in persons with advanced HIV disease: derivation, reliability, and validity. *Med Care.* 1994;32(7):716-731.

76. Bozzette SA, Hays RD, Berry SH, et al. Derivation and properties of a brief health status assessment instrument for use in HIV disease. *J Acqui Immune Defic Synd Hum Retrovirol.* 1995;8(3):253-265.

77. Devinsky O, Vickrey BG, Cramer J, et al. Development of the quality of life in epilepsy inventory. *Epilepsia.* 1995;36(11):1089-1104.

78. Vickrey BG, Hays RD, Harooni R, et al. A health-related quality of life measure for multiple sclerosis. *Qual Life Res.* 1995;4:187-206.

79. Revicki DA, Sorensen S, Wu AW. Reliability and validity of physical and mental health summary scores from the Medical Outcomes Study HIV Health Survey. *Med Care.* 1998;36(2):126-137.

80. Varni JW, Seid M, Kurtin PS. PedsQL 4.0: reliability and validity of the Pediatric Quality of Life Inventory version 4.0 generic core scales in healthy and patient populations. *Med Care.* 2001;39(8):800-812.

81. Simon GE, Revicki DA, Grothaus L, et al. SF-36 summary scores: are physical and mental health truly distinct? *Med Care.* 1998;36(4):567-572.

82. Rubenach S, Shadbolt B, McCallum J, et al. Assessing health-related quality of life following myocardial infarction: is the SF-12 useful? *J Clin Epidemiol.* 2002;55:306-309.

83. Fairclough DL. Summary measures and statistics for comparisons of quality of life in a clinical trial of cancer therapy. *Stat Med.* 1997;16:1197-1209.

84. Beusterien KM, Nissenson AR, Port FK, et al. The effects of recombinant human erythropoietin on functional health and well-being in chronic dialysis patients. *J Am Soc Nephrol.* 1996;7(5):763-73.

85. Jenkinson C, Gray A, Doll H, et al. Evaluation of index and profile measures of health status in a randomized controlled trial. Comparison of the Medical Outcomes Study 36-item Short Form Health Survey, EuroQoL, and disease specific measures. *Med Care.* 1997;35(11):1109-1118.

86. Ruhland JL, Shields RK. The effects of a home exercise program on impairment and health-related quality of life in persons with chronic peripheral neuropathies. *Phys Ther.* 1997;77(10):1026-1039.

87. Gartsman GM, Brinker MR, Khan M, et al. Self-assessment of general health status in patients with five common shoulder conditions. *J Shoulder Elbow Surg.* 1998;7(3):228-237.

88. Gartsman GM, Khan M, Hammerman SM. Arthroscopic repair of full-thickness tears of the rotator cuff. *J Bone Joint Surg Am.* 1998;80-A(6):832-840.

89. Wiklund I, Junghard O, Grace E, et al. Quality of Life in Reflux and Dyspepsia Patients. Psychometric documentation of a new disease-specific questionnaire (QOLRAD). *Eur J Surg Suppl.* 1998;583:41-49.

90. Reuben DB, Frank JC, Hirsch SH, et al. A randomized clinical trial of outpatient comprehensive geriatric assessment coupled with an intervention to increase adherence to recommendations. *J Am Geriatr Soc.* 1999;47:269-276.

91. Revicki DA, Crawley JA, Zodet MW, et al. Complete resolution of heartburn symptoms and health-related quality of life in patients with gastro-oesophageal reflux disease. *Aliment Pharmacol Ther.* 1999;13:1621-1630.

92. Stavem K, Erikssen J, Boe J. Performance of a short lung-specific health status measure in outpatients with chronic obstructive pulmonary disease. *Respir Med.* 1999;93:467-475.

93. Taylor SJ, Taylor AE, Foy MA, et al. Responsiveness of common outcome measures for patients with low back pain. *Spine.* 1999;24:1805-1812.

94. Adler D, Bungay KM, Cynn DJ, et al. Patient-based health status assessments in an outpatient psychiatry setting. *Psychiatr Serv.* 2000;51(3):341-348.

95. Hughes SL, Weaver FM, Giobbie-Hurder A, et al. Effectiveness of team-managed home-based primary care: A randomized multicenter trial. *JAMA.* 2000;284(22):2877-2885.

96. Mayo NE, Wood-Dauphinee S, Cote R, et al. There's no place like home: an evaluation of early supported discharge for stroke. *Stroke.* 2000;31:1016-1023.

97. Nortvedt MW, Riise T, Myhr KM, et al. Performance of the SF-36, SF-12 and RAND-36 summary scales in a muliple sclerosis population. *Med Care.* 2000;38(10):1022-1028.

98. Hobart J, Freeman J, Lamping D, et al. The SF-36 in multiple sclerosis: why basic assumptions must be tested. *J Neurol Neurosurg Psychiatry.* 2001;71(3):363-370.

99. Roznowski M, Tucker L, Humphreys L. Three approaches to determining the dimensionality of binary items. *Applied Psychological Measurement.* 1991;15(2):109-127.

100. US Department of Health and Human Services. Healthy People 2010. Available at: http:/www.health.gov/healthypeople/document/html/uih/uih_2.htm. Accessed April 23, 2002.

101. National Institutes of Health. Addressing health disparities: The NIH program of action. Available at: http://healthdisparities.nih.gov. Accessed July 16, 2002.

102. McKinlay JB, McKinlay SM, Beaglehole R. A review of the evidence concerning the impact of medical measures on recent mortality and morbidity in the United States. *Int J Health Serv.* 1989;19(2):181-208.

103. Evans RG, Stoddart GL. Producing health, consuming health care. *Soc Sci Med.* 1990;31:1347-1363.

104. US Department of Health and Human Services. Risks to good health. *Healthy People: The Surgeon General's Report on Health Promotion and Disease Prevention.* Washington, DC: US Department of Health, Education, and Welfare; 1979:2-1-2-8.

105. McHorney CA. Concepts and measurement of health status and health-related quality of life. In: Albrecht GL, Fitzpatrick R, Scrimshaw S, eds. *The Handbook of Social Studies in Health & Medicine.* London: SAGE, 1999:339-358.

106. Blaxter M. Evidence on inequality in health from a national survey. *Lancet.* 1987;4:30-33.

107. House JS, Kessler RC, Herzog AR. Age, socioeconomic status, and health. *Milbank Q.* 1990;68(3):383-411.

108. Arber S, Ginn J. Gender and inequalities in health in later life. *Soc Sci Med.* 1993;36(1):33-46.

109. Guralnik JM, Land KC, Blazer D, et al. Educational status and active life expectancy among older blacks and whites. *N Engl J Med.* 1993;329(2):110-116.

110. Sorlie PD, Backlund E, Keller JB. US mortality by economic, demographic, and social characteristics: the National Longitudinal Mortality Study. *Am J Public Health.* 1995;85(7):949-956.

111. Schoenbaum M, Waidmann T. Race, socioeconomic status, and health: accounting for race differences in health. *J Gerontol B Psychol Sci Soc Sci.* 1997;52(Special Issue):61-73.

112. Kind P, Dolan P, Gudex C, et al. Variations in population health status: results from a United Kingdom national questionnaire survey. *BMJ.* 1998;316:736-741.

113. Adams P, Hendershot GE, Marano M. Current Estimates from the National Health Interview Survey, 1996. Hyattsville, Md: US Department of Health and Human Services; 1999; (PHS) 99-1528. (Data from the National Health Survey; Vol. 10).

114. Stewart AL, Napoles-Springer A. Health-related quality-of-life assessments in diverse population groups in the United States. *Med Care.* 2000;38(9 Suppl):II102-II124.

115. Day J. *Population Projections of the United States by Age, Sex, Race, and Hispanic Origin: 1995 to 2050.* Washington, DC: US Government Printing Office, 1996.

116. Liang J. Assessing cross-cultural comparability in mental health among older adults. *J Ment Health Aging.* 2001;7(1):21-30.

117. Herdman M, Fox-Rushby J, Badia X. "Equivalence" and the translation and adaptation of health-related quality of life questionnaires. *Qual Life Res.* 1997;6:237-247.

118. Byrne BM, Campbell TL. Cross-cultural comparisons and the presumption of equivalent measurement and theoretical structure. A look beneath the surface. *J Cross Cult Psychol.* 1999;30(5):555-574.

119. Kleinman AM. Depression, somatization and the "new cross-cultural psychiatry." *Soc Sci Med.* 1977;11:3-10.

120. Flaherty JA, Gaviria FM, Pathak D, et al. Developing instruments for cross-cultural psychiatric research. *J Nerv Ment Dis.* 1988;176(5):257-263.

121. Pelto PJ, Pelto GH. Studying knowledge, culture, and behavior in applied medical anthropology. *Med Anthropol Q.* 1997;11(2):147-163.

122. Lutner RE, Roizen MF, Stocking CB, et al. The automated interview versus the personal interview. Do patient responses to preparative health questions differ? *Anesthesiology.* 1991;75:394-400.

123. Rozensky RH, Honor LF, Rasinski K, et al. Paper-and-pencil versus computer-administered MMPIs: a comparison of patients' attitudes. *Comput Human Behav.* 1986;2:111-116.

124. Greist JH, Klein MH, Erdman HP, et al. Comparison of computer- and interviewer-administered versions of the Diagnostic Interview Schedule. *Hosp Community Psychiatry.* 1987;38(12):1304-1311.

125. Kobak K, Reynolds W, Rosenfeld R, et al. Development and validation of a computer-administered version of the Hamilton Depression Rating Scale. *Psychol Assess.* 1990;2(1):56-63.

126. Baer L, Brown-Beasley MW, Sorce J, et al. Computer-assisted telephone administration of a structured interview for obsessive-compulsive disorders. *Am J Psychiatry.* 1993;150:1737-1738.

127. Kobak KA, Reynolds WM, Greist JH. Development and validation of a computer-administered version of the Hamilton Anxiety Scale. *Psychol Assess.* 1993;5(4):487-492.

128. Baer L, Jacobs DG, Cukor P, et al. Automated telephone screening survey for depression. *JAMA.* 1995;273(24):1943-1944.

129. Roizen MF, Coalson D, Hayward RS, et al. Can patients use an automated questionnaire to define their current health status? *Med Care.* 1992;30(5 Suppl):MS74-MS84.

130. Newell S, Girgis A, Sanson-Fisher RW, et al. Are touchscreen computer surveys acceptable to medical oncology patients? *J Psychosoc Oncol.* 1997;15(2):37-46.

131. Taenzer PA, Speca M, Atkinson MJ, et al. Computerized quality-of-life screening in an oncology clinic. *Cancer Pract.* 1997;5(3):168-175.

132. Buxton J, White M, Osoba D. Patients' experiences using a computerized program with a touch-sensitive video monitor for the assessment of health-related quality of life. *Qual Life Res.* 1998;7:513-51.

133. McBride JS, Anderson RT, Bahnson JL. Using a hand-held computer to collect data in an orthopedic outpatient clinic: a randomized trial of two survey methods. *Med Care.* 1999;37(7):647-651.

134. Velikova G, Wright EP, Smith AB, et al. Automated collection of quality-of-life data: a comparison of paper and computer touch-screen questionnaires. *J Clin Oncol.* 1999;17(3):998-1007.

135. Lofland J, Schaffer M, Goldfarb N. Evaluating health-related quality of life: cost comparison of computerized touch-screen technology and traditional paper systems. *Pharmacotherapy.* 2000;20(11):1390-1395.

136. Taenzer P, Bultz BD, Carlson LE, et al. Impact of computerized quality of life screening on physician behaviour and patient satisfaction in lung cancer outpatients. *Psychooncology.* 2000;9:203-213.

137. Carlson LE, Speca M, Hagen N, et al. Computerized quality of life screening in a cancer pain clinic. *J Pallia Care.* 2001;17(1):46-52.

138. Ernst ME, Doucette WR, Dedhiya SD, et al. Use of point-of-service health status assessments by community pharmacists to identify and resolve drug-related problems in patients with musculoskeletal disorders. *Pharmacotherapy.* 2001;21(8):988-997.

139. Brodie M, Fournoy RE, Altman DE, et al. Health information, the Internet, and the digital divide. *Health Aff (Millwood).* 2000;19(6):255-265.

140. Burstin, H. Traversing the digital divide. *Health Aff (Millwood).* 2000;19(6):245-249.

141. Ree MJ. Foreward: differential item functioning (DIF): a perspective from the Air Force Human Resources Laboratory. In: Holland PW, Wainer H, eds. *Differential Item Functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates; 1993:xi-xii.

142. Hambleton RK. The next generation of the ITC test translation and adaptation guidelines. *Euro J Psychol Assess.* 2001;17(3):164-172.

143. Teresi JA, Cross PS, Golden RR. Some applications of latent trait analysis to the measurement of ADL. *J Gerontol.* 1989;44(5):S196-S204.

144. Groenvold M, Bjorner JB, Klee MC, et al. Test for item bias in a quality of life questionnaire. *J Clin Epidemiol.* 1995;48(6):805-816.

145. Avlund K, Era P, Davidsen M, et al. Item bias in self-reported functional ability among 75-year-old men and women in three Nordic localities. *Scand J Soc Med.* 1996; 24(3):206-217.

146. Kempen GI, Miedema I, Ormel J, et al. The assessment of disability with the Groningen Activity Restriction Scale. Conceptual framework and psychometric properties. *Soc Sci Med.* 1996;43(11):1601-1610.

147. Spector WD, Fleishman JA. Combining activities of daily living with instrumental activities of daily living to measure functional disability. *J Gerontol B Psychol Sci Soc Sci.* 1998;53(1):S46-S57.

148. Custers JW, Hoijtink H, van der Net J, et al. Cultural differences in functional status measurement: analyses of person fit according to the Rasch model. *Qual Life Res.* 2000;9:571-578.

149. Wolfe F, Hawley DJ, Goldenberg DL, et al. The assessment of functional impairment in fibromyalgia (FM): Rasch analyses of 5 functional scales and the development of the FM Health Assessment Questionnaire. *J Rheumatol.* 2000;27(8):1989-1999.

150. Jagger C, Arthur AJ, Spiers NA, et al. Patters of onset of disability in activities of daily living with age. *J Am Geriatr Soc.* 2001;49(4):404-409.

151. Teresi JA, Golden RR, Cross P, et al. Item bias in cognitive screening measures: Comparisons of elderly white, Afro-American, Hispanic and high and low education subgroups. *J Clin Epidemiol.* 1995;48(4):473-483.

152. Teresi JA, Kleinman M, Ocepek-Welikson K. Modern psychometric methods for detection of differential item functioning: Application to cognitive assessment measures. *Stat Med.* 2000;19:1651-1683.

153. Teresi JA, Kleinman M, Ocepek-Welikson K, et al. Applications of item response theory to the examination of the psychometric properties and differential item functioning of the comprehensive assessment and referral evaluation dementia diagnostic scale among samples of Latino, African American, and white non-Latino elderly. *Res Aging.* 2000;22(6):738-773.

154. Teresi JA, Holmes D, Ramirez M,et al. Performance of cognitive tests among different racial/ethnic and education groups: findings of differential item functioning and possible item bias. *J Ment Health Aging.* 2001;7(1):79-89.

155. Bjorner JB, Kreiner S, Ware JE, et al. Differential item functioning in the Danish translation of the SF-36. *J Clin Epidemiol.* 1998;51(11):1189-1202.

156. Collins WC, Raju NS, Edwards JE. Assessing differential functioning in a satisfaction scale. *J Appl Psychol.* 2000;85(3):451-461.

157. Morales LS, Reise SP, Hays RD. Evaluating the equivalence of health care ratings by whites and Hispanics. *Med Care.* 2000;38(5):517-527.

158. Schaeffer NC. An application of item response theory to the measurement of depression. In: Clogg C, ed. *Sociological Methodology.* San Francisco, Calif: Jossey-Bass; 1988:271-307.

159. Ellis BB, Minsel B, Becker P. Evaluation of attitude survey translations: an investigation using item response theory. *Int J Psychol.* 1989;24:665-684.

160. Ellis BE, Kimmell HD. Identification of unique cultural response patterns by means of item response theory. *J Appl Psychol.* 1992;77:177-184.

161. Stommel M, Given BA, Given CW, et al. Gender bias in the measurement properties of the Center for Epidemiologic Studies Depression Scale (CES-D). *Psychiatry Res.* 1993;49:239-250.

162. Dancer LS, Anderson AJ, Derlin RL. Use of log-linear models for assessing differential item functioning in a measure of psychological functioning. *J Consult Clin Psychol.* 1994;62(4):710-717.

163. Flannery WP, Reise SP, Widaman KF. An item response theory analysis of the general and academic scales of the Self-Description Questionnaire II. *J Res Pers.* 1995;29:168-188.

164. Hammond SM. In IRT investigation of the validity of non-patient analogue research using the Beck Depression Inventory. *Eur J Psychol Assess.* 1995;11(1):14-20.

165. Huang CD, Church AT, Katigbak MS. Identifying cultural differences in items and traits: differential item functioning in the NEO Personality Inventory. *J Cross Cult Psychol.* 1997;28(2):192-218.

166. Panter AT, Swygert KA, Grant Dahlstrom W, et al. Factor analytic approaches to personality item-level data. *J Pers Assess.* 1997;68(3):561-589.

167. Suh T, Gallo JJ. Symptom profiles of depression among general medical service users compared with specialty mental health service users. *Psychol Med.* 1997;27:1051-1063.

168. Coelho VL, Strauss ME, Jenkins JH. Expression of symptomatic distress by Puerto Rican and Euro-American patients with depression and schizophrenia. *J Nerv Ment Dis.* 1998;186(8):477-483.

169. DeRoos Y, Allen-Meares P. Application of Rasch analysis: exploring differences in depression between African-American and white children. *J Soc Serv Res.* 1998;23(3/4):93-107.

170. Gallo JJ, Cooper-Patrick L, Lesikar S. Depressive symptoms of whites and African Americans aged 60 years and older. *J Gerontol B Psychol Sci Soc Sci.* 1998;53(5):P277-P286.

171. Santor DA, Ramsay JO. Progress in the technology of measurement: applications of item response models. Psychol Assess 1998;10:345-359.

172. Smith LL, Reise SP. Gender differences on negative affectivity: an IRT study of differential item functioning on the Multidimensional Personality Questionnaire Stress Reaction Scale. *J Pers Soc Psychol.* 1998;75(5):1350-1362.

173. Christensen H, Jorm AF, Mackinnon AJ, et al. Age differences in depression and anxiety symptoms: a structural equation modeling analysis of data from a general population sample. *Psychol Med.* 1999;29(2):325-339.

174. Cooke DJ, Michie C. Psychopathy across cultures: North America and Scotland compared. *J Abnorm Psychol.* 1999;108(1):58-68.

175. Grayson DA, Mackinnon A, Jorm AF, et al. Item bias in the Center for Epidemiologic Studies Depression scale: effects of physical disorders and disability in an elderly community sample. *J Gerontol B Psychol Sci Soc Sci.* 2000;55(5):P273-P282.

176. Waller NG, Thompson JS, Wenk E. Using IRT to separate measurement bias from true group differences on homogeneous and heterogeneous scales: an illustration with the MMPI. *Psychol Methods.* 2000; 5(1):125-146.

177. Azocar F, Arean P, Miranda J, et al. Differential item functioning in a Spanish translation of the Beck Depression Inventory. *J Clin Psychol.* 2001;57:355-365.

178. Santor DA, Coyne JC. Evaluating the continuity of symptomology between depressed and nondepressed individuals. *J Abnorm Psychol.* 2001;110(2):216-225.

179. Drasgow F, Hulin CL. Cross-cultural measurement. *Interamerican J Psychol.* 1987;21(1-2):1-24.

180. Walstad WB, Robson D. Differential item functioning and male-female differences on multiple-choice tests in economics. *J Econ Educ.* 1997;28:155-171.

181. Guarnaccia PJ. Anthropological perspectives: the importance of culture in the assessment of quality of life. In: Spilker B, ed. *Quality of Life and Pharmacoeconomics in Clinical Trials.* 2nd ed. Philadelphia, Pa: Lippincott-Raven Publishers; 1996:523-527.

182. Rogler LH. The meaning of culturally sensitive research in mental health. *Am J Psychiatry.* 1989;146(3):296-303.

183. McHorney CA, Ware JE Jr, Rogers W, et al. The validity and relative precision of MOS short- and long-form health status scales and Dartmouth COOP Charts. Results from the Medical Outcomes Study. *Med Care.* 1992;30(5 Suppl):MS253-MS265.

184. McHorney CA, Tarlov AR. The use of health status measures for individual patient level applications: problems and prospects. *Qual Life Res.* 1994;3:43-44.

185. McHorney CA. Measuring and monitoring general health status in elderly persons: practical and methodological issues in using the SF-36 Health Survey. *Gerontologist.* 1996;36(5):571-583.

186. Haley SM, McHorney CA, Ware JE Jr. Evaluation of the MOS SF-36 physical functioning scale (PF-10): I. Unidimensionality and reproducibility of the Rasch item scale. *J Clin Epidemiol.* 1994;47(6):671-684.

187. Prieto L, Alonso J, Lamarca R, et al. Rasch measurement for reducing the items of the Nottingham Health Profile. *J Outcome Meas.* 1998;2(4):285-301.

188. Finch M, Kane RL, Philp I. Developing a new metric for ADLs. *J Am Geriatr Soc.* 1995;43:877-884.

189. Grimby G, Andrén E, Holmgren E, et al. Structure of a combination of Functional Independence Measure and Instrumental Activity Measure items in community-living

persons: a study of individuals with cerebral palsy and spina bifida. *Arch Phys Med Rehabil.* 1996;77:1109-1114.

190. Nordenskiold U. Daily activities in women with rheumatoid arthritis. Goteborg, Sweden: Goteborg University; 1996.
191. Doble SE, Fisher AG. The dimensionality and validity of the Older Americans Resources and Services (OARS) Activities of Daily Living (ADL) Scale. *J Outcome Meas.* 1998;2(1):4-24.
192. Grimby G, Andren E, Daving Y, et al. Dependence and perceived difficulty in daily activities in community-living stroke survivors 2 years after stroke: a study of instrumental structures. *Stroke.* 1998;29:1843-1849.
193. Carter J, Wilkinson L. A latent trait analysis of the MMPI. *Multivariate Behav Res.* 1984;19:385-407.
194. Steinberg L, Thissen D. Item response theory in personality research. In: Shrout P, Fiske S, eds. *Personality Research, Methods, and Theory: A Festscrift Honoring Donald W. Fiske.* Hillsdale, NJ: Lawrence Earlbaum Associates; 1995:161-181.
195. Gray-Little B, Williams VSL, Hancock TD. An item response theory analysis of the Rosenberg Self-Esteem Scale. *Pers Soc Psychol Bull.* 1997;23(5):443-451.
196. Payne SL. *The Art of Asking Questions.* Princeton, NJ: Princeton University Press; 1951.
197. Wesman AG. Writing the test item. In: Thorndike RL, ed. *Educational Measurement.* Washington, DC: American Council on Education; 1971:81-129.
198. Andrich D. Distinctive and incompatible properties of two common classes of IRT models for graded responses. *Appl Psychol Meas.* 1995;19(1):101-119.
199. Wright BD, Masters GN. *Rating Scale Analysis.* Chicago, Ill: MESA Press; 1982.

200. Gibbons RD, Clark DC, VonAmmon Cavanaugh S, et al. Application of modern psychometric theory in psychiatric research. *J Psychiatr Res.* 1985;19(1):43-55.

201. Steinberg L. Context and serial-order effects in personality measurement: limits on the generality of measuring changes the measure. *J Pers Soc Psychol.* 1994;66(2):341-349.

202. Kirisci L, Clark DB, Moss HB. Reliability and validity of the State-Trait Anxiety Inventory for Children in adolescent substance abusers: confirmatory factor analysis and item response theory. *J Child Adolesc Subst Abuse.* 1996;5(3):57-69.

203. Kirisci L, Moss HB, Tarter RE. Psychometric evaluation of the Situational Confidence Questionnaire in adolescents: fitting a graded item response model. *Addict Behav.* 1996;21(3):303-317.

204. Cooke DJ, Michie C. An item response theory analysis of the Hare Psychopathy Checklist-Revised. *Psychol Assess.* 1997;9(1):3-14.

205. Kim Y, Pilkonis PA. Selecting the most informative items in the IIP scales for personality disorders: an application of item response theory. *J Personal Disord.* 1999;13(2):157-174.

206. Waller NG. Searching for structure in the MMPI. In: Embretson S, Hershberger S, eds. *The New Rules of Measurement: What Every Psychologist and Educator Should Know.* Mahwah, NJ: Lawrence Earlbaum Associates; 1999:185-217.

207. Marshall GN, Orlando M, Jaycox LH, et al. Development and validation of a modified version of the Peritraumatic Dissociative Experiences Questionnaire. *Psychol Assess.* 2002; 14(2):123-134.

208. Masters GN. A Rasch model for partial credit scoring. *Psychometrika.* 1982;47(2):149-174.

209. Kolen MJ, Brennan RL. *Test Equating: Methods and Practices.* New York, NY: Springer-Verlag; 1995.

210. Feuer MJ, Holland PW, Green BF, et al, eds. *Uncommon Measures: Equivalence and Linkage among Educational Tests.* Washington, DC: National Academies Press; 1999.

211. Reise SP, Henson JM. Computerization and adaptive administration of the NEO PI-R. *Assessment.* 2000;7(4):347-367.

212. Green BF. The promise of tailored tests. In: Wainer H, Messick S, eds. *Principles of Modern Psychological Measurement.* Hillsdale, NJ: Lawrence Earlbaum Associates; 1983:69-80.

213. Hambleton RK, Jones RW. Comparison of classical test theory and item response theory and their applications to test development. *Educ Meas Issues Pract.* 1993;12:38-47.

214. Waller NG, Reise SP. Computerized adaptive personality assessment: an illustration with the Absorption Scale. *J Pers Soc Psychol.* 1989;57(6):1051-1058.

215. Koch WR, Dodd BG, Fitzpatrick SJ. Computerized adaptive measurements of attitudes. *Meas Eval Couns Dev.* 1990;23:20-30.

216. Dodd BG, De Ayala RJ, Koch WR. Computerized adaptive testing with polytomous items. *Appl Psychol Meas.* 1995;19(1):5-22.

217. Hambleton RK, Slater SC. Item response theory models and testing practices: current international status and future directions. *Eur J Psychol Assess.* 1997;13(1):21-28.

218. Shapiro MF. Is the spirit of capitalism undermining the ethics of health services research? *Health Serv Res.* 1994;28(6):661-672.

Contact Information:

Colleen A. McHorney, Ph.D., Merck, Inc., colleen_mchorney@merck.com.

Acknowledgment:

Preparation of this manuscript was supported by the Department of Veterans Affairs (RR&D C-2488-R)

Thoughts on Assorted Issues in Health-Related Quality of Life Assessment

Ron D. Hays, Ph.D.

This paper provides a written record of comments made at the National Center for Health Statistics (CDC) Summary Measures Workshop held March 2003. Topics covered include the definition of health-related quality of life, use of qualitative and quantitative methods (mixed methods) in developing measures, assessing unidimensionality, aggregation of health-related quality of life measures, determinants of health outcomes, minimally important differences, and evaluating measures in disadvantaged populations (differential item functioning, matching items to individuals, readability).

Definition of Health-Related Quality of Life

The health-related quality of life (HRQOL) field has achieved some level of consensus on the definition of HRQOL. That is, it is generally agreed that HRQOL encompasses functioning and well-being in physical, mental, and social dimensions of life. Functioning refers to the ability to perform as well as the performance of daily activities ranging from the most basic self-care activities to very advanced activities such as running a mile. Well-being refers to perceptions such as pain and energy and how one feels about life in terms of happiness, anger, anxiety, depression, and global perceptions of quality of life (QOL). Interestingly, several authors report empirical associations between depressive symptoms and HRQOL without acknowledging that depressive symptoms are indicators of mental health and, therefore, HRQOL.[1]

Social support is not a measure of HRQOL, because HRQOL ends at the skin of the person being measured. In contrast, social function is an indicator of HRQOL because it indicates how well an individual gets along with family, friends, and others. That is, social function reflects the person's social health whereas social support represents whether the external environment is

supportive of him or her. While it may be difficult to separate social support from social function empirically, conceptually it is important to do so.

We also know that existing measures of social function are often very highly related to measures of mental health, and we have a hard time providing empirical support for a dimension of social health. This is due in part to the dearth of good measures of social functioning developed to date, but it also reflects a challenging measurement problem. For example, the SF-36 social functioning scale loads on both underlying mental and physical health factors and does not define a separate social health factor, but this is largely due to the fact that the social functioning items are worded with respect to physical and emotional problems: 1) During the past 4 weeks, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbors, or groups? 2) During the past 4 weeks, how much of the time has your physical health or emotional problems interfered with your social activities (like visiting with friends, relatives, etc.)?

HRQOL is one important component of QOL, but QOL encompasses additional constructs other than health. Patient-reported outcome (PRO) is a more general concept than HRQOL because it includes perceptions of QOL that extend beyond health and patient evaluations of care.[2]

Qualitative and Quantitative Methods

The typical HRQOL developmental cycle begins with a review of the literature, evaluation of existing measures, focus groups, compiling and drafting of items, cognitive interviews, revising the items, field testing and evaluation of psychometric properties, revising the items, and additional testing (focus groups, cognitive interviews, field testing) as needed. This mixed method approach is now the gold standard approach to instrument development. Qualitative methods are an important part of the developmental cycle, as they are used early and potentially again later. In contrast, many widely used measures such as the SF-36 concentrate more on quantitative than qualitative work.[3]

To the extent that it is possible to bring together the quantitative people with the survey research experts who are skilled at focus groups and cognitive interviewing techniques that lead to measures that are optimal in terms of respondent understanding, the better off we will be. Meshing together the different types of expertise is critical to successful instrument development.

Another important consideration is whether a measure that is to be used in different subgroups should be developed in a sequence or in a parallel process. In the United States, for example, we have developed many surveys in English first and then translated them into other languages, hoping they would work equally well. Alternatively, the WHO-QOL instrument was developed in parallel in multiple languages.[4,5] An advantage of the parallel approach is that can we anticipate better where items are going to fail because we included different subgroups early in the process. This could be especially beneficial with respect to differential item functioning.

Unidimensionality

The importance of having appropriate dimensionality is highlighted in the HRQOL field because of the increasing application of item response theory (IRT) methods. A fundamental assumption of IRT models is sufficient dimensionality (typically unidimensionality). In reality, there has long been a focus on dimensionality in HRQOL circles. We need to continue to do that with IRT and even with classical test theory methods, but this focus is not a new phenomenon.

One potential fruitful approach is to estimate a bifactor model for the data. In this model, each item is allowed to have a positive loading on a general factor that is assumed to underlie all the items. In addition, each item can load on a "group" (subscale) factor.[6] By comparing the loadings on the general factor in a bifactor model with those from a hypothesized single factor model, we can get a sense of whether the items are sufficiently unidimensional to satisfy the IRT unidimensionality assumption.

Aggregation of Health-Related Quality of Life Measures

There are many layers to consider in HRQOL measures. As we go up the hierarchy, we lose something in terms of detail in order to get at that summary level. But there are other advantages to summarization. For example, for some purposes, we need to make a bottom-line decision, and it is easier to do this with summary measures than with a plethora of profile information.

How the aggregation is done is very important. One decision is whether to derive summary indices using orthogonal (assuming underlying dimensions are uncorrelated) or oblique (estimating correlations among underlying dimensions) scoring. Some argue that your model should reflect the fact that physical health and mental health are correlated. That is, you shouldn't fit a model to the data that isn't going to fit. In fact, the use of an orthogonal model in deriving the SF-36 physical health and mental health summary scores has led to discrepancies between them and the eight scale scores.[7,8] If one's scores go up over time on all eight scales' scores, the "physical health" scales (physical functioning, role limitations due to physical health problems, pain, general health perceptions) have the expected effect of improving the physical health composite score (PCS), but the fact that the "mental health" scales (emotional well-being, role limitations due to emotional problems, social functioning, energy) also go up has the effect of lowering the PCS (see factor scoring weights in Table 1). If the mental health scales go up more than the physical health scales, this can cause the PCS to stay the same or possibly even go down. Recently, factor weights based on an oblique factor solution have been derived (see Table 2).[8]

Table 1: Standard (Uncorrelated Model) Factor Scoring Weights for the SF-36 Physical Component Summary (PCS) and Mental Component Summary (MCS) Scores

PCS_z = (PF_z * .42) + (RP_z * .35) + (BP_z * .32) + (GH_z * .25)

+ (EF_z * .03) + (SF_z * -.01) + (RE_z * -.19) + (EW_z * -.22)

MCS_z = (PF_z * -.23) + (RP_z * -.12) + (BP_z * -.10) + (GH_z * -.02)

+ (EF_z * .24) + (SF_z * .27) + (RE_z * .43) + (EW_z * .49)

Note: Weights derived by Ware and Kosinski (2001).[9]

Table 2: Alternative (Correlated Model) Factor Scoring Weights for the SF-36 Physical Component Summary (PCS) and Mental Component Summary (MCS) Scores

PCS_z = (PF_z * .20) + (RP_z * .31) + (BP_z * .23) + (GH_z * .20) + (EF_z * .13) + (SF_z * .11) + (RE_z * .03) + (EW_z * -.03)

MCS_z = (PF_z * -.02) + (RP_z * .03) + (BP_z * .04) + (GH_z * .10) + (EF_z * .29) + (SF_z * .14) + (RE_z * .20) + (EW_z * .35)

Note: Weights derived by Varon (2005).[8]

Disparities and Determinants

We have focused a great deal on medical determinants, but they don't really capture much of the variance in HRQOL. We need to expand the models we apply to our data. Donald Patrick talked about population health models. There are also social science models that can be brought to bear and additional thinking that will help to make more comprehensive models that extend beyond medicine.[10] For example, health behaviors may have important relationships with HRQOL.[11]

Minimally Important Difference

Another issue is minimally important difference (MID). We are interested in not just if a difference is statistically significant, but whether the difference is big enough to care about. Because of the uncertainty in estimating the MID, it has been recommended that multiple and preferably different kinds of anchors be examined and that bounded estimates of the MID be reported rather than forcing the MID to be a single value.[12] Interestingly, statistical significance is paramount when the focus is on individual rather than group change because the amount of change required for achieving statistical significance is so big.[13]

Evaluating Measures in Disadvantaged Populations

More attention is being paid to and needs to continue to be directed at how HRQOL measures work in different subgroups. The IRT method helps you to see if you are getting empirical equivalents. But the qualitative methods are also very useful in trying to figure out up front if

you can do it, whether a measure is going to work in different subgroups, and also, after the fact to figure out why you may be having troubles and why measures aren't equivalent.[14]

Matching Items to the Individual

IRT allows us to do a better job than classical test theory in having items that match our populations. So for some purposes, it may be fine if the items are close to the ceiling or close to the floor or in the middle of the scale, if that is the purpose they are going to be put to. But for many of the population applications, we need to cover the range quite well, because we are going to have people fitting into different places on that continuum. With IRT, we can also get better estimates of how well items discriminate between people.

Another benefit of IRT is that you can actually look to see—you are basically fitting the model to the data, and you are testing the model to see how well it corresponds to the data, which is a nice feature, whereas normally you don't actually test the fit of the assumed model.

You can also look to see how individuals fit the model, so that if you find the model doesn't fit the person, you don't necessarily use his or her data, or you may interpret the person's data in a different light.[15] It can tell you about careless responses. It can also tell you about how the model fits different types of people—you don't have to assume it works for everybody.

Readability of Surveys

Readability is something that hasn't really had enough attention so far. That is, readability of surveys and the literacy that is required and matching the items to the subpopulation you are studying. It hasn't been common practice to assess how readable surveys are, and we don't even know exactly the best way to do that. We have some very crude methods, but there is much work that could be done to really assure that a survey is not at too high a level for the population to which it will be administered.[16]

References

1. Mancuso C A, Peterson MG, Charlson ME. Effects of depressive symptoms on health-related quality of life in asthma patients. *J Gen Intern Med.* 2000;15:301-310.

2. Willke R J, Burke LB, Erickson P. Measuring treatment impact: a review of patient-reported outcomes and other efficacy endpoints in approved product labels. *Control Clin Trials.* 2004;25:535-552.

3. Stewart AL, Sherbourne CD, Hays RD, et al. (1992). Summary and discussion of MOS measures. In: Stewart AL, Ware JE, eds. *Measuring Functioning and Well-Being: The Medical Outcomes Study Approach.* Durham, NC: Duke University Press; 1992:345-371.

4. WHOQOL Group. Study protocol for the World Health Organization project to develop a Quality of Life assessment instrument (WHOQOL). *Qual Life Res.* 1993;2:153-159.

5. WHOQOL Group. Development of the World Health Organization WHOQOL-BREF quality of life assessment. *Psychol Med.* 1998;28:551–558.

6. Reise SP, Morizot J, Hays RD. The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Qual Life Res.* (in press).

7. Taft C, Karlsson J, Sullivan M. Do SF-36 summary component scores accurately summarize subscale scores? *Qual Life Res.* 2001;10:395-404.

8. Varon SF. Interpreting health-related quality of life measures—summary scores and the minimally important difference [dissertation]. Los Angeles: University of California; 2005.

9. Ware JE, Kosinski M. *SF-36 Physical and Mental Health Summary Scales: A Manual for Users of Version 1.* 2nd ed. Lincoln, RI: QualityMetric, Inc;2001.

10. Glanz K, Rimer BK, Lewis FM, eds. *Health Behavior and Health Education: Theory, Research, and Practice.* 3rd ed. San Francisco: Jossey-Bass;2002.

11. Stewart AL, Hays RD, Wells KB, et al. Long-term functioning and well-being outcomes associated with physical activity and exercise in patients with chronic conditions in the Medical Outcomes Study. *J Clin Epidemiol.* 1994;47:719-730.

12. Hays RD, Farivar SS, Liu H. Approaches and recommendations for estimating minimally important differences for health-related quality of life measures. *COPD.* 2005;2:63-67.

13. Hays RD, Brodsky M, Johnston MF, et al. Evaluating the statistical significance of health-related quality of life change in individual patients. *Eval Health Prof.* 2005;28:160-171.

14 Scott NW, Fayers PM, Bottomley A, et al. Comparing translations of the EORTC QLQ-C30 using differential item functioning analyses. *Qual Life Res.* 2006;15:1103-1115.

15. Emons WHM., Glas CAW, Meijer RR, et al. Person fit in order-restricted latent class models. *Appl Psychol Meas.* 2003;27, 459-478.

16. Calderón JL, Morales LS, Liu H, et al. Variation in the readability of items within surveys. American Journal of Medical Quality. *Am J Med Qual* 2006;21:49-56.

Contact Information:

Ron D. Hays, Ph.D., Professor of Medicine, UCLA School of Medicine; Senior Health Scientist, RAND Health Program, drhays@ucla.edu.

Acknowledgment:

Preparation of this paper was supported in part by the UCLA/DREW Project EXPORT, National Institutes of Health, National Center on Minority Health & Health Disparities, (P20-MD00148-01) and the UCLA Center for Health Improvement in Minority Elders/Resource Centers for Minority Aging Research, National Institutes of Health, National Institute of Aging (AG-02-004).

Current State of the Art in Preference-Based Measures of Health and Avenues for Further Research

John Brazier, Ph.D., M.Sc.

Introduction

Preference-based measures of health (PBMH) have been developed primarily for use in economic evaluation. They have two components: a standardized, multidimensional system for classifying health states and a set of preference weights or scores[1] that generate a single index score for each health state defined by the classification, where full health is one and zero is equivalent to death. A health state can have a score of less than zero if regarded as worse than being dead. These PBMH can be distinguished from non–preference-based measures by the way the scoring algorithms have been developed, in that they are estimated from the values people place on different aspects of health rather than a simple summative scoring procedure or weights obtained from techniques based on item response patterns (e.g., factor analysis or Rasch analysis).

The use of PBMH has grown considerably over the last decade with the increasing use of economic evaluation to inform health policy, for example, through the establishment of bodies such as the National Institute for Clinical Excellence in England and Wales[2] and the Health Technology Board in Scotland[3] as well as similar agencies in Australia[4] and Canada.[5] Preference-based measures have become a common means of generating health state values for calculating quality-adjusted life years (QALY). The status of PBMH was considerably enhanced by the recommendations of the U.S. Public Health Service Panel on Cost-Effectiveness in Health and Medicine to use them in economic evaluation.[6] A key requirement for PBMH in economic evaluation is that they allow comparison across programs.

Although PBMH have been developed primarily for use in economic evaluation, they have also been used to measure health in populations. PBMH provide a better means than a profile measure of determining whether there has been an overall improvement in self-perceived health.

The preference-based nature of their scoring algorithms also offers an advantage over non–preference-based measures since the overall summary score reflects what is important to the general population. A non–preference-based measure does not provide an indication to policy makers of the overall importance of health differences between groups or of changes over time.

The purpose of this paper is to critically review methods of designing preference-based measures. The paper begins by reviewing approaches to deriving preference weights for PBMH, and this is followed by a brief description and comparison of five common PBMH. The main part of the paper then critically reviews the core components of these measures, namely the classifications for describing health states, the source of their values, and the methods for estimating the scoring algorithm. The final section proposes future research priorities for this field.

Approaches to Obtaining Preference Weights for Measures of Health

There are three empirical approaches to deriving preference weights for measures of health: (1) empirical mapping onto an existing PBMH, (2) mapping onto a respondent's own health valuation, and (3) asking respondents to value states defined by the health measure. Although existing preference-based measures are currently based on the third approach, it is important to understand the potential role of the alternatives.

Mapping between measures

The approach of empirically mapping a health measure onto a PBMH so as to obtain preference weights for the former has been used in numerous studies. This approach requires the PBMH and the non–preference-based measure to be administered to the same population. The approach was used by Fryback et al.[7] in the Beaver Dam study and Nichol et al.,[8] both of whom mapped the generic SF-36 onto preference-based measures. Tsuchiya et al.[9] and Brazier et al.[10] have mapped PBMH onto condition-specific measures. These studies have regressed the dimension scores of the non–preference-based measure onto the preference-based measure.

This approach can be pragmatically useful, but it makes several assumptions. First, it assumes that the items used to score the dimensions have equal importance, and second, that the intervals between the response choices are equally important to people. These assumptions can be relaxed by modelling each item response as a dummy variable.[9] A more important limitation is the assumption that the preference-based measure covers all the important aspects of health covered by the non–preference-based measure. If it does not, important dimensions of health might not be valued appropriately. Mapping onto an existing PBMH should be viewed as a second best compared to the direct valuation of the health measure.

Valuing a health measure using direct values for health states

This approach involves administering the health measure alongside a valuation question about the respondent's own health. Such an approach was taken by Lundberg and colleagues[11] and involved administering the SF-12 health status questionnaire alongside a self-administered version of the time trade-off (TTO) technique in a postal survey of 8,000 members of the general population. The TTO question asks respondents to consider the number of years of life they would be willing to sacrifice in order to live the remainder of their life in full health. SF-12 item responses (among other variables such as age) were regressed against their TTO value to provide a set of preference weights for the SF-12.

This study was limited because the health states were those of a random sample of the general population. The sample of states valued in the survey was therefore not determined by any statistical design but by their natural occurrence in the population. As many severe states are quite rare, this reduces the ability of the model to predict values for more severe states. These limitations could be partly overcome by careful sampling of people with a wide range of conditions, but there will always be people with some medical conditions who cannot participate in such a survey, necessitating the use of proxies.

In conclusion, this approach captures only the values of those people in the states. This may be seen as a problem for those seeking to implement the Washington Panel's recommendations to use "societal values." The issue of whose values is examined later in this paper.

Hypothetical valuation of health states

The approach adopted by PBMH developers to date involves asking respondents (typically adult members of the general population) to imagine being in a health state. They could be asked to imagine the state on someone else's behalf, perhaps as a proxy or by taking a third-party perspective (such as behind a 'veil of ignorance'), but most applications ask respondents to imagine they are in the state. This has a logistical advantage over the previous direct approach because it allows a single respondent to value numerous states, and the researcher can select the states being valued according to a proper statistical design. This approach conforms most closely to the recommendations of the Washington Panel.

Existing Preference-Based Measures of Health

The five preference-based measures considered in this section are the Quality of Well-Being (QWB) scale,[12] the Health Utility Index (HUI) versions two and three (HUI-2 and HUI-3),[13, 14] EQ-5D,[15, 16] and the SF-6D—a derivative of the SF-36 and SF-12.[17, 18] These instruments were chosen because they are the most widely used. In Table 1, which summarizes their characteristics, we see that existing preference-based measures differ in the content of their descriptive systems, the valuation technique, and the method of extrapolation. And yet their descriptive systems share a common structure, as they all have multilevel dimensions. Furthermore, despite differences in methods of valuation, all preference data were obtained from a sample of the general population (although the HUI-2 was valued by a random sample of parents from Hamilton, Ontario).

Table 1: Characteristics of multi-attribute utility scales

MAUS	*Descriptive characteristics*			*Valuation characteristics*			
	Dimension	**Levels**	**Health states**	**Valuation technique**	**Method of extrapolation**	**Sample**	**Country**
QWB	Mobility, physical activity, social functioning	3	1,170	VAS	Statistical	866 (general population)	USA (San Diego)
	27 symptoms/problems	2					
HUI-2	Sensory, mobility, emotion, cognitive, self-care, pain	4–5	24,000	VAS tranformed into SG	MAUT	203 (parents)	Canada (Hamilton), UK
	Fertility	3					
HUI-3	Vision, hearing, speech, ambulation, dexterity, emotion, cognition, pain	5–6	972,000	VAS tranformed into SG	MAUT	504 (general population)	Canada (Hamilton), France
EQ-5D	Mobility, self-care, usual activities, pain/discomfort, anxiety/depression	3	243	TTO and VAS	Statistical	3,395 (general population)	UK, Japan, Spain, USA (among others)
SF-6D	Physical functioning, role limitation, social functioning, pain, energy, mental health	4–6	18,000	SG	Statistical	611 (general population)	UK, Japan, Hong Kong

Note: VAS—visual analogue scale, TTO—time trade-off, SG—standard gamble, MAUT—multi-attribute utility theory.

All five instruments purport to be generic. Even so, they differ in defining health and in the dimensions they cover. The developers of the HUI instruments restrict health to "beneath the skin" and exclude those consequences for quality of life (QOL) affecting the person's functioning in society, such as role and social functioning. By contrast, the QWB, SF-6D, and EQ-5D include "out-of-skin" aspects of health. The HUI-2 was designed for children and the rest for adults. The instruments also differ greatly in size, with between five and eight dimensions and two to six levels on each dimension, resulting in the number of potential health states ranging from 243 for the EQ-5D to 972,000 for the HUI-3.

The instruments use different valuation techniques to elicit preferences, including the visual analogue scale (VAS), standard gamble (SG), and TTO (see definitions in next section). The HUI-2 and HUI-3 health states were not directly valued using SG but rather via a transformation of VAS values. The five measures were also valued using different variants of the valuation techniques. Finally, all five measures have too many states to be valued in one survey (with the possible exception of EQ-5D), and thus values must be extrapolated from a sample of states defined by their respective classifications. This was done for three of the measures by statistical modelling (EQ-5D, QWB, and SF-6D), and the HUI-2 and -3 use multi-attribute utility theory (MAUT). Valuation work has been replicated across a number of countries for the EQ-5D, SF-6D, and HUI-2 and -3.

Comparison of Measures

This section briefly compares the five PBMH in terms of practicality, reliability, and degree of agreement.

There is little to choose between the instruments on the basis of practicality, with each now having self-completed versions. The EQ-5D has just five questions, but it is closely followed by the 15 items for the HUIs and 36 (or just the 11 used to construct the classification) for the SF-6D. Stability over time in those whose health has not changed has been demonstrated for most instruments, and there is no reason to suppose them to be worse than non–preference-based

measures.[19] The standard deviations of the preference scores, however, are larger for the EQ-5D and HUI-3 than the SF-6D,[20–22] because the EQ-5D and HUI-3 cover a larger range of states.

Content validity depends on the aspects of health the user wishes to cover and also on the disease group and age of the patients. There are also issues about perspective and whether social health is relevant. It is also evident from the descriptions that the SF-6D may suffer from "floor effects" on dimensions of physical functioning and role limitation, where many respondents choose the lowest response category, and this has been borne out in recent empirical comparisons.[20–22] Conversely, concern has been expressed that the EQ-5D suffers from a ceiling effect because large numbers of patients are given states of full health.[23]

Among those who use a choice-based valuation technique, the HUI-2 and -3 and SF-6D might be preferred to the EQ-5D by those who regard the SG as the "gold standard" and EQ-5D by those who prefer TTO. The SG utilities for the HUIs have been derived from VAS values using a power transformation that has been criticized (see above). The valuation of the HUI has also been obtained from a smaller and less representative sample of the general population than the valuation survey of the EQ-5D in the United Kingdom and United States. A further difference between them is the methods of modelling the values for health states, with the HUIs using MAUT and the rest using statistical inference. The review in the next section concludes that there is little evidence on their relative predictive performance.

The measures seem to be moderately correlated at around 0.5 to 0.6, and the differences in mean values for patient groups are often just 0.05 on the zero-to-one utility scale, but this makes for significant systematic variation between instruments.[20, 21] Comparisons of the SF-6D with the EQ-5D and HUI-3 have shown the former to generate higher values for more severe states and lower values for the mildest states. Furthermore, these cross-sectional differences were found to translate into significant differences in the size of change measured over time in one patient group.[20] The variation is a product of the differences between the instruments, but it is not possible to determine whether this variation is driven by differences in their descriptive systems, techniques of valuation, or methods of extrapolation.

The next section critically reviews the three components of PBMH: the descriptive systems, valuation techniques, and methods of extrapolation.

Review of Methods

Descriptive systems

Two important issues in designing a descriptive system for a generic preference-based measure are (1) the definition of health and (2) the construction of descriptive systems for PBMH.

Definition of health

The developers of the HUI advocate a "beneath the skin" definition of health that excludes social activities and work (e.g., social and role dimensions on the SF-6D and usual activities in the EQ-5D). Social and role activities are deemed the result of the personal preferences of the respondent as well as his/her state of health and, it has been argued, should be excluded from the descriptive system. For the HUI measures, this also helps to achieve orthogonality or independence between attributes in the classification of health state, which is important for the application of MAUT and, to a lesser extent, the statistical approaches.

There is a long history of having "out of skin" consequences in measures of health, due in part to the original WHO (World Health Organization) definition that included social health.[24] It could also be argued that the impact on role and social activities is important in helping respondents in a valuation survey to fully understand the impact of a health state on their QOL. It effectively reduces the imaginative workload being demanded of respondents.

This debate has interesting parallels in the QOL literature in general through the notion of "response shift,"[25] where the impact of a health state on a person's QOL is not stable. A sudden change in health may initially have a substantial impact, but gradually people learn to cope and adapt to their limitations in a number of ways, and over time the impact lessens. Those measures having role and social dimensions are likely to be even more prone to this "response shift."

Whether adaptation should be excluded from the final values given to states is a normative question addressed in a later section.

Most PBMH have a generic descriptive system.[14, 17, 26] These have been found, however, to be inappropriate or insensitive for many medical conditions.[27–29] Condition-specific descriptions may be more sensitive to changes in the condition than generic measures and more relevant to the concerns of patients.[30] There is a concern, however, that condition-specific PBMH fail to achieve comparability. Differences in scores between measures (whether generic or condition specific) result from differences in the methods of valuation as well as the descriptive system. In principle, if the descriptive system is valued on the same full health–dead scale using the same variant of the same valuation technique employing a comparable population sample, the valuations should be comparable. Any remaining differences in values should be a legitimate consequence of the descriptive system. This, however, assumes that the value of a dimension is independent of those dimensions outside of the descriptive system, and this requires empirical testing.

Construction and validation of descriptive systems for PMBH

Although there are difficulties in establishing the validity of the values generated by a preference-based measure (as discussed below), it is important to show that a descriptive system accurately describes the health state. Little has been written in the economics literature about this aspect of validity.[31] Published economic evaluations rarely address the issue even though small differences in descriptions of health states can substantially alter the results. There has been some criticism expressed of specific generic preference-based measures but little systematic evaluation of their descriptive systems. The assessment of the validity of the descriptive system of a preference-based measure should be undertaken with the same rigor as would be applied to non–preference-based measures. There are some important differences of principle, however, and these are explained below.

Content and face validity

The psychometric criteria of content and face validity, although subjective, are important in assessing the comprehensiveness, relevance, and sensitivity of the dimensions of preference-based measures. Content, in terms of dimensions and items, limits the attributes being covered by the measure. Economists, who are concerned with ensuring that the measure correctly reflects a person's utility function, will prefer an approach that generates items and dimensions from patients. This has been pursued to a limited extent in the development of some existing instruments, but most preference-based measures were constructed primarily by teams of experts.

Item Response theory (IRT)

IRT provides a powerful technique for understanding the relationship between items. It is based on the assumption that item responses are determined by their degree of difficulty along some spectrum of a unidimensional construct. This has been very helpful in understanding the relationship between items and has been used, among other purposes, to assist in selecting items for an instrument and in scoring the items. Recent years have seen a rapid expansion of the application of IRT to the construction and testing of non–preference-based measures.

IRT has potential in helping to construct the descriptive systems of preference-based measures. It was used in the selection of items for the physical functioning dimension of the SF-6D from the 10 items of this dimension in SF-36[17] By pooling items from different PBMH and non–preference-based measures, IRT can assist in understanding the severity range covered by the descriptive systems of the preference-based measures. IRT may identify significant gaps in the descriptive classifications and provide candidate items for improving them.

It is important to understand the limitations of IRT, however. It would make no sense to apply IRT across dimensions for a preference-based measure, anymore than it would for a non–preference-based measure. Furthermore, while the advocates of IRT claim it generates an interval scale, and this might be true for the health concept being measured, it does not provide a measure of strength of preference with the required interval properties.

Construct validity

Having an empirically based means of testing the descriptive validity of an instrument is important. Construct validation is appropriate for testing the validity of the description of health or health change underlying a PBMH. The ability of an instrument to reflect known or expected differences in health is an essential precursor to its ability to reflect preferences. Such tests should be undertaken on the unscored descriptions of an instrument, however. Otherwise, there is a danger that the failure of a score to detect a difference found by a condition-specific QOL measure is incorrectly interpreted to imply that the descriptive component of the preference-based instrument is insensitive. The score of a preference-based measure may fail to detect the difference simply because this difference is not valued by patients.

Valuation

The methods of valuation underpin preference-based measures and have been a topic of considerable debate in the health economics literature. The key areas of concern have been the choice of technique, the variant of the technique, the problem of states worse than death, and the appropriate sources of values.

Valuation Technique

The five preference-based measures reviewed in this paper have used VAS, SG, or TTO, and much has been written on the relative virtues and flaws in each of these techniques. This section attempts to summarize this well-trodden path.

Several informative reviews have compared the VAS, SG, and TTO techniques in terms of practicality, reliability, and validity.[32–36] Generally, all three techniques have been reported to be practical and acceptable for most populations, but VAS is considered marginally better in terms of response rate, cost, and consistency of responses.[33] On the other hand, the validity of VAS as a measure of the strength of preference has been challenged.[37–39] The main criticism is that a rating scale does not confront the respondent with the notion of opportunity cost and so does not reflect the economist's notion of strength of preference. Interviews with respondents indicate that

they did not intend it to reflect their preferences.[37, 39, 40] When asked, respondents talk about concepts of fitness or the natural history of illness and not the value of health.

SG and TTO present respondents with a choice and offer a more theoretically appealing measure of strength of preference (i.e., involving opportunity cost). SG asks respondents to make a choice between alternative outcomes where one of them involves uncertainty. Respondents are asked how much risk in terms of probability of death or some other bad outcome they are willing to accept to avoid living in the certainty of the health state being valued. This technique is based on the Expected Utility Theory of decision making under uncertainty developed by Von Neumann and Morgenstern,[41] which rests on a set of axioms about the nature of individual preferences when prospects are uncertain. The TTO technique, developed as an alternative to SG,[32] was designed to overcome the problems of explaining probabilities to respondents. The TTO technique asks the respondent to choose between two alternatives with certain prospects, that is, shorter years (x) in full health and longer years (t) in the health state being valued. Respondents are asked to consider trading a reduction in their length of life ($t - x$) for an improvement in health. The health state valuation is the fraction of healthy years equivalent to a year in a given health state, or x/t.

SG has the most rigorous foundation in theory in the form of the Expected Utility Theory of decision making under uncertainty. There are theoretical arguments against SG in valuing health states, however,[34] and little empirical support for Expected Utility Theory.[42,43] There are also concerns about the empirical basis of the TTO technique. Some evidence suggests that the time spent in a health state and the time at which it occurs affect TTO values.[44, 45] A key review by Bleichrodt[46] summarized the different sources of bias in each of these techniques and concluded that SG is subject to mainly upward bias, whereas TTO is subject to upward and downward bias, and so he concludes that TTO may be preferred overall. There is currently no consensus regarding the best technique. One solution might be to try to correct for these biases.[47]

A further consideration is that the SG "utilities" of the HUI-2 and -3 are derived from VAS valuations on the basis of an estimated power function where the difference between VAS ratings and SG utilities is assumed to be a person's attitude toward risk. This conclusion was

based on the suggestion of Dyer and Sarin,[48] but the validity of the power transformation has been questioned in the literature. Dolan and Sutton[49] and Stevens et al.[50] have all demonstrated that other specifications fit the data as well as or in some cases better than a power function.

More recently, there has been interest in basing valuations on ordinal data from ranking and discrete choice experiments. The use of ordinal individual data to generate cardinal health state values draws on random utility theory. It may prove to be a promising alternative to SG and TTO, particularly in more vulnerable populations, because empirical work has found that it can produce similar values for the EQ-5D, SF-6D, and HUI-2.[51, 52]

Variant of the technique

Differences in variant may prove more important than choice of technique. It has been shown in numerous studies that SG responses are subject to framing effects, such as whether the probabilities are expressed in terms of success or failure. SG has been found to generate inconsistent valuations with changes to the lower anchor.[53]

There is a wide range of variants, including (1) mode of administration (interview or self-completion, computer or paper administration), (2) the use of props, (3) presentation of probabilities, (4) time allowed for reflection, and (5) individual versus group interviews. Few publications in the health economics literature have compared these alternatives, although several researchers have undertaken considerable efforts to try to improve the quality of the data from SG and TTO.[35, 54] The evidence is that values for health states vary considerably between variants of the same technique.[49, 55] Indeed, it has been found that between-variant differences can be more important than between-technique differences.

Evidence that the way people are asked about their preferences has a major impact on the results raises questions about the nature of people's preferences for health states. It suggests that people do not have well-defined preferences about health before the interview; rather, their preferences are constructed during the interview. This would account for the apparent willingness of respondents to be influenced by the precise framing of the question.

A common criticism of the current methods for eliciting preference is that the tasks are cognitively complex, with respondents being asked to consider variations in up to eight health dimensions alongside a life-and-death scenario involving probabilities of survival. Evidence from psychology suggests that respondents faced with such complex problems would tend to adopt simple heuristic strategies.[56] This would be particularly true where respondents have little time to consider their real underlying values.

The foregoing underscores the need to develop respondent-friendly methods. A well-conducted study involving the elicitation of preferences should fully explain the task to the respondent and undertake a practice question. Unfortunately, respondents are typically expected to evaluate health states in one sitting, with little time for reflection. Respondents need more time and support to reflect on their values in order to process such complex information. It has been suggested that respondents could be re-interviewed after they have had time to deliberate on the health state in question. Shiell and colleagues[57] found significant differences between interviews for the same states, with values tending to be higher at subsequent sittings. An implication may be to move away from the current large-scale surveys of members of the general public involving a single sitting to smaller-scale studies of panels from the general public who are better trained and more experienced in the techniques and who are given time to fully reflect on their valuations.

States worse than dead

For states deemed worse than dead, the SG technique asks the respondent to choose between the prospect of death for certain and the uncertain prospect of full health or the state being valued. The probability of full health is varied until the point of indifference where the value of the state worse than death is $-P/(1 - P)$. The analogous TTO question asks respondents to choose between the first alternative of dying immediately and the second alternative of some number of years in the state being valued followed by a number of years in full health (where the two periods sum to t). The time in full health x is varied until the person is indifferent between these alternatives. The value for the state worse than death is then given by $-x/(x - t)$. These two formulas, together with the method for calculating the value of states better than dead, produce a range of values between +1 and $-\infty$, which gives greater weight to negative values in the calculation of mean

health state scores and presents problems in statistical analysis. To reduce the influence of extreme negative outliers, Patrick and colleagues[58] proposed transforming the values to limit the range from +1 to −1.

It is important to consider states worse than being dead because they are common among preference-based measures. The U.K. TTO valuation of the EQ-5D, for example, produced mean scores below zero for one third of all states. Evidence from the U.K. valuation of the EQ-5D suggests a discontinuity around zero that appears to indicate a special significance is attached to this value. Once people regard a health state as worse than death, they are willing to give quite a low value. Doubts must exist about whether the scale has the same interval properties either side of zero. More research is needed into the valuation of states worse than death.

Source of values

There is evidence of significant variation in values by disease experience, age, and education. In general the evidence points to patients giving health states a higher value than do members of the general population.[59–61]

The Washington Panel argued that using the values of the general population favors patients because general population values give a larger value to treatments that restore patients to full health. Lenert and colleagues, however, have demonstrated that values of the general population may be less sensitive to movements between points at the lower end of health because of a reference point effect.[62] Furthermore, the lower health state valuations of the general population work against the interests of patients for life-saving interventions since these lower values result in a smaller gain in QALY.

The main normative argument for using general population values seems to hinge on the view that in a publicly funded health care system, it is society's resources that are being allocated, and therefore it is the view of the general population that is relevant. In a similar way, it can be argued that enrollees of insurance schemes should be asked to provide values in the context of decisions within a private insurance program rather than patients. By contrast, it has been

suggested that the values of patients should be used because they are in the best position to know their own state.[63]

The choice of viewpoint is ultimately a value judgement, but it may depend on the reasons for the discrepancy. The main cause of the difference between the values of patients and the general population is adaptation to the condition. Patients experiencing long-term conditions might also change their life goals and expectations. These are aspects of the "response shift" recognized in the QOL research literature mentioned earlier in this paper. These adaptations will be related to the length of time a patient experiences the condition.

Members of the general population know little about such adaptation. The choice between patient and general population values really comes down to the extent to which these changes should be taken into account. Menzel et al.[64] tried to distinguish between "laudable" adaptations, such as enhancing one's skills, adjusting activities, and even altering perceptions of health, and less desirable changes, such as cognitive denial of functional health and suppressed recognition of full health. There is also a genuine concern that many of the changes listed are the result of laudable effort, and incorporating them into health state values in resource allocation may work against patients' interests, which seems in some sense unfair.

It seems difficult to justify using just patient values or uninformed members of the general population to obtain preferences for health measures. Menzel and colleagues[64] suggest that more research is required into the causes of adaptation. They also suggest that patients should be consulted on the extent to which they want their adapted values to be used in decision making. This empirical research would not address the normative question of what aspects of adaptation should ultimately be used. Menzel et al. suggest rather ambitiously that perhaps the general population may be able to disentangle appropriate from inappropriate adaptation. This is consistent with the recommendations of the Washington Panel, which advocate the use of informed general population values. A middle way could be to provide members of the general population with patients' values before asking them for their own values. There is important empirical work to be done to develop methods for conveying such information to the general population and to measure its impact on health state values.

The question of whether values from one country or culture can be used in another is also an important one. The emerging evidence suggests that VAS valuations do not vary much between countries, but there are significant differences between countries in TTO values for EQ-5D[9, 16, 65] and SG valuations of the SF-6D[27] and HUI-2.[52] It seems there are significant differences between countries in health state values and important variations by sociodemographic characteristics and ethnic group.[16, 66]

Method of extrapolation

There have been two approaches to estimating a function for valuing states from a health state classification system, the decomposed and composite approaches.[33]

The decomposed approach employs MAUT to determine the functional form and the sample of states to be valued. MAUT substantially reduces the valuation task by making simplifying assumptions about the relationship between dimensions. The most commonly used specifications are the additive and multiplicative functional forms. The simple additive functional form assumes dimensions to be independent and hence permits no interaction. This was found by Torrance et al.[13] to be invalid, and the multiplicative function has been used to value the HUI-2 and -3. The multiplicative function permits a very limited form of interaction between dimensions by assuming the interdependency is the same between all dimensions and for all levels of each dimension.

The application of MAUT decomposes the valuation task into three parts. First, each dimension is valued separately to estimate single-attribute utility functions. Second, “corner states” are valued; these are states where one dimension is at one extreme (usually the worst level) and the rest are set at the other (usually the best) level. Such corner states may represent infeasible combinations of dimension levels unless the descriptive system ensures the dimensions are truly independent. A failure to achieve this with the HUI-2 resulted in a complex “backing-off” procedure.[13] Third, a set of multi-attribute states determined by the model specification is valued. A single respondent undertakes all tasks for the HUI-2 and -3.

The composite approach requires a larger sample of states to estimate a function by regression. A common method for sampling states is to use an orthogonal design for estimating an additive model. There are problems, however, with determining the states required to estimate a model with interaction terms. To date, researchers have typically added extra states at random. An important piece of research will be to develop more sophisticated algorithms for sampling health states in order to value key interactions. The statistical models developed for the EQ-5D and SF-6D have estimated crude summary terms for interactions, such as dummy variables taking a value of one for states containing at least one dimension at its worst level.[15, 17]

The composite approach requires more states than can be valued by a single respondent. The sample states are allocated between respondents, and thus it is necessary to disentangle variation between respondents from variation between states. The statistical modelling has to cope with this hierarchical structure to the data set, and researchers have done this using random effects techniques or by modelling mean health state values.[15, 17] Modelling also has to cope with a highly skewed data set. Work in this area has explored a range of transformations to overcome this problem, but none has been found to improve the models. Recently, work has been undertaken to use a Bayesian approach that applies a nonparametric method to estimating posterior mean health state values with variances, and the first application to the SF-6D has proved very successful.[67]

There has been little written comparing statistical and MAUT approaches. The MAUT multiplicative models are (in a limited sense) more sophisticated than the additive EQ-5D, but they are based on the valuation of a far smaller number of health states. The MAUT approach uses deterministic models and does not allow for the pattern of the error structure. For HUI-2 and -3, VAS was used to value the health states, which means the transformation may introduce another source of error. In principle, however, it is possible to apply the MAUT approach using SG (or TTO) directly.

The choice between these approaches must rest on their ability to predict health state values in an independent sample. A comparison of the MAUT and statistical approaches was undertaken two

decades ago in a study of job choice by Currim and Sarin.[68] The authors found the statistical approach outperformed a multiplicative algebraic model: the correlation between actual and predicted choices across jobs using SG utility values was 0.64 and 0.16, respectively. This was a very limited study, however, and not in the context of health. A recent study by McCabe[69] applied the MAUT and statistical approaches to the U.K. valuation of the HUI-3 and found that the statistical approach was marginally better in terms of absolute mean error and percent within the range of plus or minus 0.1 and 0.05 of actual values. This evidence is also not conclusive because it is also influenced by the VAS-SG mapping.

The ensuing debate between MAUT and the statistical approach requires a head-to-head comparison where the two are used optimally, with the statistical approach being based on a better sampling procedure and the MAUT using SG or TTO in a direct fashion.

Future Avenues for Research

Descriptive systems of existing PBMH

The psychometric properties of existing PBMH need to be better understood through head-to-head comparisons with each other and with non–preference-based measures of health. These data would permit the application of IRT and classical psychometric assessments of the validity of the descriptive systems of these instruments.

Comparison of existing measures

There is currently research under way to compare the PBMH in different patient groups. This will provide a better understanding of the relationship between existing measures and should contribute to an understanding of the reasons for the differences between these measures.

Source of values

There are two related pieces of research into this component. The first will be to elicit patient health values alongside PBMH across a wide range of patient groups. This will help us understand the differences between patient and general population values and how the differences vary between medical condition and other background variables. The second would be to examine the use of informed general population values. This would require research into using patient values in valuing health states among the general population.

Methods for eliciting preferences

There is increasing interest in using ordinal tasks, like ranking or pairwise comparison, to derive the values of health states, and these could prove particularly valuable in enfranchising the more vulnerable groups. Also, the potential role of reflection and deliberation in the valuation of health states needs to be explored in future valuation surveys of PBMH.

States worse than being dead

Research is needed into developing ways to value states worse than being dead that lie on the same scale as states better than being dead.

Estimation

Further research is needed into estimating preference-based index measures from a sample of health state valuations. MAUT should be applied directly using SG and TTO, and statistical modelling can be improved including the use of Bayesian approaches. The different approaches then need to be compared.

Compare existing measures or develop a new one?

For any major program of research, an important question is whether to use existing measures. In the short-to-medium term, the use of existing measures, such as the recently completed valuation

of the EQ-5D and SF-36, makes good use of existing data sets. In the longer term, however, there could be a case for developing new measures drawing on more recent psychometric literature (including IRT) and valuation (such as the use of ordinal methods). This research is particularly important in the more vulnerable groups such as children, the very elderly, and people with major mental health problems, where existing measures are often inappropriate.

Conclusion

PBMH have come a long way over the last 20 years and offer an important set of tools for economic evaluation and other uses of summary health measures. Existing instruments differ in their descriptive systems and methods of valuation, and so they often generate different scores. Recent developments in assessing and valuing health status provide an important basis for improving preference-based health measurement and for developing more appropriate instruments for special groups, including vulnerable groups such as the very young, the very elderly, and people with serious mental health problems. The challenge is to ensure that new instruments provide a means of generating standardized and comparable scores across populations.

Acknowledgements

I am grateful to colleagues at HEDS who contributed to a 'brainstorm' session on an outline draft of this paper and to subsequent drafts. I would particularly like to acknowledge Professors Chris McCabe, Paul Dolan, and Jennifer Roberts and Dr. Aki Tsuchiya and Katherine Stevens. The views expressed in this paper and all remaining errors are of course mine. The author gratefully acknowledges funding from the U.K. MRC HSRC.

References

1. Drummond MF, O'Brien B, Stoddart GL, et al. *Methods for the Economic Evaluation of Health Care Programmes.* 2nd ed. Oxford: Oxford University Press; 1997.

2. National Institute for Clinical Excellence. *Guide to the Technology Appraisal Process.* London: National Institute for Clinical Excellence; 2001.

3. Health Technology Board for Scotland. *Guidance for Manufacturers on Submission of Evidence Relating to Clinical and Cost-Effectiveness in Health Technology Assessment.* Glasgow: Health Technology Board for Scotland; 2002.

4. Commonwealth Department of Health, Housing and Community Services. *Guidelines for the Pharmaceutical Industry on the Preparations of Submissions to the Pharmaceutical Benefits Advisory Committee.* Canberra: Australian Government Printing Service; 1992.

5. Ontario Ministry of Health and Long-Term Care. *Ontario Guidelines for Economic Analysis of Pharmaceutical Products.* Toronto: Queen's Printer for Ontario; 1994.

6. Gold MR, Siegel JE, Russell LB, et al., eds. *Cost-Effectiveness in Health and Medicine.* New York: Oxford University Press; 1996.

7. Fryback DG, Dasbach ED, Klein R, et al.Health assessment by SF-36, quality of well-being index and time tradeoffs: predicting one measure from another. *Med Decis Making.* 1992;12:348-356.

8. Nichol MB, Sengupta N, Globe DR. Evaluating quality-adjusted life years: estimation of the health utility index (HUI2) from the SF-36. *Med Decis Making.* 2001;21:105-112.

9. Tsuchiya A, Ikeda S, Ikegami N, et al. Estimating an EQ-5D population value set: the case of Japan. *Health Econ.* 2002;11:341-353.

10. Brazier JE, Kolotkin RL, Crosby RD, et al. Estimating a preference-based single index for the Impact of Weight on Quality of Life-Lite (IWQOL-Lite) instrument from the SF-6D. *Value Health.* 2004;7:490-498.

11. Lundberg L, Johannesson M, Isacson DG, et al. The relationship between health-state utilities and the SF-12 in a general population. *Med Decis Making.* 1999;19:128-140.

12. Kaplan RM, Anderson JP. A general health policy model: update and applications. *Health Serv Res.* 1988;23:203-235.

13. Torrance GW, Feeny DH, Furlong WJ, et al. Multiattribute utility function for a comprehensive health status classification system. Health Utilities Index Mark 2. *Med Care.* 1996;34:702-722.

14. Feeny D, Furlong W, Torrance GW, et al. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. *Med Care.* 2002;40:113-128.

15. Dolan P. Modeling valuations for EuroQol health states. *Med Care.* 1997;35:1095-1108.

16. Shaw JW, Johnson JA, Coons SJ. US valuation of the EQ-5D health states: development and testing of the D1 valuation model. *Med Care.* 2005;43:203-220.

17. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. *J Health Econ.* 2002;21:271-292.

18. Brazier JE, Roberts J. The estimation of a preference-based measure of health from the SF-12. *Med Care.* 2004;42:851-859.

19. Brazier J, Deverill M, Green C. A review of the use of health status measures in economic evaluation. *J Health Serv Res Policy.* 1999;4:174-184.

20. Hatoum HT, Brazier JE, Akhras KS. Comparison of the HUI3 with the SF-36 preference based SF-6D in a clinical trial setting. *Value Health.* 2004;7:602-609.

21. Brazier J, Roberts J, Tsuchiya A, et al. A comparison of the EQ-5D and SF-6D across seven patient groups. *Health Econ.* 2004;13:873-884.

22. Longworth L, Bryan S. An empirical comparison of EQ-5D and SF-6D in liver transplant patients. *Health Econ.* 2003;12:1061-1067.

23. McDowell I, Newell C. *Measuring Health: A Guide to Rating Scales and Questionnaires.* New York: Oxford University Press; 1996.

24. World Health Organization. *Constitution of the World Health Organization. Basic documents.* Geneva:World Health Organization; 1948.

25. Schwartz CE, Sprangers MA. Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. *Soc Sci Med.* 1999;48:1531-1548.

26. Brooks R. Euroqol: the current state of play. *Health Policy.* 1996;37:53-72.

27. Brazier J, Fukuhara S, Ikeda S, et al. *The Japanese Valuation of the SF-6D and Comparison to the UK Values.* HEDS DP 01/05. Sheffield, UK: University of Sheffield; 2005.

28. Barton GR, Bankart J, Davis AC, et al. Comparing utility scores before and after hearing-aid provision: results according to the EQ-5D, HUI3 and SF-6D. *Appl Health Econ Health Policy.* 2004;3:103-105.

29. Kobelt G, Kirchberger I, Malone-Lee J. Review. Quality-of-life aspects of the overactive bladder and the effect of treatment with tolterodine. *BJU Int.* 1999;83:583-590.

30. Guyatt G. Commentary on Jack Dowie, "Decision validity should determine whether a generic or condition-specific HRQOL measure is used in health care decisions". *Health Econ.* 2002;11:9-12.

31. Brazier J, Deverill M. A checklist for judging preference-based measures of health related quality of life: learning from psychometrics. *Health Econ.* 1999;8:41-51.

32. Torrance GW. Measurement of health state utilities for economic appraisal. *J Health Econ.* 1986;5:1-30.

33. Froberg DG, Kane RL. Methodology for measuring health-state preferences—II: Scaling methods. *J Clin Epidemiol.* 1989;42:459-471.

34. Richardson J. Cost-utility analysis: what should be measured? *Soc Sci Med.* 1994;39:7-21.

35. Dolan P, Gudex C, Kind P, et al. Valuing health states: a comparison of methods. *J Health Econ.* 1996;15:209-231.

36. Green C, Brazier J, Deverill M. Valuing health-related quality of life. A review of health state valuation techniques. *Pharmacoeconomics.* 2000;17:151-165.

37. Robinson A, Loomes G, Jones-Lee M. Visual analog scales, standard gambles, and relative risk aversion. *Med Decis Making.* 2001;21:17-27.

38. Bleichrodt H, Johannesson M. An experimental test of a theoretical foundation for rating-scale valuations. *Med Decis Making.* 1997;17:208-216.

39. Nord E. The validity of a visual analogue scale in determining social utility weights for health states. *Int J Health Plann Manage.* 1991;6:234-242.

40. Robinson A, Dolan P, Williams A. Valuing health status using VAS and TTO: what lies behind the numbers? *Soc Sci Med.* 1997;45:1289-1297.

41. von Neumann J, Morgenstern O. *Theory of Games and Economic Behavior.* Princeton, NJ: Princeton University Press;1944.

42. Camerer C. Individual decision-making. In: Kagel J, Roth A, eds. *Handbook of Experimental Economics.* Princeton, NJ: Princeton University Press;1995.

43. Schoemaker PJH. The expected utility model: its variants, purposes, evidence and limitations. *J Econ Lit.* 1982;20:529-563.

44. Sutherland HJ, Llewellyn-Thomas H, Boyd NF, et al. Attitudes toward quality of survival. The concept of "maximal endurable time". *Med Decis Making.* 1982;2:299-309.

45. Dolan P, Gudex C. Time preference, duration and health state valuations. *Health Econ.* 1995;4:289-299.

46. Bleichrodt H. A new explanation for the difference between time trade-off utilities and standard gamble utilities. *Health Econ.* 2002;11:447-456

47. Oliver A. The internal consistency of the standard gamble: tests after adjusting for prospect theory. *J Health Econ.* 2003;22:659-674.

48. Dyer JS, Sarin RK. Relative risk aversion. *Manage Sci.* 1982;28:875-886.

49. Dolan P, Sutton M. Mapping visual analogue scale health state valuations onto standard gamble and time trade-off values. *Soc Sci Med.* 1997;44:1519-1530.

50. Stevens K, McCabe C, Brazier J. Mapping between Visual Analogue Scale and Standard gamble data: Results from the UK study using the Health Utilities Index II Framework. Health Economics Study Group. Leeds, UK; January 2003.

51. Salomon JA. Reconsidering the use of rankings in the valuation of health states: a model for estimating cardinal values from ordinal data. *Popul Health Metr.* 2003;1:12.

52. McCabe C, Brazier J, Gilks P, et al. Estimating population cardinal health state valuation models from individual ordinal (rank) health state preference data. *Sheffield Health Economics Group Discussion Paper* 04/02;2004.

53. Llewellyn-Thomas H, Sutherland HJ, Tibshirani R, et al. The measurement of patients' values in medicine. *Med Decis Making.* 1982;2:449-462.

54. Furlong W, Feeny D, Torrance GW, et al. Guide to design and development of health state utility instrumentation. Hamilton, Ontario: McMaster University Centre for Health Economics and Policy Analysis. Working Paper Series 1990;90-9.

55. Brazier JE, Dolan P. *Evidence of Preference Construction in a Comparison of SG Methods.* Sheffield, UK: Health Economics and Decision Science Department, University of Sheffield; 2004.

56. Lloyd AJ, Hutton J. Do decision making heuristics distort efforts to elicit preferences? Paper presented at: Developing Economic Evaluation Methods workshop; 2002; York, UK.

57. Shiell A, Seymour J, Hawe P, et al. Are preferences over health states complete? *Health Econ.* 2000;9:47-55.

58. Patrick DL, Starks HE, Cain KC, et al. Measuring preferences for health states worse than death. *Med Decis Making.* 1994;14:9-18.

59. Sackett DL, Torrance GW. The utility of different health states as perceived by the general public. *J Chron Dis.* 1978;31:697-704.

60. Boyd NF, Sutherland HJ, Heasman ZK, et al. Whose utilities for decision analysis? *Med Decis Making.* 1990;10:58-67.

61. Hurst NP, Jobanputra P, Hunter M, et al. Validity of Euroqol—a generic health status instrument—in patients with rheumatoid arthritis. Economic and Health Outcomes Research Group. *Br J Rheumatol.* 1994;33:655-662.

62. Lenert LA, Treadwell JR, Schwartz CE. Associations between health status and utilities: implications for policy. *Med Care.* 1999;37:479-489.

63. Buckingham K. A note on HYE (healthy years equivalent). *J Health Econ.* 1993;12:301-309.

64. Menzel P, Dolan O, Richardson J, et al. The role of adaptation to disability and disease in health state valuation: a preliminary normative analysis. *Soc Sci Med.* 2002;55:2149-2158.

65. Badia X, Roset M, Herdman M, et al. A comparison of United Kingdom and Spanish general population time trade-off values for EQ-5D health states. *Med Decis Making.* 2001;21:7-16.

66. Dolan P, Roberts J To what extent can we explain time trade-off values from other information about respondents? *Soc Sci Med.* 2002;54:919-929.

67. Kharroubi SA, O'Hagan A, Brazier JE. Estimating utilities from individual health preference data: a nonparametric Bayesian approach. *Appl Stat.* In press.

68. Currim IS, Sarin RK. A comparative evaluation of multiattribute consumer preference models. *Manage Sci.* 1984;30:543-561.

69. McCabe C. *Estimating Preference Weights for a Paediatric Health State Classification (HUI-2) and a Comparison of Methods.* [dissertation]. Sheffield, UK: University of Sheffield; 2003.

Contact Information:

John Brazier, Ph.D., M.Sc., Professor of Health Economics, University of Sheffield, j.e.brazier@sheffield.ac.uk.

Commentary

Article: Current State of the Art in Preference-Based Measures of Health and Avenues for Further Research, By John Brazier, Ph.D., M.Sc.

Pennifer Erickson, Ph.D.

In his article, "Current State of the Art in Preference-Based Measures of Health and Avenues for Further Research," John Brazier Ph.D. has provided an excellent review of methods of designing preference-based measures. In this commentary, I expand on several important ideas he raised concerning the development and application of these measures.

To translate the technology of preference measurement into application, it is important to remember that preference-based summary measures of health were developed to inform policies along a continuum ranging from macro- to micro-level decisions. In a 2001 report on measuring health, the Institute of Medicine identified three research approaches that provide data across this continuum, namely, population studies, health services research, and clinical/biomedical investigations. While each of these approaches is distinguished by features that set it apart, such as purpose of measurement, statistical design, and target population, taken together they provide a hierarchy of health information. That is, data collected using one approach may be used not only for decision making within a level but also for understanding treatment and policy impacts between levels. For example, clinical trial efficacy data may be combined with administrative and survey data for understanding treatment effectiveness and developing marketing strategies. Thus, including a preference-based measure, ideally the same one, in studies in each of the three research approaches is essential for making the best use of the unique information that these measures provide for informing macro- to micro-level decisions.

Preference-based measures differ from other health indicators in that they incorporate two types of stakeholder information: 1) individual health status as reflected by the multi-dimensional health states; and 2) community-based preferences for these states. Thus, through the use of these

measures, both individuals and society have direct input into the decision-making process, which in turn increases the validity of the resultant policies. Further, policy makers benefit from having a single measure that summarizes both health states and societal trade-offs between alternative health states.

Currently, preference-based measures rely on a single set of weights that aim to be representative of a population as a whole. These societal weights provide a common metric for studying relationships between health and its determinants, including access to care and behavioral and environmental factors, and for using the findings to guide the development and implementation of treatment and prevention policies. Also, to the extent that the preference weights are constructed to be consistent with economic theory, they can be used for allocating resources across all members of a given community, whether at the federal, state, or local level. Thus, research on methods for preference elicitation in diverse settings, including sample surveys, or development of new weights needs to result in a single set of preferences that is relevant for all stakeholders, e.g., patients and non-patients, young and old.

Multiple sets of preferences, however, may be necessary if we are to fully understand health disparities within a population. For example, a major goal of Healthy People 2010 is to reduce inequalities in health observed in groups such as those defined by gender, race/ethnicity, income and education, and disability. Preferences for health, like many other factors, may differ by sociodemographic characteristics. Although a few studies support the use of a single set of preferences, more research is needed to examine the adequacy of a single set of weights. If sociodemographic groups are found to have distinctive preferences, then additional research will be needed to understand how this information might be used to reduce inequalities in health without compromising either efficiency or equity.

Concerns about the need for preferences that are relevant to various segments of the population raise similar issues about the meaningfulness of the domains used to define the health states. Most currently available preference-based measures define health states in terms of basic physical, mental, and social functioning. Yet, these may inadequately represent the health of some. For example, function-based measures are likely to underestimate the health status of

people with disabilities since they may have relatively little limitation in non-physical aspects of health. Findings associated with the development of the Health and Activity Limitation Index (HALex), a preference-based measure of health that includes self-rated health, found a meaningful number of persons who were dependent in activities of daily living rating themselves in good to excellent health. Thus, for disabled people, inclusion of self-rated health resulted in a more representative measure of health than one based solely on function. These analyses suggest that self-rated health deserves serious consideration as a domain in any summary measure of health. Additional research to evaluate the utility of this concept as a domain can be done using the HALex and data from the National Health Interview Survey (NHIS). Focus groups and cognitive interviewing might be used to further understand the contribution of self-reported health relative to other domains, for example, physical, mental, and social function.

Lastly, use of self-report instruments for collecting stakeholder positions about both health states and preferences raises concerns about the stability of the information over time. A repeated, cross-sectional analysis of Years of Healthy Life (YHL), a measure that includes life expectancy adjusted by HALex scores, over the interval from 1984 to 1994, indicated that after 6 years of steady increase, YHL were found to decline from 1992 to 1994. This decline was found to be due, in part, to the economic recession that occurred in the early 1990s, suggesting that respondent-reported health is influenced by factors outside of the health care system, such as income and unemployment. This finding indicates that to the extent that we want to use data from various studies, either normatively to aid in interpretation of findings or analytically to understand health outcomes from alternative interventions, the data need to be from the same point in time. For example, comparing mean scores collected in 1992 with normative scores from 1990 will exaggerate the difference due to the overall decline in health attributable, in part, to the general economic downturn that occurred within the interval. Thus, if a preference-based measure of health is to be widely used in health services research and clinical/biomedical research as well as population studies to guide decision making, then the data will need to be collected continually. The NHIS is uniquely suited for this, since it is an ongoing survey that provides timely and normative data that can be used for interpreting findings across a wide range of applications.

As outlined here, preference-based measures of health have much to offer in terms of providing unique information for decision making. To realize their full potential, however, additional research is needed. The issues are challenging, and now is the time to begin to meet these challenges.

Contact Information:

Pennifer Erickson, Ph.D., Pennsylvania State University, pae6@psu.edu.

On the Policy Implications of Summary Measures of Health Status

Michael C. Wolfson, Ph.D., B.Sc.

Introduction

The most fundamental challenge in the health sector of most wealthy societies is achieving a reorientation in emphasis. At present, the dominant focus is on the operation of the health care sector—how much it costs and who has access to what kinds of care. Far less attention is paid to the overall health of the population and to the health outcomes of the myriad interventions provided by the health care system.

The evidence for this claim is simple—in the United States, for example, where insurance coverage is the principal marker of access to services, it is relatively straightforward to get data on who is covered; on surgical procedures performed, such as heart bypass or hip replacement; and on the total costs of the system. But comprehensive indicators of whether the population is getting healthier (not just living longer) and the incremental effects of heart procedures and virtually all other interventions on overall population health status are simply nonexistent. The regular production of summary measures of population health (SMPHs), as part of a coherent underlying statistical system, is essential to remedying this imbalance—to shift from a preoccupation with inputs, throughputs, and proximate outputs to the ultimate bottom line, people's health.

The idea of SMPHs has become much more widely known since they were discussed in the *World Development Report*[1] and the *World Health Report 2000*.[2] At the same time, these publications have generated substantial controversy, which in turn has tainted this important idea.

In this paper, I will first give an overview of the construction of such measures, pointing out the key areas where controversy has emerged. This rather lengthy overview is needed to provide an appropriate basis for the subsequent discussion. Second, I review several examples of the

application of such measures, with a particular focus on their roles in health policy. I conclude with a few comments on the implications for statistical development, health research, and exchange of knowledge.

Measuring Health Status—from Questionnaires to an SMPH

Examples of SMPHs include DALYs (disability-adjusted life years), which were estimated in the *World Development Report*[1]; DFLE (disability-free life expectancy), which dates back at least to Wilkins and Adams[3]; and HALE (health-adjusted life expectancy), as proposed and estimated, for example, by Wolfson.[4, 5] More generally, estimates for a summary measure that integrates both length of life and health status during life go back at least to Sullivan.[6]

Much of the discussion of SMPHs is unfortunately marred by a preoccupation with the characteristics of one or another specific measure, such as DALYs. It is useful to describe SMPHs at two levels. The first is the main kinds of elements or components required for their construction; the second is the specific choices to be made for each component. In this section, I focus on the former, and hence I provide an overview of the structure of SMPHs.[††]

The usual key components of an SMPH are:

- An individual-level, standardized, generic, short, and "questionnaire ready" description of health status.
- An individual-level numerical index of health status, based on the generic descriptive system.
- A method for aggregating individual-level numerical indices over everyone in a population (or a representative sample thereof) to construct an overall index.

[††] It is important to distinguish two broad groups of SMPHs, so-called "gap" measures, which include DALYs, and "expectancy" measures, essentially the HALE family.[7] While the two groups of SMPHs share components, specifically the first two listed in the following paragraphs, I focus here mainly on the HALE group.

In addition, it has become obvious from some of the concerns raised about DALYs, for example, that two other features are necessary:

- A "drill down" capacity to disaggregate overall results, for example to describe subgroups.
- A "what if" capacity to support quantitative estimates of health outcomes resulting from specific interventions.

I shall discuss each of these in turn.

Individual-Level Health Domains—*Questionnaire*

There is good evidence that health, as conceived by the general public, is multidimensional.[8, 9] With the rise of clinical medicine, the dominant mode of description by health professionals is in terms of diseases, for example as classified by the World Health Organization's (WHO's) International Classification of Diseases (ICD).[10] In contrast, when asked in open-ended questions what they think of when considering their health, people often respond in terms like pain, energy, mood, and the ability to get around (mobility). None of these terms has a unique one-to-one relationship to ICD-defined diseases; many diseases, for example, can cause pain and limit mobility.

As a result, there has been increasing appreciation of the need for non–disease-based approaches to conceptualizing and classifying health. One such approach is the International Classification of Functioning, Disability, and Health (ICF).[11, 12] The ICF and the ICIDH from which it evolved (International Classification of Impairments, Disabilities, and Handicaps[13]) provide very useful concepts. They focus on and distinguish among impairments, limitations on activity, and restrictions on participation. The ICF also makes a useful distinction between limitations or restrictions that derive principally from the intrinsic characteristics of the individual person and the extrinsic circumstances of the physical and social environment (e.g., barriers to mobility, social stigma) within which the person lives. The ICF is far too long and detailed, however, to

serve as the basis for a short-form series of questions that would constitute a generic health status instrument for widespread application.

Several such instruments have evolved that are in fairly wide use, including the SF-36 ("short form" with 36 questions[14]), the HUI (Health Utilities Index[15, 16]), and the EQ-5D.[17, 18] Subsequently, WHO[19, 20] selected six domains as its focus for a short-form questionnaire, the World Health Survey (WHS). Although WHO claims that this choice is consistent with and is based on the ICF, it is not clear in detail how, either statistically or methodologically, the connection was made. Nor is there an obvious crosswalk between the six questions in the WHS and the domains and chapters in the ICF.

Table 1, in its first two columns, shows in highly summarized form the results of two more open-ended analyses that sought to determine what domains are most often used by the general public to describe health status.[8, 9] In the next four columns, the table gives an overview of the domains covered by the most widely used individual-level descriptive systems for generic health status. The last column shows the descriptive system currently under development at Statistics Canada. This system was motivated by concerns with all of the extant systems, including the newly created WHO system adopted for the WHS. In the case of the WHS questions, these concerns include the absence of a clear methodology underlying the choice of specific domains,[‡‡] a lack of published results from testing the new WHO questions, and a lack of analysis of the questions' appropriateness as the basis for expressing preferences for health states or assigning weights to those states (see below).

[‡‡] Sadana and colleagues [20] describes the derivation of a longer list of 18 domains but not the selection of the subset of the 6 specific domains used in the WHS for valuing health states.

Table 1: Domains of Health and Functioning

Open-Ended / Factor Analysis (descending order)		**Specific Questionnaires** (ordered to highlight parallels)				
van Dalen	Bernier	SF 36	HUI 3	EuroQol	WHO WHS	Stats Can New
Function (38%)	Functional limitations	Physical function**	Ambulation	Mobility	Mobility	Physical Function
Not Ill (33%)	Chronic conditions	Role limits—physical		Self-care	Self-care	
Psychological well-being (32%)	Psychological well-being	Mental Health	Emotion**		Affect	Emotion
Energy / vitality (26%)	Sensory impairment	Role limits—emotional		Anxiety**/ depression		Anxiety*
Reserve (23%)	Disability days	Pain	Pain/ discomfort**	Pain/ discomfort**	Pain	Pain/ discomfort
Behaviour (7%)	Depression	Social function**		Usual activities	Usual activities	Social relationships
	Stress	General health				
		Energy**				Fatigue
			Memory and thinking**	Cognition (Dutch ver.)	Cognition	Memory and thinking
			Vision**			Vision*
			Hearing**			Hearing*
			Speech**			Speech*
			Use of hands and fingers**			Use of hands and fingers*

* Secondary domain in new Statistics Canada system.

** Main source (at least in general terms) for domain in new Statistics Canada system.

The Statistics Canada system under development is based on (a) a review of the literature, especially the results of open-ended questions asked of the general public to elicit their thinking on what constitutes health and hence what should constitute the basic domains[8, 9]; (b) a factor analysis of a range of existing measures[21] as well as their component domains applied in "head-to-head" comparison surveys by Statistics Canada[22]; and (c) a carefully structured series of focus group discussions conducted by Statistics Canada.[23]

I should note that the original DALYs[1, 2] made no use of generic health state descriptions. Instead, these measures were based first on more than 100 ICD-based diseases and then on a

mapping from each disease to a univariate "disability" level between zero and one.[24] This latter set of mappings was in turn based on a series of small expert group panels. This approach had the major advantages that (a) it could build on the very wide range of pre-existing data and clinical knowledge available that was classified in terms of the ICD and (b) the resulting DALY estimates could be disaggregated in order to attribute the global burden of disease to specific "causes"—necessarily defined in terms of clinical disease. On the other hand, this disease-based approach suffers from several serious limitations:

- The original mapping from disease to a number between zero and one was not based on any preferences or weights elicited in a systematic manner from a representative sample of the population.[§§]
- No practical method exists to elicit the population's preferences in terms of diseases—most of the general population is not familiar with diseases as clinically defined, and there are too many kinds and manifestations of diseases for the task to be practical.
- No account was taken of comorbidities, and doing so would be difficult.
- The general population does not think of its health solely, or even predominantly, in terms of ICD-defined diseases, as shown in the first two columns of Table 1 above.

Of course, we still need to be able to connect health status, measured in some generic manner, to conventionally defined diseases. But approaches other than that taken by the original DALYs are preferable. In the first instance, the Statistics Canada efforts are relying on consensus panels of medical experts to define mappings from clinical diseases into the generic descriptions of health status. In the future, by putting the generic health status questions onto large surveys of population health and linking the surveys (subject to respondents' consent) to administrative data on health care encounters that include ICD codes, we will be able to underpin these initial mappings with rich empirical evidence (essentially a "Rosetta stone" analysis).[***]

[§§] Subsequent mappings were based on panels of knowledgeable persons in several countries. Since then, WHO's WHS provides a population basis, not linked to ICD diseases but rather to the six health states in the second-to-last column of Table 1. No such analyses have yet been published, however.

[***] For some very serious diseases, such as dementia and terminal cancer, survey responses will be unlikely, and thus some form of proxy mapping, for example by clinicians, will remain essential.

In addition to the factors discussed so far, several other criteria should guide the selection of domains for the basic descriptive system and the construction of the specific questions to capture health status within each of the domains. One, already implicit in the discussion above, is that the domains should *span* the most important kinds of health problems experienced by people in the society. Another is that there should not be too many domains; the domains should be *parsimonious*, not only in number, but also in the descriptions of the levels of functioning or health status within each domain (ranging from very modest to very severe health problems). One reason is to keep the questionnaire short, so that it can be widely adopted without large costs in surveying or undue burdens on respondents. Another, more technical reason is that the methods for eliciting preferences among health states simply will not work if these multidimensional health states (one dimension for each domain) are too complex.

A third criterion is to focus on functional limitations in the sense of *disabilities* in the older ICIDH lexicon.[†††] Clearly, this omits handicaps in the performance of social roles (i.e., limitations in social roles that result from disabilities only because of the external social or physical environment) from the core concept, but it seems better to consider such handicaps as sequelae of disabilities (to continue with the older ICIDH terms). This omission, as an ultimate outcome, is troubling. But the focus on functional limitations ("within the skin" health characteristics) offers several advantages—it is more amenable to validation; it is less likely to pose problems of translation across languages and cultures; it may prove a less complex task from the viewpoint of achieving the consensus needed for widespread adoption; and it is more likely to allow structurally independent dimensions of health status in the individual-level descriptive system, a technical benefit for the derivation of the preference or weighting function (see below).[‡‡‡]

Relatedly, there is no point in wasting scarce questionnaire time on domains unless they are generally *uncorrelated*. Similarly, there are opportunity costs to devoting time to domains that

---

[†††] We would say "activity limitations" in the context of the new ICF, as opposed to "participation restrictions," were there not still some debate about how these two broad concepts should be defined.

[‡‡‡] Nevertheless, the way people value or express preferences about various health states, even if these states are constructed as much as possible to be "within the skin," will inevitably reflect to some degree the social milieu within which they live.

are necessarily or structurally linked, such as physical mobility and activities of daily living. An impairment of mobility will generally restrict a range of activities of daily living, so that the responses will tend to be correlated. Of course, a lack of such correlation, in itself, would be useful information, because it would reveal that a given "disability" (in the original ICIDH sense) does not always express itself in the same handicaps. But for reasons mentioned in the following section, domains like these—mobility and activities of daily living—are not *structurally independent,* and thus both should not be used.

More difficult challenges in conceptualizing health domains relate to the distinction between capacity and performance ("can do" or "is able to do" versus "does do"). In principle, it is preferable to assess capacity rather than performance. The simple reason is that reductions in actual performance may arise from either individual factors, especially intrinsic limitations in people's capacity but also possibly from lack of motivation, on the one hand or from external factors such as environmental barriers on the other. As a result, measures of performance are at greater risk of ambiguity as to the source of any restrictions and hence not as useful in providing measures of the health status of individual people. Of course, it would be most valuable if the health statistics system could also provide information on the extent to which performance, or "participation restrictions" in ICF parlance, is attributable to aspects of the environment. But this must start with measures of individual health that are essentially intrinsic, or "within the skin."

There is a broader aspect of the social and physical environment that poses problems. In its most general form, this is the issue of cross-cultural comparability. As explored in Sadana et al, [25] there is good evidence that responses to such basic questions as the reporting of pain vary systematically across countries and across cultural subpopulations within countries. These problems likely transcend the difficulties of translating the questions across languages. They may also be related to cultural norms or deeply rooted perspectives on life. This remains an open area where further research is badly needed.

Aggregating over Health Domains at the Individual Level—Preferences

Given a parsimonious set of health state domains, incarnated as a health status descriptive system by means of a set of well-tested survey questions, the next step in constructing an SMPH is a method for mapping any set of responses into a number between zero and one, where one represents fully healthy, and zero represents dead.[§§§]

There is general consensus that these preferences for health states are best based on the views of a representative sample of the population, using carefully designed and sensitive procedures. Within this ambit, however, there is a range of possibilities. One concerns the approach used to elicit the preferences. There are four main approaches: visual analogue scale (VAS), time trade-off (TTO), standard gamble (SG), and person trade-off (PTO). As these methods are well described elsewhere, I do not go into detail here.

There is considerable evidence that the kind of method used materially affects the preference function estimated.[26] For example, PTO tends to count very mild health problems as much smaller decrements (i.e., closer to one) than TTO or SG.

In the latest Statistics Canada project,[23] all of these approaches were assessed through a series of focus groups for their cognitive difficulty and acceptability in the general Canadian population. My conclusion is that Canadians feel most comfortable with the SG approach—notwithstanding the facilitator having explained that the resulting preference weights could be used to inform social decisions on resource allocation for the general population and not just the person responding. Operationally, I found that the SG was best preceded by a VAS exercise to acquaint respondents with the multidimensional descriptions of health states and the basic idea of ranking these states.

In addition to the kind of approach for measuring preferences, there is the question of the best mode for eliciting preferences. The initial approach was to draw a sample of hundreds or

[§§§] In studies that have included extremely poor health states, many people will assign a value less than zero; in other words, they consider that health state to be worse than death.

thousands of people and ask each individually a series of questions (framed using either SG or TTO). The results from the sample were then collected and analyzed statistically to produce an overall average or representative preference function.[16, 27, 28] Murray[24] and the Dutch, [29] in contrast, have used much smaller populations (e.g., a dozen) in a group setting. In the WHO case,[24] the persons in the group discussed the questions together and were then led to a consensus response to each question. There was also a real-time check on the internal consistency of the responses.

This latter approach has the major advantage of giving respondents some opportunities to learn more about what is, in the end, a rather challenging exercise. (The response rates on the population surveys just noted for the HUI in Canada and the EuroQol in Britain were on the order of 50%.) On the other hand, the groups were too small to be representative of their populations, and forcing the group toward consensus obscures what may be very real heterogeneities in preferences among members of the group.

Statistics Canada has addressed these issues by doing the following:

- Using a relatively large number of full-day focus groups with participants from heterogeneous backgrounds, and conducting the groups across the country.
- Using a group setting with a facilitator to offer respondents an opportunity to learn about the motivation for collecting their preferences for health states, and to discuss with others what is meant by the health states themselves and the process for ranking them.
- Allowing each person within the group to record her/his own preferences, without any obligation to share them with others in the group or to form a group consensus.

Practically speaking, even with a small number of health status domains (e.g., five to eight in Table 1) and only a few levels of functioning within each domain (e.g., three to five), there are hundreds to tens of thousands of possible combinations. For example, the eight domains in the HUI, each with five or six levels, result in almost 1 million possible combinations. As a result, virtually all protocols for eliciting preferences have relied on approaches in which everyone evaluates a common handful of composite health states (a specific profile or vector of responses to each of the questions). This common set of health states often includes all the “corner states”

(i.e., health states where the profile of responses is at the top level in terms of healthiness for all but one of the domains). Typically, everyone also evaluates a sample of other health states, although the specific health states generally vary across respondents. Finally, the result is values for only dozens of the hundreds to thousands of all possible health states within the descriptive system, and thus values for all the (large majority of) remaining possible health states are imputed by means of some mathematical formula. This formula is best estimated statistically. This process of selecting a small set of health states from the set of all possible health states to be directly valued and then extrapolating to the rest generally presupposes a mathematical form of the preference function. In turn, this form typically assumes independence between any pair of health domains. This is the main reason that the criterion of "structural independence" mentioned above is important although often neglected.[****]

Even within the class of mathematical functions that assume independence among health domains, a variety of specific functional forms are possible. Developers of the EQ-5D used a simple linear regression to estimate its valuation formula (which they call a "tariff"[28]), but they found that for some of the worst states of health, a simple linear formula did not fit the data. As a result, they added (arbitrarily) an extra interaction term for some of the worst health states. The developers of the HUI have tried a variety of forms, including both additive and multiplicative. The data underlying the latest HUI3 were carefully designed to allow an empirical assessment of the most appropriate form, and this turned out to be multiplicative.[16]

In empirical work to date, there is a suggestion that preferences for health states may vary systematically across subgroups (e.g., rich and poor) and it is well-known that they vary with personal or close experience of disease or disability. This is a matter warranting further research.

Still, such variation is analogous to exploring the robustness of price indices across various population sub-groups like the poor or elderly. Just as virtually all statistical offices worldwide

---

[****] This criterion has been an explicit part of the design in only the HUI3 and the Statistics Canada descriptive systems. The other three systems in Table 1 do not meet this criterion. Structural independence is also implicit if corner states are used as part of the process for eliciting weights, because respondents will find it unreasonable to value an obviously impossible state, such as being completely impaired in terms of mobility and yet having no limitations at all in usual activities.

use an "average" expenditure basket for constructing consumer price indices and then measuring inflation, even though there are important differences in expenditure patterns across population subgroups, it should be satisfactory to use an "average" preference function, at least as a starting point.

It is also of practical importance to realize that progress on consensus for an individual-level system for describing health status (the previous section) can be *decoupled* from that for an individual-level preference function (this section). Similarly, the individual-level descriptive system and preference function are logically distinct from, and their development need not be coupled with, that of the formula for aggregating individual people's health status to form a population-level index (the following section).

Aggregating Over Individual People—the Population Index

The third component of any SMPH is a formula for aggregating health status across a set of people representing a population. It is only after this step that we would have the basis to conclude for a country's population, say, that not only are we "adding years to life" but also "adding life to years" (to paraphrase the Rochon Commission report in Quebec, 1987[30]), or in Fries'[31] sense, we are witnessing a "compression of morbidity."

Assuming we already have a representative population health survey that included the standard questions for the generic system for describing health status, and we already had estimated a preference function, then each person in the sample could be assigned an index between zero and one, representing that person's individual-level index of health status. Several approaches are then available to construct an overall index.

One is simple cross-tabulation, for example to generate the average level of health status of the population. Comparisons of this statistic over time, or across subpopulations, however, could be substantially affected by other factors, such as differing proportions of elderly. In this case, the usual approach is some sort of age standardization. Mechanically, this corresponds to re-

weighting the descriptive data on individual-level health status before cross-tabulating to represent some "standard" or reference population in terms of its distribution by age groups.

Alternatively, the averaging can build on the reference population coming from a life table, based on contemporaneous mortality rates. Indeed, if instead of dividing by the number of person-years in the life table, one uses the radix, we have the Sullivan method and what is widely referred to as HALE.[32] This is the most commonly used approach for the last step in constructing SMPHs.

More sophisticated approaches are possible. In essence, they involve creating a representative sample of individual-level health trajectories using either multistate life tables or microsimulation. Indeed, these extensions of the Sullivan method are essential for the kinds of "what if" capacities for SMPHs described below.

Ethical Concerns

Numerous ethical concerns have been raised with regard to SMPHs—particularly with regard to distributional and equity matters.[33, 34] These ethical concerns are in turn motivated by an expectation that SMPHs (plus their associated statistical infrastructure) will have as their primary use in policy the informing of decisions on allocating resources within the health sector. I return to this point in a later section.

To some extent, these concerns reflect a failure to carefully distinguish steps 2 and 3 above in the construction of SMPHs: aggregating the questionnaire responses over health domains to assign any given *person* a summary health status score (typically between zero and one), and the subsequent aggregation of these individual-level scores to represent a summary *population* index (typically measured in years in a form of adjusted life expectancy), the SMPH.

For example, Daniels[34] asks, "How much priority should we give to the sickest or worse off patients? When should we allow modest benefits to many people to outweigh significant benefits to fewer? When should we allocate resources to produce "best outcomes" and when should we

give people fair chances at some benefit?" His questions are motivated by a concern about the blind application of SMPHs in decisions on allocating resources, where some patients will get help while others will do without.

Daniels' questions go to the character of the (implicit) ethical judgments regarding distributional equity embodied in Sullivan-style SMPHs. Put simply, the HALE summary measure gives "one person-year one vote." An increment of 0.1 (let us say) in individual health status for 1 year increases HALE by exactly the same amount no matter who the person is—age; sex; income; and current, previous, and projected health status all make no difference. This may be a problem if a matter of indifference in the calculation of HALE[††††] would not be a matter of indifference to the people concerned (e.g., whether they or someone else received the treatment) or to certain population subgroups (e.g., the disabled).

Concerns about the indifference implicit in a given HALE measure, however, are tantamount to a challenge to the individual-level preference function described earlier. For example, it is equivalent to asserting that an increase in person A's summary health score from 0.4 to 0.5 is not the same as person B's increase from 0.4 to 0.5 or person C's increase from 0.8 to 0.9. But the preference function that determines when a given health state is scored as a 0.4 or a 0.5, say, has presumably been elicited from a representative population using a reasonable series of methods. If so, those expressing such ethical concerns are saying, in effect, that they disagree with these persons' preferences as elicited and summarized into a mathematical formula. This seems a rather weak philosophical argument.

A stronger argument is that the preferences, as described above, have been elicited piecemeal from individual people (i.e., one health state at a time). But once people see the implications of their expressed preferences played out on the broader canvas of a specific SMPH-based analysis of resource allocation, they may well disagree with the final results. Methodologically, it would suggest that the processes involved in each component, as well as the way they are put together, need to be re-examined. Practically, it means simply that SMPH-based analyses should never be

[††††] "Indifference" is here being used in the same sense as indifference curves in the economics of individual behavior—that is, contours in the space of inputs to the HALE calculation where the result is a constant.

applied mechanistically—a highly unlikely event in any case. There should always be supplementary statistical information and a range of other knowledge and common sense brought to bear, as well as appropriately designed decision-making processes. All we seek here is information suitable for "evidence-considered" (as opposed to "evidence-based") decision making.

A related ethical challenge concerns population subgroups who, given their characteristics (e.g., disabled), would be treated differently if resources were allocated solely based on an SMPH. One response to this challenge is to appeal to the Rawlsian notion of the "veil of ignorance." The fundamental question is whether people would say that a given pattern of financing health interventions is fair *before* knowing what specific health problems they will encounter during their lifetimes. This seems a better criterion of the fairness of a system of social allocation of resources for health-related interventions than one that polls people *after* they know they have disease x or health problem y.

Another response is to ensure that any SMPH has an associated capacity to generate measures of the distributional impacts of health-related interventions, broken down for example by age, health problem, socioeconomic position, or geographic locale. Then, any decision making based, for example, on the changes in HALE expected as a consequence of a proposed new health intervention would be complemented by more detailed estimates of the impacts on a range of population subgroups.

Levels of Health and Inequalities in Health

The ethical concerns just discussed reinforce the general importance of complementing SMPHs with measures of the distribution of health within a population—including the extent of health inequalities. The *World Health Report 2000*[2] was a major advance insofar as it included in its basic summary measures not only the average level of population health but also the distribution. The way WHO did this,[35] however, has been controversial.

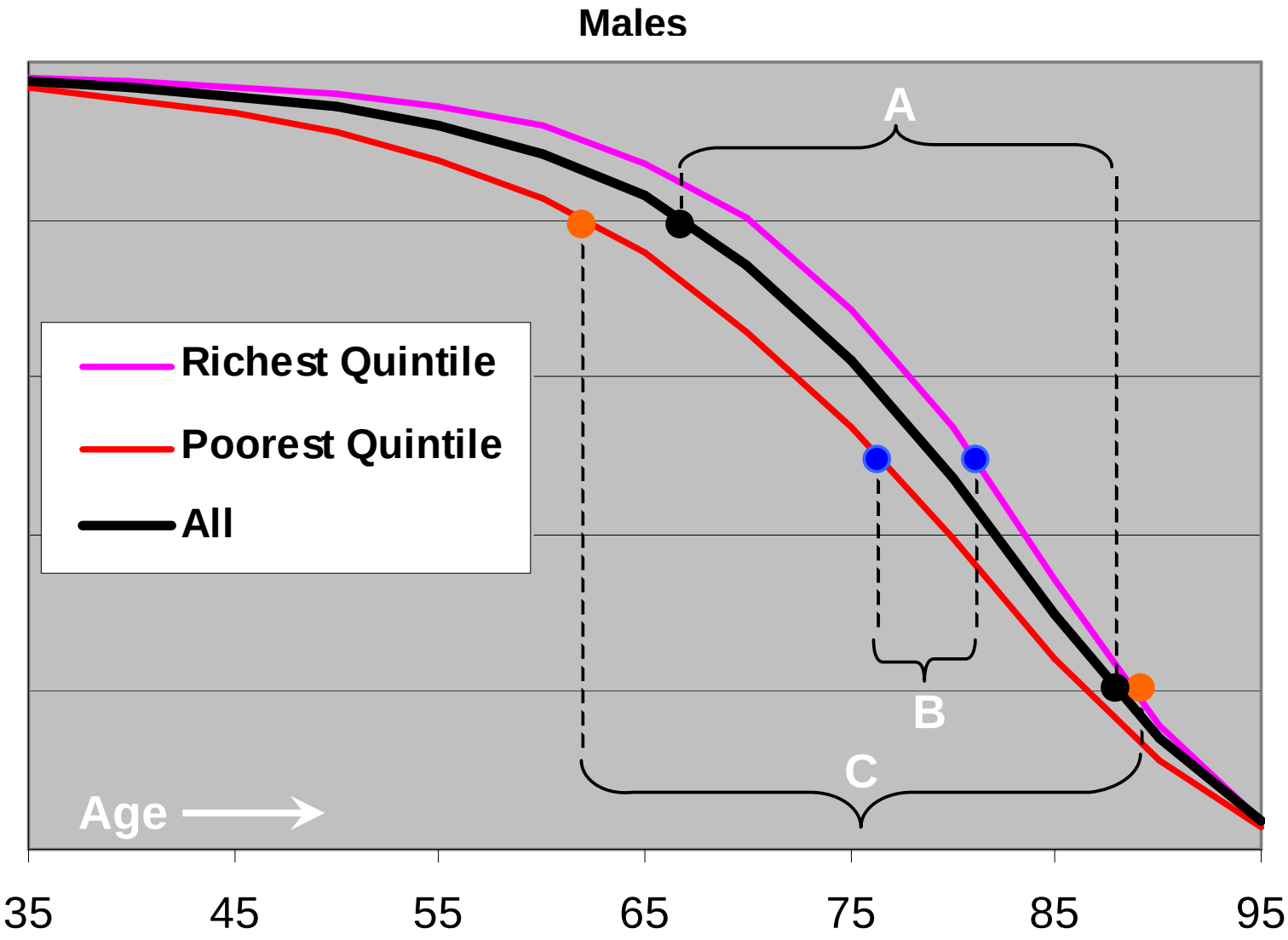
Essentially, there are two broad ways of characterizing health inequalities. The WHO approach considers health as analogous to income and looks at the extent of variations or disparities in health within a population. For example, some people die at age 40 (or have a health-adjusted life length of 38), while others die at age 80. This is much more unequal than, say, everyone living to age 75 (or having health-adjusted life lengths of 71).

Most of the literature on health inequalities, however, looks at the correlation between health status and one or another marker of socioeconomic status. In a wide range of circumstances, and for different measures of both health status and socioeconomic status, every step up the socioeconomic ladder is associated with better health, and there is good evidence that causality proceeds primarily from wealth to health.[36] Hence, measures of health inequalities in this sense are essentially based on the bivariate distribution of health and socioeconomic status, while those of WHO are basically univariate or marginal.

It is possible[37] to develop summary indicators of health inequalities in both the univariate or marginal sense, and in the bivariate sense, where the index whose inequality is being measured is analogous to SMPH but at the individual level, for example as sketched in Murray, Salomon, and Mathers.[38] Such indicators, however, will not be relevant for policy or the broader public if they are technically complex and difficult to understand.

As a result, it would be most useful for SMPHs to be accompanied by relatively simple indicators of inequality. One way to capture both the overall (WHO's univariate or marginal) and the social (i.e., bivariate) approaches in a common framework, instead of debating which is better, is shown in Figure 1. In this case, straight survival curves and life expectancies are displayed, although health-adjusted versions of these (i.e., leading to HALE) could be developed. (Note that this may be technically more difficult than it first appears.)

Figure 1 -- Inequality and Survival Curves, Canada, 1996 (Urban)



These are male survival curves for Canada in 1996, both overall (black) and for the top and bottom income quintiles (blue and red, respectively). In turn, the survival curve (turned on its side) represents the (cumulative) distribution within a cohort of people of their life lengths. One simple indicator of inequality in these life lengths is the range between the 20th and the 80th percentiles. For the overall population, this 80–20 range is shown by the distance "A," which is about 20 years. If there were no inequality in life lengths, we would have a rectangular survival curve, as everyone would die at the same age. When we divide the population into income quintiles, however, and examine the top and bottom income quintile survival curves, the 80–20 range in median life expectancy is shown by the distance "B," which is about 5 years. This is clearly a coherent framework for examining both overall and social inequalities in health. It suggests that for men in Canada, income-related health inequality accounted for about one quarter of overall health inequality.

A diagram like this[‡‡‡‡] shows how, rather simply, indicators of the average level of population health as well as the two main concepts of inequalities in health can be combined coherently in a single framework—all building on the core ingredients of the HALE family of SMPHs. Of course, the curves in Figure 1 presuppose that the SMPHs are decomposable, in this case by income group.

Disaggregating the Summary Index – a "Drill Down" Capacity

As noted both in the previous section and earlier, it is most desirable for an SMPH to be constructed in such a way that the overall index, or changes in the index over time or across groups, can be decomposed. SMPHs will be most relevant for public policy if their movements or variations can be differentiated among population subgroups, including those varying in socioeconomic position. It is also important that such movements or variations can be attributed to causes and that their likely responses to health policy interventions can be explicitly quantified.

[‡‡‡‡] Please ignore the interval shown as C.

One approach to inferring causality is to decompose overall changes into components. Some kinds of decomposition analysis, or disaggregation or "drilling down" into the underlying data, are straightforward—for example, estimating breakdowns by age, sex, geographic area, and sociodemographic group. These disaggregations rely on the fact that SMPHs can be produced for constituent subpopulations[39] and also are additive over age groups that together cover the full life cycle.[5]

It is not a given that this kind of drill-down capacity will be available, however. It depends on the way the SMPH has been estimated and what is published. WHO, for example, has not published all the underlying details of its DALY and HALE estimates (e.g., by age and disability level), and even if it had, its methods of estimation, which are often indirect or by means of imputation, do not allow other kinds of disaggregation (e.g., by socioeconomic group). In contrast (albeit with a bit of work), the Healthy Years Equivalent in the U.S. *Healthy People* series,[40, 41] the closest the United States has to an official SMPH, can be disaggregated by drawing on the underlying National Health Interview Survey and mortality data.

Influencing the Summary Index—a "What If" Capacity

The statistical system or framework just described for measuring health status at both the individual and population levels would represent a major advance in our ability to monitor population health (both levels and distributions) and to accumulate knowledge about causal factors. There are further needs, however. One is that certain kinds of breakdowns, for example, the portion of deaths or the reduction in HALE attributable to smoking tobacco cannot be derived by straightforward processes of disaggregation. Instead, the "estimation" of such a number requires the positing of a hypothetical "what if" scenario and the careful, quantitative tracing through of the main causal pathways that connect smoking to health status and mortality.

Another basic requirement is that this statistical framework be able to connect to health and health-related policy. The "bottom line" is informing decisions—whether by national or local government agencies or large private providers—on which of the myriad known health-affecting interventions merit a share of society's (or the insured's) limited resources. Similarly, the

statistical framework should also assist people's health choices, both in terms of behavior (e.g., diet and exercise) and use of various interventions (e.g., medication, screening). Thus, at the macro level, an obvious desideratum is to be able to link the SMPH back to factors amenable to policy.

In one sense, however, this is an unreasonable expectation. For example, in the economic sphere, no one seriously questions the regular production of data on unemployment rates, inflation, or income inequality, even though there is no simple "policy lever" to which each responds. It is recognized that all are influenced by a wide variety of factors, only some of them under the control of governments. At the same time, data on a large part of the range of influential factors (subject to budgetary constraints for the statistical system and limits in knowledge) are routinely collected in most countries (e.g., educational attainment, productivity, income tax liabilities). And finance and treasury ministries have for decades constructed sophisticated policy simulation models to enable policy analysts to project the likely impacts of moving one or another of the available policy levers (e.g., income tax provisions).

An analogous approach is feasible in the area of health policy, given agreement on an operationalized summary measure of population health, HALE. The main addition to the statistical framework would be embedding the HALE family of indicators in a computer simulation model. This would allow a generalization of the concept of "population attributable fractions"—for example, recent estimates that "40,000 deaths in Canada are due to smoking" could be extended to support a range of estimates like "x years of HALE are lost due to the pain and mobility impairment aspects of arthritis."

Such "cause-deleted" and "attributable fraction" analogues are effectively answers to "what if" questions. They therefore require credible statistical representation of the causal pathways that are implicated in answering the "what if" questions as well as capacity for sophisticated simulation modeling to generate plausible versions of the implied hypothetical or counterfactual scenarios. This is feasible, as demonstrated by WHO's most recent report[42] and by Statistics Canada's POpulation HEalth Model (POHEM[43, 44]).

Using Measures of Health

In this section, we show how measures of population health can be used, with a particular focus on public policy. In general, uses fall into two main groups. One is to monitor the evolution of levels and patterns of population health, including inequalities in the distribution of health. The other is to support decisions on the allocation of resources to health-affecting interventions.

Another use, which is relevant to public policy over the longer term, is to support research and indeed help define research priorities regarding the causal pathways underlying the kinds of "drill down" and "what if" analyses just described.

WHO's Global Burden of Disease

As noted earlier, the "global burden of disease" estimates published in WHO's *World Health Reports* since 2000, as well as the earlier *World Development Report* in 1993,[1] are likely the most widely known applications of health status measures to try to influence health policy. As far as the advanced economies are concerned, however, the influence has been indirect. Several countries (e.g., Australia, the Netherlands) have published more detailed country-specific estimates of the global burden of disease using essentially the same methodology as WHO. These results have presumably been noted by health ministries, although the implications for policy have most likely been more subtle—for example, in shifting the conventional wisdom about the relative importance of infectious and chronic diseases and of fatal and nonfatal chronic diseases such as arthritis and mental illness.

There are at least two main reasons for the (so far) relatively limited policy impact of WHO-style estimates of SMPH on countries' health policies. One is that it would be extraordinary for any publication of socioeconomic statistics to have an obvious and direct influence on public policy. For example, there are long-standing data series on the prevalence of children living in low-income families, and yet policy actions are often slow at best. Perhaps the strongest example of a direct link between the publication of socioeconomic data and policy is inflation rates, which do have a close link to monetary policy. But in this case, there is well-established (albeit imperfect) theory—macroeconomics, a rich and multifaceted statistical framework within which changes in

the inflation rate can be readily understood (the System of National Accounts), a decades-long tradition of policy-oriented macroeconometric simulation modeling, and a powerful agency whose specific mandate is to keep the inflation rate low (the central bank).

With WHO's analysis of the global burden of disease, many of these features are missing. The indicators are at too high a level to be useful for the actual practice of health policy; they are not connected to the kinds of policy levers available to ministries of health, let alone to local public health authorities and managers of health care establishments. Indeed, some of the discussion has instead suggested a "dashboard" approach, with many more indicators and the choice of those indicators more closely coupled to current issues of health policy in member countries.

The analysis of global burden of disease has no comprehensive underlying statistical framework akin to the System of National Accounts, although WHO is not to blame for this; the analysis had to rely on available data, and WHO did do a remarkable job with the limited material at hand. As well, there is no tradition nor capacity in the health sector for policy-oriented simulation modeling. Ministries of health, while in principle in a position analogous to that of central banks, are often more accurately characterized as ministries of sickness care or as ministries to pay health care providers—they do not operate with an almost single-minded focus on maintaining and improving population health that would parallel the devotion of central banks to maintaining price stability.

The other main reason for a weaker than expected policy impact is that WHO's analysis of the global burden of disease has been tainted by numerous controversies and problems, even though these do not concern anything inherent in the use of SMPHs in countries' health policies. For example, WHO proclaimed five fundamental goals, one of which was the overall level of population health measured by an SMPH (DALYs). But it then proceeded to aggregate these five diverse indicators, in conjunction with a crude economic production function, to generate an even higher-level overall indicator to rank the efficiency of each country's health system. This last step was highly questionable methodologically, based on completely inadequate data, too abstract for any practical policy application, and generally unnecessary.

Another controversial aspect of WHO's analysis of the global burden of disease is related to its second broad goal—inequalities in health status. Giving high prominence not only to the level of a population's health but also to its distribution is a major advance. As noted above, however, the way this indicator was conceptualized ignored the preponderance of existing work and policy focus, certainly in developed countries like Canada, the United Kingdom, United States, and Sweden, which have been concerned with *social* inequalities in health. It was also based on weak data for only a minority of countries, and yet results were extrapolated to all member countries. The net effect has been much controversy and the marginalization of its specific indicator of health inequality.

Canada's "Accountability Agenda"

In Canada, the policy role of measuring health status has been following a far less dramatic but possibly more penetrating evolution. Throughout the health system, including the highest political levels, there has been a slow but noticeable shift in focus in health policy away from complete preoccupation with the costs of inputs and counts of throughputs of the health (actually illness) care system toward measuring health outcomes. Since the late 1980s, there has been a long but rather weak history of developing health indicators. This trend received a much-needed boost with the signing in 2000 of a health accord (First Ministers' Meeting Communique on Health) by all the first ministers in the country (i.e., the Prime Minister plus all the provincial and territorial premiers[45]).

The 2000 accord was primarily about the funding of health care (the nation's universal Medicare system), a complex matter in Canada. Much of the jurisdiction for health is in the provincial domain, but for historical reasons, a substantial portion of the funding, as well as core overall guiding principles, is at the federal level.

Leading up to this accord, the federal government sought to ensure that its new spending on health care, which for constitutional reasons must largely take the form of block transfers to the provinces, actually purchased both reforms of the health care system and improvements in population health. As a result, the accord prescribed a series of indicators on which all

jurisdictions agreed to report regularly, starting in September 2002. This was a first, not only in health policy but also across the spectrum of social policy.

Subsequently, the Royal Commission on the Future of Health Care in Canada[46] and a special Senate committee[47] both issued major reports calling for substantial reforms of Canada's health care system as well as major additional injections of federal funds. Both reports argued forcefully that any new federal spending on health care be structured in a way that "would buy change." A major study of public attitudes undertaken for the Commission further determined that Canadians, beyond their well-known concerns regarding access, were concerned that monies be spent on *effective* interventions.[48] Most recently, there was a second meeting of first ministers in February 2003. The agreement coming out of this meeting, the First Ministers' Accord on Health Care Renewal,[49] not only spells out a major injection of new federal funding for various components of health care (e.g., funds earmarked for reformed primary care, additional diagnostic imaging equipment, and universal catastrophic drug insurance), but it also builds on the indicators in the initial accord of 2000.

As a result, these two first ministers' agreements give increasing prominence to the role of reporting information, especially indicators, to the general public. They signal another step in the federal government's strategy in funding health care. In the 1960s and 1970s, cash transfers to the provinces were based on a 50-50 sharing of costs for specific kinds of expenditures, and thus the provinces could essentially pay for hospitals and physicians with 50-cent dollars. In the 1970s and 1980s, the federal government shifted to block funding, abandoning the 50-50 cost sharing because it had become an open-ended obligation over which it had no spending control. In the 1990s, as part of a broad strategy of controlling deficits, the federal government began cutting back on the growth rate of these block fund transfers to the provinces. The health accords of 2000 and 2003 reflect both the success of the federal government's strategy to control deficits —the federal budget has been in surplus for 7 years at this point, although there was also a period of essentially zero real growth in health care spending during the mid-1990s—and the growing pressure to increase, at least back to the levels of the early 1990s, the level of federal funding of health care.

At the same time, with the 2000 health accord, and more so with the 2003 health accord, the injection of new federal funds is being accompanied by increased accountability based on explicit health indicators to be reported regularly by each jurisdiction to its public.[§§§§] Of course, some of these indicators are still of the old school, for example to track spending on new diagnostic imaging equipment and to track the pace at which new kinds of primary health care services are made available within each province and territory.

But the required indicators also include explicit measures of health status and the effects the health system is having in terms of "changes in life expectancy, improved quality of life, and reduced burden of disease and illness.[45]"

About a week after the 2003 health accord, the federal government brought down its budget. The most prominent parts were the authority to increase fiscal transfers to the provinces and territories for spending on health care and the agreed-upon health reforms. While these amounts are measured in tens of billions of dollars, the budget also provided tens of millions of dollars to improve the data and statistical foundations for the expanded range of reporting mandated in the accord.

As with any exercise in developing indicators, there is always a difficult trade-off between focusing on those indicators for which data are already available but which at best poorly capture the ideas or concepts that need to be tracked and the opposite, looking at better indicators for which there are little or no data. Fortunately, in this case a mixed strategy seems likely to continue to be followed—getting results out quickly where data are available while at the same time moving to put in place new kinds of data collection and analytical capacities to support indicators of greater relevance.

In the latter case, informal conversations suggest, for example, a clear sense that life expectancy is a 'tired" indicator. Canadians generally (aboriginals being the one unfortunate exception) live long lives by international standards. The population is much more interested in, and can relate

[§§§§] This reporting is to be reinforced by a new "health council" with a very high-profile mandate for national reporting. The precise roles of this new council were being worked out as this paper was being drafted.

better to, data that reflect their lived experiences (their own or those of people close to them) of chronic disease and disability. There is, as a result, a greater interest in SMPHs that combine both length of life and health-related quality of life. As a result, DFLE (disability-free life expectancy) was included in the reporting pursuant to the 2000 accord, while HALE was adopted for the second round of reporting following the 2003 health accord.

Importantly, the novel and much higher-profile reliance on indicators in the health accords (in turn flowing from the broader "A Framework to Improve the Social Union for Canadians" that set forth the basic idea[50]) represents another possible step in the evolution of a new strategy for managing Canada's health system. While not entirely explicit, this may represent an attempt to (a) shift the pathways of accountability for the health system away from typically fractious and highly political disputes among key interest groups (e.g., various providers and payers—the loudest voices) toward more neutral and technocratic grounds, and (b) move away from attempting to control the system via direct spending powers (which for constitutional reasons have become increasingly constrained at the federal level) toward using the provision of strategic kinds of information, in particular on the performance of the health system and health outcomes, communicated directly and broadly to the public. The intent with this new strategy is to exploit the power of statistical information to illuminate underlying possible approaches to improving the system and to use public interest to steer the system toward practices that are likely to result in better outcomes. If so, this means that SMPHs and the analytical tools needed to attribute changes in overall health status to specific health policies and interventions will come to play a much stronger role in health policy in Canada. In effect, the health accords may signal a fundamental shift in the use of health-related statistical indicators from a relatively passive monitoring role toward a more active role through engaging the general public.

Emerging Work on "Healthy Lifestyles"

One of the primary intended policy uses of SMPHs is estimating the relative benefits of allocating resources to alternative health interventions. These decisions occur at many levels. They may be narrowly focused, such as determining when it would be appropriate to use alternative "clot busting" drugs (streptokinase and recombinant tissue plasminogen activator

[rtPA]) following heart attacks. They may be at intermediate levels, such as deciding the relative amounts of hospital resources to devote to orthopedic versus cardiac procedures. They may be at the level of a broad disease process, such as gauging the relative importance of screening for hypertension and rehabilitation following stroke, or at a very broad level, such as deciding between screening for cancer and providing education during early childhood.

A prominent example of the use of SMPHs at the broader level is the most recent *World Health Report,*[42] which focuses on risk factors. For developed economies, the WHO report (page 87) estimates that the top five modifiable risk factors, in terms not only of their effects on life expectancy but also DALYs, are tobacco use, hypertension, alcohol, cholesterol, and overweight. Moreover, WHO goes on to assess dozens of possible interventions targeted at reducing these and other risk factors, both in terms of their beneficial effects on an SMPH (DALYs in this case) and their cost—yielding a coherent framework for exploring and comparing the gamut of interventions using a common metric, costs per DALY gained.

While this is a bold and innovative analysis, it would not be directly usable in a country like Canada. For example, the results show twice the DALY burden in the region composed of Canada and the United States from overweight as from physical inactivity. Yet, in Canada, the growing attention to the increasing prevalence of overweight and obesity looks toward *both* of their major causes—overeating and physical inactivity. Indeed, there is epidemiological evidence that physical fitness is *more* important than not being overweight[51] in terms of risk for mortality, contrary to the WHO estimates. As a result, it would be most useful for Canada, and any other country wishing to pursue the innovative WHO analysis, to conduct a range of sensitivity or exploratory analyses. These might include basic disaggregations of the results by age group as well as more complex analyses exploring alternative causal pathways among, for example, obesity, fitness, and heart disease. Unfortunately, WHO has not provided member states direct access to the underlying spreadsheet and statistical models it used to develop its estimates, which greatly limits the policy use of its efforts and illustrates, in an unfortunate direction, the importance of the "drill down" and "what if" capacities described above.

Canada is seeking to remedy these deficiencies with the development of a Population Health Impact (PHI) framework. This joint project of Statistics Canada; Health Canada, a regional health authority; and several health researchers is an extension of Statistics Canada's POHEM (see below). It is being designed explicitly not only to be able to nest and replicate the WHO kind of analysis just described, but also to be publicly available as a tool anyone with a personal computer can use (free of charge).

Evaluation of New Interventions

POHEM has already been used for the prospective evaluation of interventions. This model is needed because the vast bulk of research evidence (e.g., from clinical trials) focuses on relatively small populations, which are not always representative of the full population to which the intervention would apply. The evidence also tends, of necessity, to be piecemeal (e.g., because of the large costs of mounting more omnibus studies). As a result, before judging the net health benefit of most interventions, it is necessary to use some sort of melding process for a variety of data and analytical results, based on a computer simulation model to derive useful results. We refer to this as "meta-synthesis," which combines data from diverse studies, as compared to meta-analysis, which combines data only from studies of the same topic.

One recent example of this kind of application is the choice of whether to institute a program of screening for colorectal cancer in Canada and, if so, for what age groups and according to what protocol. This is a somewhat complex decision, not least because there are two stages to the screening (fecal occult blood testing and colonoscopy). False positives and false negatives may occur at both stages. There are no clinical trials at all in Canada of screening for colorectal cancer; and the second-stage test can have adverse effects, including death. Moreover, to form a judgment of the appropriate age ranges for screening, in addition to the expected benefits in terms of life years gained,[*****] it is necessary to bring in the costs not only of the screening tests themselves but also the various sequelae. The results of such an analysis have been incorporated by a national expert consensus panel into a new set of guidelines.[52]

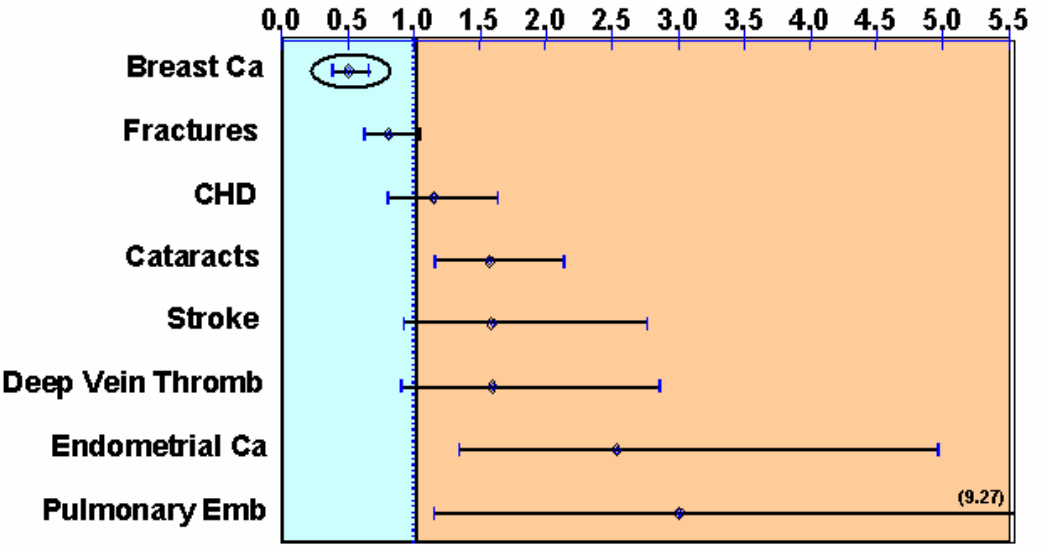
Another example is an analysis of the use of tamoxifen, a well-established secondary preventive agent for contralateral breast cancer in women who already have the disease. The new idea was to assess whether tamoxifen would also be beneficial in preventing breast cancer in women who are otherwise healthy but who, according to specified criteria, are at higher risk, because, for example, they have a close relative with breast cancer.

Figure 2[53] displays the main results of a major clinical trial designed to assess the efficacy of tamoxifen in this new primary prevention mode. The horizontal axis shows relative risks based on almost 5 years of follow-up: 1 is neutral; less than 1 means the chances of the event happening were reduced. The horizontal bars show the 95% confidence intervals for the effects of this drug used in this preventive manner.

The incidence of breast cancer, shown at the top of the figure, was cut in half, and the confidence interval was quite tight. The drug clearly has a beneficial impact on incidence of breast cancer for the population targeted in the clinical trial. For other clinical endpoints examined in the trial, however, such as the incidence of coronary heart disease, stroke, deep vein thrombosis, and

[*****] At this stage, POHEM has not yet incorporated the weights for health status needed to estimate effects in terms of health-adjusted life years gained. This work is under way, however, based on the system for describing health states outlined earlier.

Figure 2 -- Relative Risks of "Preventive" Tamoxifen for Various Diseases



endometrial cancer, there are highly uncertain but often substantially adverse effects. Given these results, using the drug tamoxifen to prevent breast cancer in otherwise healthy women is a challenge to assess, as it has both positive and negative effects. Because most of the trial was not in Canada, and to some extent people self-select to participate, the population enrolled in the clinical trial was not representative of the target population in Canada. And as clearly evident in Figure 2, even these best-available clinical trial results have a substantial degree of uncertainty, represented by the wide statistical confidence intervals for some endpoints.

These kinds of interventions pose numerous challenges to making a reasoned judgment as to whether, or how, the intervention should be adopted. If a large proportion of the population is potentially affected, costs to governments, insurance plans, or individual women and men, depending on how these are financed, will likely be important. Because these interventions may have both positive and negative effects, an informed judgment should be based on a careful and explicit assessment of the likely joint or overall impacts of all these effects on the population of intended beneficiaries. The tamoxifen trial conducted by Fisher et al.[53] did not find any

statistically significant differences in overall mortality rates between the tamoxifen and the control groups over the study period of less than 5 years, and thus the trial itself offers no guidance on this most basic question. When there are mixed results like these and heterogeneities of populations matter, differences between the trial population and the intended group of beneficiaries can be important.

In this case, a detailed analysis of the likely net impacts of preventive tamoxifen was undertaken for the population of Canadian women, using POHEM.[54] This model was used to bring together data and modules to simulate explicitly the results for each of the diseases shown in Figure 1 in a competing-risks framework and over the full remaining lifetimes of the women who would be involved. The results were sobering, contrasting sharply with the favorable conclusions drawn by the principal investigators of the randomized trial,[53] which were so favorable that the trial was even stopped early on the grounds that the benefits were evident enough to make it unethical to continue the placebo arm of the trial. In contrast, Will et al.[54] found that for a representative cohort of Canadian women, targeted at the risk levels approved by the U.S. Food and Drug Administration, preventive tamoxifen under the preferred scenario in Fisher et al.[53] would involve having 40% of all women being on the drug at some point in their lives, and it could have a net *negative* effect on their life expectancy.

This kind of result can generally not be obtained by a single clinical trial. Such extrapolations to actual intended populations require the use of simulation models like POHEM, which in turn draw on a base of coherent and extensive (longitudinal) data on the dynamics of disease processes and health interventions, including both their costs and impacts. At the same time, such analytical tools are long overdue to underpin myriad choices and policy decisions in the health sector.

As noted in connection with the study of colorectal cancer screening mentioned above, POHEM cannot yet produce results in terms of an SMPH, specifically HALE. With the system for describing health states described in the right-most column of Table 1, however, its implementation in numerous Statistics Canada health surveys, and the linkage of these survey results (given respondents' consent) to health care administrative records, we expect to have the

kinds of "Rosetta stone" data sets within a few years that would provide the empirical foundations to extend POHEM in this direction. We will then be much closer to achieving the objective of a coherent system of health statistics that would have at its apex an SMPH and at the same time provide policy-relevant "drill down" and "what if" capacities.

What Remains to Be Done?

So far, this paper has described in some detail just how SMPHs are constructed and has sketched how these summary measures of population health, in combination with an underlying statistical framework and capacity for simulation modeling, can play fundamental roles in health policy. While extraordinarily promising, these statistical tools are still in their infancy; several developments are essential to their growth and use.

Development of Health Information Systems

One necessary development is the continued evolution of health information systems. Repeated cross-sectional population health surveys are clearly fundamental, but there is a need to move beyond these to longitudinal and hybrid data—connecting data on individual persons' many health care encounters into longitudinal trajectories of care, and linking these administrative data to self-reported information from population surveys, respectively. The move toward electronic health records offers not only the prospect of major improvements in the quality of patient care but also an extraordinary foundation of data for statistical analysis, including management information to support improvements in system performance and, more important for this discussion of the policy application of SMPHs, the data needed to build models like POHEM that can connect broad SMPH results to the kinds of policy levers that are practical within the health system. One key to this evolution will be the widespread use of a common generic description of health status, not necessarily on a census basis for all health care encounters in all kinds of settings, but certainly for a strategically selected sample, as proposed by Gold et al.[55]

Support for Academic Research

Given a growing mass of richly multivariate, longitudinal person-level data on health and health care trajectories, the next challenge is distilling useful information—nuggets of empirical regularity (e.g., relative risks, survival curves, disease progression rates, success rates, and quantified benefits for various health interventions). One of the biggest challenges is explicating the key causal pathways, or "web of causality".[56] For this, there needs to be support for academic research, but this research cannot be entirely driven by investigators. First, there is no strong academic tradition in this area. Second, the research needs a degree of coherence, if not explicit coordination, to ensure that its overall coverage and breadth is appropriate. As a result, these endeavors will need to be supported by specially targeted research funding.

For-Profit Research

A very substantial portion of research into the impacts of health interventions is undertaken by the private sector, for example, by major pharmaceutical firms. The only common endpoint or outcome measure at present is mortality. Much of the benefit of these interventions however, is not so much in saving lives but in improving health-related quality of life. To provide a common metric for comparing these interventions, and to enable them to be placed within the framework of SMPHs, there must be widespread use of a generic description of health status, again as recommended in Gold et al.[55] Ultimately, this might occur only if the relevant regulatory authorities (e.g., the U.S. Food and Drug Administration) mandated such measures in all reports on efficacy or effectiveness required for approval of a drug or medical device.

The benefits of the widespread adoption of a common generic description of health status (not exclusively, of course; other measures tailored to the specific kind of intervention should also be included in the studies) would be substantial. It would open up a broad body of scientific research to cross-fertilization, to meta-analysis, and to synthesis into larger statistical and "what if" modeling frameworks.

Exchange of Knowledge

This paper has developed a vision of health information, with an SMPH at its apex and a greatly expanded commonality of purpose in a diverse range of research and analytical activities, with the objective of developing a much stronger evidentiary base for health policy. This vision implies large-scale and coherent efforts. It is unlikely that any one agency can coordinate efforts like this, but with strategic investments and a breadth of sharing knowledge, myriad efforts can come together with a coherence that would be tantamount to coordination.

The requisite sharing of knowledge will require new mechanisms and networks. It will involve more than an increased volume of articles in peer-reviewed journals or even the creation of new journals. A joint effort is required, bridging the worlds of health policy, national statistical systems, and academic research. Efforts to realize the full potential of SMPHs—defined to include a broad statistical infrastructure of which they are only the most visible part—will require the leadership of key agencies, resources, and a shared vision of the benefits.

References

1. World Bank. *World Development Report: Investing in Health.* New York: Oxford University Press; 1993.
2. World Health Organization. *World Health Report 2000—Health Systems: Improving Performance.* Geneva: WHO; 2000. Available at: http://www.who.int/whr/2000/en/index.html.
3. Wilkins R, Adams O. *Healthfulness of Life.* Montreal, Quebec: Institute for Research on Public Policy; 1983.
4. Wolfson MC. A system of health statistics toward a new conceptual framework for integrating health data. *Rev Income Wealth.* 1991;37:81-104.
5. Wolfson MC. Health-adjusted life expectancy. *Health Rep.* 1996;8:41-46.
6. Sullivan DF. A single index of mortality and morbidity. *HSMHA Health Rep.* 1971;86:347-354.
7. Murray CJL. Rethinking DALYs. In: Murray CJL, et al., eds. *Summary Measures of Population Health: Concepts, Ethics, Measurement and Applications.* Geneva: WHO; 2002.
8. van Dalen H, Williams A, Gudex C. Lay people's evaluations of health: are there variations between different subgroups? *J Epidemiol Community Health.* 1994;48:248-253.
9. Blaxter M. *Health and Lifestyles.* London and New York: Tavistock-Routledge; 1990.
10. WHO. *International Statistical Classification of Diseases and Related Health Problems, Tenth Revision.* Vols 1-3. Geneva: WHO; 1992-1994.
11. WHO. *International Classification of Functioning, Disability, and Health (ICF).* Geneva: WHO; 2001.
12. Ustun B. The international classification of functioning, disability and health — a common framework for describing health states. In: Murray CJL, et al., eds. *Summary Measures of Population Health: Concepts, Ethics, Measurement and Applications.* Geneva: WHO ;2002:343-348.
13. WHO. *International Classification of Impairments, Disabilities, and Handicaps (ICIDH).* Geneva: WHO; 1980.

14. Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36): I. Conceptual framework and item selection. *Med Care.* 1992;30:473-483.
15. Feeny D. The utility approach to assessing population health. In: Murray CJL, et al., eds. *Summary Measures of Population Health: Concepts, Ethics, Measurement and Applications.* Geneva: WHO; 2002:515-528.
16. Feeny D, et al. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. *Med Care.* 2002;49:113-128.
17. Brooks R. EuroQoL: the current state of play. *Health Policy.* 1996;37:53-72.
18. EuroQol Group. EuroQoL—a new facility for the measurement of health-related quality of life. *Health Policy.* 1990;16;199-208.
19. WHO. World Health Survey. Available at: www.who.int/whs.
20. Sadana RA, et al. Describing population health in six domains: comparable results from 66 household surveys. *GPE Discussion Paper,* No. 43. Geneva: WHO/Global Programme on Evidence for Health Policy; 2002. Available at: http://www.hsph.harvard.edu/burdenofdisease/publications/papers/Describing%20Population%20Health.pdf.
21. Bernier J, Wojciechowski T. The dimensions of health: a study using the Canadian National Population Health Survey. Mimeo: Health Analysis and Measurement Group, Statistics Canada. Ottawa, Canada; 2003.
22. Houle C, Berthelot JM. A head-to-head comparison of the health utilities index mark 3 and the EQ-5D for the population living in private households in Canada. *Quality of Life Newsletter.* 2000;5-6.
23. Gorber S. Measurement of health state preferences in Canadians: introduction and administration of the feeling thermometer exercise. Mimeo: Health Analysis and Measurement Group, Statistics Canada: Ottawa, Canada; 2003.
24. Murray CJL. Rethinking DALYs. In: Murray CJL, Lopez AD, eds. *The Global Burden of Disease.* Vol 1. Cambridge, Mass: WHO, Harvard University Press; 1996.
25. Sadana R, et al. Comparative analyses of more than 50 household surveys on health status. In: Murray CJL, et al., eds. *Summary Measures of Population Health: Concepts, Ethics, Measurement and Applications.* Geneva: WHO; 2002:369-386.

26. Salomon JA, Murray CJL. Estimating health state valuations using a multiple-method protocol. In: Murray CJL, et al., eds. *Summary Measures of Population Health: Concepts, Ethics, Measurement and Applications.* Geneva: WHO; 2002:487-499.
27. Feeny D, et al. A multiattribute approach to population health status. In: *Proceedings of the 153rd Annual Meeting of the American Statistical Association, August 8-12, 1993, San Francisco, CA.* Alexandria, VA: American Statistical Association; 1994.
28. The MVH Group. *The Measurement and Valuation of Health. First Report on the Main Survey.* York, UK: Centre for Health Economics, University of York; 1994.
29. Stouthard MEA, Essink-Bot ML, Bonsel GJ. Dutch Disability Weights Group. Disability weights for diseases: a modified protocol and results for a western European region. *Eur J Public Health.* 2000;10:24-30.
30. Province of Quebec. *Report of the Commission of Inquiry on Health and Social Sciences (Rochon Commission).* La Commission D'enquete Sure Les Services de Sante et Les Services Sociaux. Quebec: Gouvernement de Quebec; 1987.
31. Fries J. Aging, natural death, and the compression of morbidity. *N Engl J Med.* 1980;303:130-135.
32. Mathers CD, Robine JM, Wilkins R. Health expectancy indicators: recommendations for terminology: In: Mathers CD, McCallum J, Robine JM, eds. *Advances in Health Expectancies: Proceedings of the 7th Meeting of the International Network on Health Expectancy (REVES), February 1994, Canberra, Australia.* Canberra: Australian Institute of Health and Welfare; 1994.
33. Brock DW. Ethical issues in the development of summary measures of population health status. In: Field MJ, Gold MR, eds. *Summarizing Population Health: Directions for the Development and Application of Population Metrics.* Washington, DC: Institute of Medicine, National Academy Press; 1998:73-86.
34. Daniels N. Distributive justice and the use of summary measures of population health status. In: Field MJ, Gold MR, eds. *Summarizing Population Health: Directions for the Development and Application of Population Metrics.* Washington, DC: Institute of Medicine, National Academy Press; 1998:58-72.
35. Gakidou EE, Murray CJ, Frenk J. Defining and measuring health inequality: an approach based on the distribution of health expectancy. *Bull World Health Organ.* 2000;78:42-54.

36. Wolfson MC, et al. Career earnings and death: a longitudinal analysis of older Canadian men. *J Gerontol.* 1993;48:S167-S179.
37. Wolfson M, Rowe G. On measuring inequalities in health. *Bull World Health Organ.* 2001;79:553-560.
38. Murray CJL, Salomon JA, Mathers CD. The individual basis for summary measures of population health. In: Murray CJL, et al., eds. *Summary Measures of Population Health: Concepts, Ethics, Measurement and Applications.* Geneva: WHO; 2002:41-51.
39. Nault F, Roberge R, Berthelot JM. Espérance de vie et espérance de vie en santé selon le sex, l'état matrimoniale et le statut socio-économique au Canada (in French). *Cahiers Quebecois de Demographie.* 1996;25:241-259.
40. Sondik E. Summary measures of population health: applications and issues in the United States. In: Murray CJL, et al., eds. *Summary Measures of Population Health: Concepts, Ethics, Measurement and Applications.* Geneva: WHO; 2002:75-81.
41 National Center for Health Statistics. *Healthy People 2000 Final Review.* Hyattsville, Md: Public Health Service; 2001. Available at: http://www.cdc.gov/nchs/data/hp2000/hp2k01.pdf.
42. WHO. *World Health Report 2002—Reducing Risks, Promoting Healthy Life.* Geneva: WHO; 2002.
43. Wolfson MC. POHEM—a framework for understanding and modelling the health of human populations. *World Health Stat Q.* 1994;47:157-176.
44. Will BP, et al. Canada's Population Health Model (POHEM): a tool for performing economic evaluations of cancer control interventions. *Eur J Cancer.* 2001;37:1797-1804.
45. First Ministers' Meeting Communique on Health [press release]. 2000. Available at: http://www.scics.gc.ca/cinfo00/800038004_e.html.
46. Commission on the Future of Health Care in Canada. *Final Report: Building on Values: The Future of Health Care in Canada.* 2002. Available at: http://www.hc-sc.gc.ca/english/pdf/romanow/pdfs/HCC_Final_Report.pdf.
47. Senate of Canada. *The Health of Canadians—The Federal Role. Final Report.* Vol 6. 2002. Available at: http://www.parl.gc.ca/37/2/parlbus/commbus/senate/com-e/soci-e/rep-e/repoct02vol6-e.htm.

48. Maxwell J, et al. Commission on the Future of Health Care in Canada. *Report on Citizens' Dialogue on the Future of Health Care in Canada.* 2002. Available at: http://www.hc-sc.gc.ca/english/pdf/romanow/pdfs/Dialogue_E.pdf.
49. 2003 First Ministers' Accord on Health Care Renewal. 2003. Available at: http://www.hc-sc.gc.ca/hcs-sss/delivery-prestation/fptcollab/2003accord/index_e.html.
50. Health Canada. A framework to improve the social union for Canadians. 1999. Available at: http://www.scics.gc.ca/cinfo99/80003701_e.html.
51. Wei M, et al. Relationship between low cardiorespiratory fitness and mortality in normal-weight, overweight, and obese men. *JAMA.* 1999;282:1547-1553.
52. Health Canada. Reducing Canadian colorectal cancer mortality through screening. 2003. Available at: http://www.phac-aspc.gc.ca/publicat/ncccs-cndcc/ccsrec_e.html.
53. Fisher B, et al. Tamoxifen for prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. *J Natl Cancer Inst.* 1998;90:1371-1388.
54. Will BP, et al. First do no harm: extending the debate on the provision of preventive tamoxifen. *Br J Cancer.* 2001;85:1280-1288.
55. Gold MR, et al. *Cost-Effectiveness in Health and Medicine.* New York: Oxford University Press; 1996.
56. Krieger N. Epidemiology and the web of causation: has anyone seen the spider? *Soc Sci Med.* 1994;39:887-903.

Contact Information:

Michael C. Wolfson, Ph.D., B.Sc., Statistics Canada, wolfson@statcan.ca.

Commentary

Article: On the Policy Implications of Summary Measures of Health Status, by Michael C. Wolfson, Ph.D., B.Sc.

Robert M. Kaplan, Ph.D.

There is an obvious need for comprehensive ways of examining health outcomes. Unfortunately, progress in this area has been slow. In this commentary, I will discuss some reasons why this has been the case.

Part of the problem is that we have been distracted by disagreements about the merits of alternative methods. Developers typically think their measures are the best. Moreover, we have had disagreements on the general philosophy of outcome measurement—for example, generic versus disease-specific measures, psychometric versus utility measures, etc. We also have discipline-related differences. The field includes contributors from statistics, economics, medicine, psychology, and anthropology. Different training has sent us to chase different issues.

There have been three major distractions. First, we have differences of opinion about how to measure preference or utility. Dr. Wolfson mentioned this in his article. Utilities have been measured using the standard gamble, the time trade-off, and the visual analog scale.

Different scaling methods do produce different results. However, this need not halt efforts to quantify population health. We should think of different scaling techniques as different scoring systems. All approaches require the measurement of health states. Once health states have been classified, we can apply scoring functions based on standard gamble, time trade-off, or visual analog scales. Sensitivity analysis can be used to evaluate the impact of the different weighting systems upon scores.

The response shift problem has been the second major distraction. Preferences change over time, and preferences are different between patients and non-patients. As a result, some people argue

that preference weights have no meaning. The response shift literature focuses on subjective reports by patients about their own well-being. People adapt to illnesses and this adaptation is reflected in higher self-ratings, even though underlying objective health status has not changed. The distraction is that the response shift literature is asking a different set of questions. Utility-based, population-based measures often concentrate ratings on health states along a continuum ranging from death to wellness. Using these methodologies and scaling techniques such as the visual analog scale, patients and non-patients are very similar in the way they rate cases. The response shift is an important, but different phenomenon. It need not distract us from developing population health status measures.

The third major distraction has been that policy options cut across different components of a very fragmented health care system. In order to make broad policy comparisons, measures must be generic, so that all options can be quantified using the same measurement unit. The distraction has been the urge to apply disease-specific outcome measures. It is true that these disease-specific measures are often more sensitive. However, concentration on disease-specific outcomes gives up the primary advantage of generic utility-based outcome measures. Only generic measures can be used for broad policy comparisons.

Progress in the development of comprehensive health outcome measures has been slow. However, we have made much more progress than people realize. We do agree on some of the core issues. In fact, most of the measures can be traced back to Daniel F. Sullivan. Sullivan proposed indexes of adjusted survival and argued for the need to create separate components for mortality, morbidity, and preference. Most authors do not cite Sullivan, and I was very pleased to see that Dr. Wolfson's article acknowledged his work. Dr. Wolfson also discussed some of the important conceptual work on the concept of health status. He cited qualitative studies from Europe that identify important notions of health. These concepts match the content of many of our current measures. When asked what is important to them, focus group participants identify ability to function, freedom from symptoms, freedom from psychological distress, and so forth. The content of our measures is similar and includes most of the concepts identified in qualitative studies.

We have come a long way, but there is still much to be done. However, there are methods that can be applied right now. For example, in the United States, we can apply the HALex measure, NAFIS HUI, and QWBX1. Despite the many distractions, well-validated tools are available today for the assessment of population health status.

Contact Information:
Robert M. Kaplan, Ph.D., Professor and Chair, Department of Health Services
UCLA School of Public Health, rmkaplan@ucla.edu.