Summarizing Population Health

Directions for the Development and Application of Population Metrics

Committee on Summary Measures of Population Health

Marilyn J. Field and Marthe R. Gold, *Editors*

Division of Health Care Services

INSTITUTE OF MEDICINE

NATIONAL ACADEMY PRESS
Washington, D.C. 1998

NATIONAL ACADEMY PRESS • 2101 Constitution Avenue, N.W. • Washington, D.C. 20418

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

The Institute of Medicine was chartered in 1970 by the National Academy of Sciences to enlist distinguished members of the appropriate professions in the examination of policy matters pertaining to the health of the public. In this, the Institute acts under both the Academy's 1863 congressional charter responsibility to be an adviser to the federal government and its own initiative in identifying issues of medical care, research, and education. Dr. Kenneth I. Shine is president of the Institute of Medicine.

Support for this project was provided by the U.S. Department of Health and Human Services (Contract Number 282-97-0019). The views presented are those of the Institute of Medicine committee and are not necessarily those of the funding organizations.

International Standard Book Number: 0-309-06099-0

Additional copies of this report are available for sale from:

National Academy Press
2101 Constitution Avenue, N.W.
Box 285
Washington, DC 20055
Call 800-624-6242 (or 202-334-3313 in the Washington metropolitan area), or visit the NAP's on-line bookstore at: **www.nap.edu**

For more information about the Institute of Medicine, visit the IOM's home page at **www2.nas.edu/iom**.

Printed in the United States of America

COMMITTEE ON SUMMARY MEASURES OF POPULATION HEALTH

Harvey V. Fineberg,* M.D., Ph.D. (*Chair*), Provost, Harvard University
Mohammed Akhter, M.D., Executive Director, American Public Health Association
Norman Daniels, Ph.D., Professor of Philosophy, Tufts University
Dennis Fryback, Ph.D., Professor of Preventive Medicine, University of Wisconsin, Madison
Dean Jamison,* Ph.D., Director, Center for Pacific Rim Studies, University of California Los Angeles
Mary Mundinger,* Dr.P.H., Dean and Centennial Professor of Health Policy, Columbia University
Paige R. Sipes-Metzler, D.P.A., Manager of Quality Management, Regence Blue Cross and Blue Shield Oregon

Staff

Marthe R. Gold, M.D., M.P.H., Project Consultant, Chair, Department of Community Health and Social Medicine, City University of New York Medical Center

Marilyn J. Field, Ph.D., Study Director and Deputy Director, Health Care Services
Dana Caines, Administrative Assistant (September–December 1997)
Cecilia Rossiter, Administrative Assistant (January–March 1998)
Kay Harris, Financial Associate

*Member, Institute of Medicine.

Acknowledgments

In developing this short report, the committee and staff particularly benefited from the experience and expertise of the presenters and other participants in the workshop the committee convened on December 12 and 13, 1997, in Washington, D.C. Appendix A lists the workshop participants, presenters, and agenda.

This report was subject to independent review by a group of outside experts chosen for their diverse perspectives and technical knowledge. The purpose of this review, which was conducted in accord with procedures established and overseen by the Report Review Committee of the National Research Council, was to assist the report authors and the IOM in producing a published report that met institutional standards for objectivity, evidence, and responsiveness to the study charge. The content of the review comments and the draft manuscript remain confidential to protect the integrity of the deliberative process. On behalf of the Institute of Medicine, the study committee wishes to think the following individuals for this participation in the review of this report: David Kindig, Ph.D., University of Wisconsin, Madison School of Medicine; John Ludden, M.D., Harvard Pilgrim Health Care; Donald Marazzo, Ph.D., University of Pittsburgh School of Public Health; Jean-Claude Poullier, Ph.D., Organization for Economic Cooperation and Development; Steven Teutsch, M.D., Merck and Company, Inc.; Hugh Tilson, M.D., Glaxo Wellcome Company; Alan Williams, Ph.D., University of York (United Kingdom); and Michael Wolfson, Ph.D., Statistics Canada. These reviewers provided many constructive comments and suggestions, but responsibility for the final content of this report rests solely with the Committee on Summary Measures of Population Health and the Institute of Medicine.

Linda Bailey, J.D., the project officer for the study at the Department of Health and Human Services, helped in many ways with the study. The committee also benefited from discussions with a DHHS working group on measures of population health. Members included Yen-pin Chiang, Ph.D., Matthew McKenna, M.D., M.P.H.., Gregory Pappas, M.D., Ph.D., James Schuttinga, Ph.D., and Fred Seitz, Ph.D. Dr. Chiang organized a meeting on topics related to summary measures that made it possible for a number of additional experts to attend the Institute of Medicine (IOM) workshop. Earlier, DHHS, Jo Ivey Boufford, M.D., and Kristine McCoy helped establish directions for the project and, generally, made the study possible. At the IOM, Christopher Howson, Ph.D., spent many months negotiating the framework for the study. Claudia Carl, Mike Edington, and Evelyn Simeon assisted at various stages in the project.

Contents

Summarizing Population Health: Directions for the Development and Application of Population Metrics

SUMMARY

Historically, policies to improve population health have focused on major causes of death such as smallpox and cholera. Policy priorities have, in turn, been guided by information on mortality and life expectancy, and governments and others have worked to collect comprehensive, reliable, and valid mortality data. As death rates have decreased and life spans have lengthened, however, people have become increasingly interested in other health goals such as preventing disability, improving functioning, and relieving pain and the distress caused by other physical and emotional symptoms. With broader goals, policymakers need additional information to help them make decisions and establish priorities for public health, biomedical research, and personal health services.

For some purposes and decisions such as making individual patient care decisions or reducing postoperative infection rates, detailed clinical, behavioral, and organizational information is required. For other purposes such as understanding broad trends in the public's health or comparing the value of population health promotion strategies, it is helpful to have some overall picture or summary measure of health and well-being in addition to information on specific aspects or dimensions of health.

The development and application of summary measures of population health present complex and intriguing methodological, ethical, and political challenges. Methodologists have taken the lead in confronting these challenges, for example, in devising ways to summarize in a single measure the impact on population health of both mortality and morbidity. They generally have used one of several different methods to attach a single number—usually ranging between 0 (death) and 1 (optimal health)—to a complex of social and personal attributes that represent health status. This number has then been linked to life expectancy to form a single integrative measure of overall health. Under the overall measurement rubrics of health-adjusted life years (HALYs) or health-adjusted life expectancy (HALE), several kinds of measures have been developed. The best known include quality-adjusted life years (QALYs), years of healthy life (YHLs), and disability-adjusted life years (DALYs). Methodologists are still refining these measures to improve their reliability, validity, credibility, and ease of use.

Important as methodological concerns and advances are, the ethical and policy implications of different measures and measurement strategies also deserve more systematic attention. In particular, alternative ways of valuing the duration of life, the quality of life, the burden of ill-health, or inequalities in health incorporate critical but not necessarily obvious or well-accepted judgments about whose life or what kind of life has meaning and worth. It is, therefore, important to examine—empirically and normatively—how the use of summary measures of population health can shape, improve, or distort decisions and how the analyses and resulting decisions compare to partial measures and to more traditional, often informal decisionmaking approaches. Much of the debate about summary measures is actually about policy choices that their use makes more explicit. Policymakers and policy analysts at all levels—international, national, regional, and local—would benefit from a better understanding of the strengths and limitations of different measures in informing decisions about how to invest limited resources to improve population health and well-being.

This report from an Institute of Medicine (IOM) committee is intended to encourage methodologists, ethicists, and policymakers to learn from each other and to work together to identify the strengths, limitations, and appropriate uses of summary measures. In addition to the committee's own expertise and experience, the report builds on discussions during a December 1997 workshop and the background papers (see appendixes) drafted for the workshop. The conclusions and recommendations that follow, each of which is discussed in more detail in subsequent sections of this report, describe directions for work to strengthen the credibility and utility of summary measures of population health.

First, mortality measures, although important, provide decisionmakers incomplete and insensitive information about overall population health. Summary measures of population health need to recognize the physical and psychological illnesses and disabilities that cause much individual suffering and limit social and economic advances within and across nations.

Second, summary measures of population health that integrate mortality and morbidity information are increasingly relevant to both public health and medical decisionmakers. Their actual and proposed uses include describing differences and trends in the health of populations; informing decisions about alternate uses of public health care dollars; and assessing the cost-effectiveness of alternative personal health care services and technologies.

Third, the similarities and differences among summary measures of population health require further examination as part of a strategy for assessing how well particular measures and measurement strategies may serve different local, national, and international purposes. These assessments also need to include comparisons with simpler measures.

Fourth, although methodological innovation in population metrics has strengthened the analytical base for health decisions, the lack of accepted standard measures often creates confusion and caution among potential users.

Fifth, 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 and those who promote their use.

From these conclusions, the committee derived three recommendations. These recommendations, which are directed primarily at the U.S. Department of Health and Human Services

(DHHS), are intended to encourage the design of summary measures that are readily understood and helpful and to promote use of these measures in ways that are accountable and credible. The committee recommends that DHHS do the following:

1. Initiate a process of analysis and public discussion to (a) clarify the ethical assumptions and value judgments embedded in different measures of population health and (b) assess the critical ethical and policy implications of differing designs, implementation approaches, and uses of these measures.

2. Create a process to establish standards for population health metrics and to investigate the value and practicality of a compatible set of summary measures of population health that could be used for different descriptive and decisionmaking purposes.

3. Invest in the education and training of public health and medical professionals to promote their understanding of the interpretation and appropriate use of summary measures of population health.

BACKGROUND

Origin of the Report

This brief report is the product of a short-term IOM project that was funded by DHHS to provide guidance on future directions for the development and application of summary measures. To meet this charge and prepare this report, the IOM established a seven-member committee of experts in public health, ethics, policy analysis, and measurement development and use.

At the request of DHHS, the committee, which was appointed in September 1997, convened a workshop that would bring together a carefully balanced group of methodologists, ethicists, and policymakers to discuss the design and use of summary measures of population health and to provide guidance on steps that would strengthen the credibility and utility of the measures. The December 1997 workshop included one day of presentations from national and international experts and a second day devoted to small-group discussion of issues and directions. The workshop agenda, list of participants, and background papers prepared for the workshop are included here as appendixes.

The IOM committee met immediately following the workshop to consider its conclusions and recommendations and to outline a short report. In addition to the committee's own expertise and perspectives, the report draws on the workshop presentations and small-group discussion as well as on the background papers drafted for the workshop by Marthe Gold, Dennis Fryback, Norman Daniels, and Daniel Brock. The focus is on the United States, but the committee benefited from the international expertise and experience represented by participants from Canada, the Netherlands, the United Kingdom, Mexico, and the Organization for Economic Cooperation and Development as well as others who advise international groups such as the World Health Organization and the World Bank.

Terminology

One problem facing the committee was that the nomenclature for concepts and measures of population health derives from a number of literature streams, and the committee found itself at a confluence of these streams, where the same names were being applied to quite different things. For purposes of this report, a *measure of population health* can involve mortality data (e.g., age-adjusted mortality rates for a given year, life expectancy at birth or age 65) or morbidity data (e.g., disability rates or quality of life indices) or both. The focus of the report is on *summary measures of population health* that combine both mortality and morbidity data to represent overall population health in a single number (e.g., a health-adjusted life year or health-adjusted life expectancy).

Figure 1 depicts how mortality and morbidity measures—as partial measures of population health—combine to form an integrative measure. The background paper by Fryback describes the process in more detail.

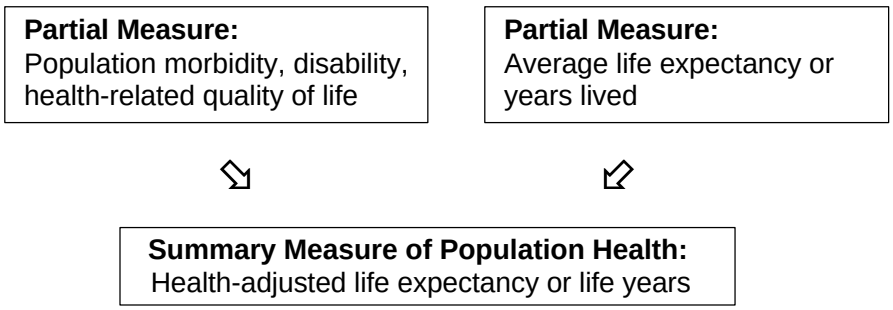


FIGURE 1 Building a summary measure from partial measures of health.

Because *measures of health-related quality of life* (HRQL) are important building blocks of these integrated measures and generate much of the controversy about these measures, such measures are also considered in this report and the background papers. To depict an individual's overall health at a particular time, these measures generate a single number on a scale anchored by 0 (state of being dead) and 1 (state of optimal health) that represents the degree or strength of preferences for one health state over another. For some purposes including cost-effectiveness analysis, measures of health-related quality of life must be based on utilities or preferences for health states that meet the conditions of welfare economics, which assumes that individuals seek to maximize *utilities* (preferences for particular outcomes) and that overall societal welfare is some function of these individual utilities.

A variety of other measures of health status have been developed. One broad category of measures constructs profiles of people's health along one or more dimensions. The measures include the SF-36, the 36-item short-form of the Medical Outcomes Study survey created from the Rand General Health Survey (Brook et al., 1979) and the Sickness Impact Profile (Bergner, 1981). Because such profiles do not yield a single summary number from, for example, scores representing physical or mental functioning, these measures of health status have not been combined with life expectancy measures to form an integrative summary measure of population

health. Measures such as the SF-36 are widely used in observational and experimental studies to characterize changes over time in the health of groups or individuals.

A glossary provides some additional definitions for several general terms and specific measures used in this report. These definitions are not, however, universally accepted, and both explicit disagreement and unrecognized differences in word use characterize the field. Work to clarify terminology is one element covered by the committee's recommendation about standard setting.

CONCLUSIONS

The committee's conclusions and recommendations are intended to provide general rather than detailed perspectives and directions for the future development and application of summary measures of population health. The IOM originally proposed a more intensive, three-year investigation of technical and policy issues and may still seek to undertake this work and prepare an in-depth report on some of the issues raised here, for example, common terminology and definitions.

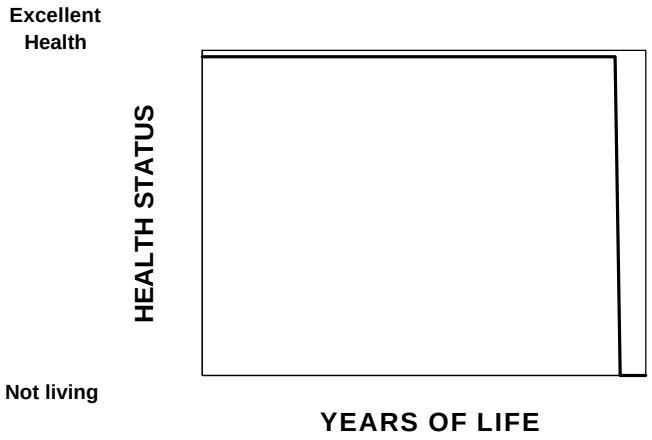
Mortality measures, although important, provide incomplete and insensitive information for decisionmaking

Many major decisions about individual and public health are today informed primarily by summary data on population mortality. The family of mortality-based summary measures includes measures that aggregate data over causes, for example, age-adjusted death rates and age-specific life expectancy. Measures may also be broken down to describe the mortality experience of different population subgroups (e.g., men and women or different ethnic or racial groups) or the number of deaths or years of life lost due to different causes (e.g., heart disease or suicide)—given suitable data for attributing cause of death.

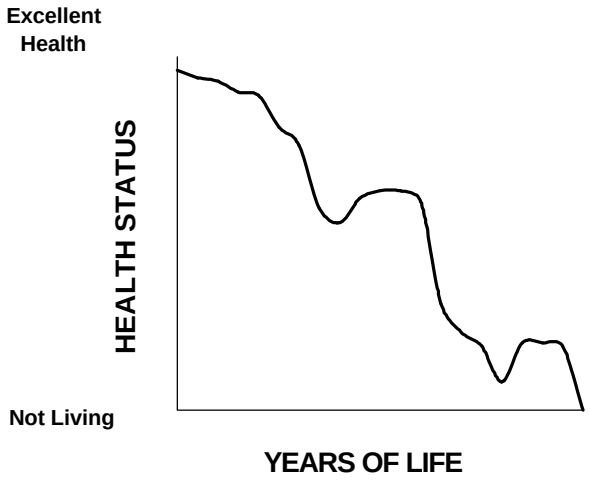
Both ordinary people and policymakers are deeply interested in extending life. At the same time, they recognize other important health goals including preventing disability, improving physical and mental functioning, and relieving pain and the distress caused by other physical and emotional symptoms. Failure to include morbidity data in summary measures of health status distorts the resulting profiles of disease burden and the information available for assessing needs for preventive, curative, palliative, and rehabilitative services.

The limits of mortality data are perhaps most evident for diseases and injuries that impose serious, continuing burdens of disability and suffering on people they do not kill. Some of these diseases, for example, arthritis, are not major sources of mortality, although others, such as heart disease, are significant contributors to both mortality and morbidity.

To illustrate, Figures 2a and 2b depict the hypothetical life paths of two people, one who dies from a condition that kills suddenly and the other who dies at the same age from a condition that causes progressive disability and distress for many years before death. A measure based only on mortality would not distinguish between the health burdens created by the two conditions because life duration is identical. Such a measure would also fail to differentiate between a health intervention for the second condition that extended life and one that both extended life and reduced the burden of disability before death.



A



B

FIGURE 2 Hypothetical life paths of (A) one person who dies suddenly after living in excellent health and (B) another who dies at the same age after living with progressive disability for many years. SOURCE: Adapted from Panel on Cost-Effectiveness in Health and Medicine (1996).

That different measures produce different pictures of health status is supported by a recent analysis of the global burden of disease that separately rank-ordered causes of death and causes of ill-health using 107 disease and injury categories and data from a number of sources (Murray and Lopez, 1996). The analysis showed considerable disparity, including 14 conditions that were ranked in the top half for burden of ill-health but in the bottom half for deaths.

Summary measures of population health that integrate mortality and morbidity information are increasingly relevant to both public health and medical decisionmakers.

As policymakers and analysts have understood the limitations of mortality data alone as a basis for decisionmaking, they have become interested in other ways of viewing overall population health. In response, many individuals and organizations have worked to develop new measures that reflect both life duration *and* morbidity or health-related life quality (see, e.g., Moriyama, 1968; Fansel and Bush, 1970; Sullivan, 1971, and more generally, the background paper by Fryback). The following sections of this report provide a very basic overview of how such integrative summary measures have been developed, examples of their uses, and simple categorizations of these uses.

Developing Summary Measures of Population Health

As noted earlier, summary measures of population health are constructed by first attaching a single number—where 0 represents death and 1 represents optimal health—to a complex of social and personal attributes that represent health status or health-related quality of life. (Negative numbers representing states regarded as worse than death are not excluded from this general conceptualization.) This number is then linked to life expectancy to form a single measure of population health that integrates morbidity and mortality information.

Health-adjusted life expectancy is arrived at by summing the products of expected years of life at each age multiplied by a numerical weight representing average health status at that age. The individual products (life expectancy at a particular age multiplied by average health status at the age) are *health-adjusted life years* (HALYs). Although measures of all types seem to become known by their acronyms or abbreviations, this report generally avoids acronyms and abbreviations.

Most of the discussion of summary measures of population health focuses on one category of measure, the *quality-adjusted life year* (QALY), and its usual building blocks, measures of health-related quality of life. Even more particularly, the focus is on QALYs built from measures of health-related quality of life that are based on utilities or preferences for health states that meet or approximate conditions of welfare economics. A more detailed discussion of the theoretical basis of preference-based measures of health status is well beyond the scope of this short report (see Panel on Cost-Effectiveness in Health and Medicine, hereafter PCEHM, 1996, for an introduction to the basic concepts).

The attributes of health (sometimes called domains or dimensions) that may be captured in a measure of health-related quality of life include physical function, mental and emotional well-being, social and role function, general health perceptions, symptoms, and vitality (see, e.g., Lohr,

1989; Stewart et al., 1989; Patrick and Erickson, 1993; Williams, 1995; Kindig, 1997). Table 1 illustrates how several instruments for measuring health-related quality of life (as described in the Glossary) vary in the health dimensions or domains they include. To the extent that the labels for subscales do not reflect their substance, however, the table may not adequately depict differences among instruments. For example, newer versions of the Quality of Well-Being scale include questions about sensation and sensory organs but do not group and label these questions as a subscale (Robert Kaplan, Ph.D., University of California San Diego, personal communication, March 23, 1998).

The health domains selected for different measures reflect both different conceptualizations of health and different purposes for which measures were developed (e.g., clinical research or resource planning). (Fryback notes that the actual health states that can, in principle, be distinguished by different measures ranges from less than a dozen to more than 1,000.) As measures of health-related quality of life have become widely used, these differences complicate comparisons of conditions or populations. The choice of what aspects of health to measure or ignore (and how to define these aspects operationally) also raises ethical questions, implying perhaps that domains not included are unimportant.

TABLE 1 Principal Concepts and Domains of Health-Related Quality of Life Contained in General Preference-Weighted Instruments for Assessing Quality-Adjusted Life Years

	Instrument				
		Health Utility Index			Quality of Well-Being
Health perceptions	EQ-5D	HUI:1	HUI:2	HUI:3	Scale
Social function					
Social relations	X	X			X
Usual social role	X	X			
Intimacy or sexual function					
Communication or speech			X	X	
Psychological function					
Cognitive function			X	X	
Emotional function	X	X	X	X	
Mood or feelings					
Physical function					
Mobility	X	X	X	X	X
Physical activity		X		X	X
Self-care		X	X		
Impairment					
Sensory function or loss		X	X		
Symptoms or impairments	X	X	X	X	X

SOURCE: Adapted from Patrick and Erickson (1993) and PCEHM (1996).

The measures of health-related quality of life listed in Table 1—the Quality of Well-Being (QWB) scale, the three versions of the Health Utilities Index (HUI), and the EQ-5D—are preference-weighted health status measures because they use some procedure for rating the relative desirability or undesirability of living with or without various functional limitations (e.g.,

an inability to dress oneself), symptoms (e.g., pain or urinary frequency), or other health states. Again, the approaches differ.

The weighting or rating approaches used have included expert panels, national or community surveys that ask people to rate health states or use utility derivation instruments (e.g., standard gamble or time trade-off methods), and deliberative discussion or other means to determine "reasoned" preferences. Choices among options typically reflect a mix of conceptual, technical, and practical (e.g., cost and convenience) considerations. In addition to critiques of specific methods, the variability in weighting techniques has—like the variability in health domains tapped—been criticized as a barrier to standardizing cost-effectiveness and other analyses across illnesses and conditions (PCEHM, 1996). In addition, as discussed in later sections of this report, much of the controversy about summary measures involves the ethical implications of different approaches to rating or valuing health states.

Examples of Applications or Tests of Summary Measures

Within the United States, the Center for Chronic Disease Prevention and Health Promotion at the Centers for Disease Control and Prevention (CDC) has been testing the DALY methodology for measuring the burden of disease to assess prevention options and priorities for the nation. The nation's disease prevention and health promotion strategy, *Healthy People 2000*, has been monitored by the National Center For Health Statistics (NCHS) at the CDC. The NCHS has used the years of healthy life (YHL) measure to chart progress in increasing the span of healthy life for Americans and to reduce disparities in health status among subgroups of the population. NCHS is currently developing its approach for *Healthy People 2010* (Sondik, 1997). The Health Care Financing Administration (HCFA) is using the SF-36, a measure described earlier that profiles health status, in the Health of Senior's Survey, which will track longitudinally the health of enrollees in managed care settings (Fried, 1997).

Although cost-effectiveness analyses often still rely on mortality measures, such analyses are increasingly using QALYs or other summary measures that cover morbidity as well as mortality. For example, the U.S. Agency for Health Care Policy and Research sponsors or directs numerous cost-effectiveness analyses that incorporate such measures of health into their estimates of effectiveness. Thus, instead of estimating the cost of an additional year of life achieved by different strategies (e.g., screening for cervical cancer every year versus screening every three years), the analyses estimate the cost of an additional quality- or disability-adjusted life year.

Composite measures of population health are also being used in other countries to supplement traditional mortality measures and to illuminate the impact of disability. A Canadian task force, for example, has recently recommended the creation of the health equivalent of the Gross Domestic Product to act as an overall indicator of trends in the nation's health (Wolfson, 1997). In the Netherlands, analysts have devised variants on the traditional survival curve (a line) that not only track the proportion of a population dying at each age but also depict levels of slight or severe disability as shaded areas within the curve (Gunnig-Schepers, 1997) (see Figure 3). Such graphs, similar versions of which have been used in Canada, help make evident the extent to which slight versus severe disability increases with age (in general or for a particular condition) for a population or population subgroup.

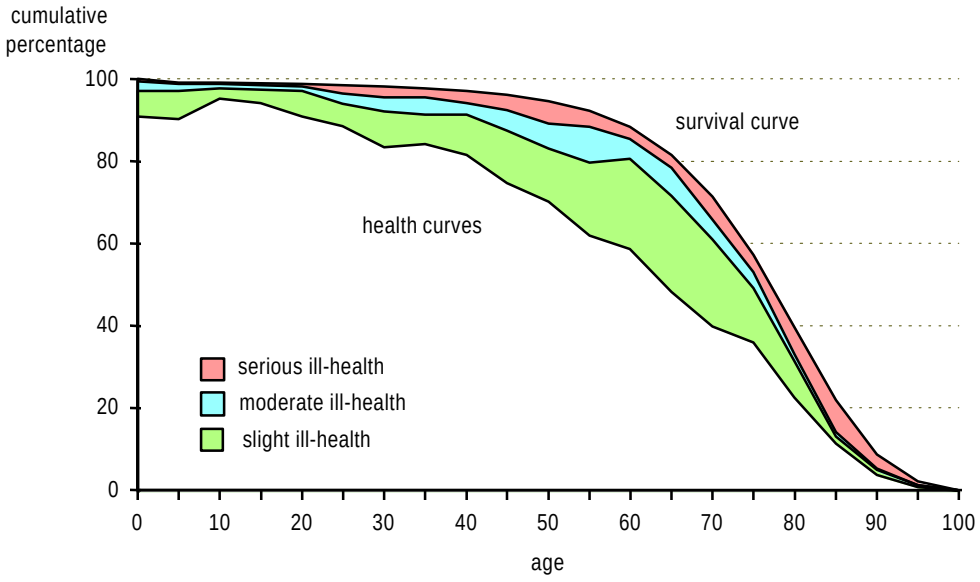


FIGURE 3 Survial curve and "health curves" by level of ill health, Dutch men, 1994. SOURCE: Dutch Public Health Status and Forecast Report, 1997.

In the international arena, comparisons of health status across countries are useful to focus attention on countries with critical needs for international assistance, to set priorities for investments in combating particular conditions and diseases, and to gain an understanding of differentials in the burden of disease among nations. The World Health Organization and the World Bank are among the organizations sponsoring such comparisons. A recent closed meeting of member countries of the Organization for Economic Cooperation and Development identified their requirements for health outcomes measures to monitor trends in population health, to strengthen the science base for evaluating the effectiveness of health interventions (e.g., screening for various illnesses), and to assist in guiding resource allocation decisions among competing health needs within member countries (Poullier, 1997).

Descriptive and Evaluative Uses of Summary Measures

Much of the discussion during the December 1997 workshop and the IOM committee deliberations focused on alternative uses of summary measures and the implications of these alternatives for the approaches used to construct various measures. Although the committee and workshop participants recognized that the categories are not independent, they found it helpful to distinguish between descriptive and decisionmaking uses of summary measures of population health.

Political and cultural values and biases, of course, play a major role in determining who benefits from public policies, but descriptive information is an important resource for those seeking to assess and modify established spending patterns. The reliable, valid description of health status is an important achievement in and of itself. For example, such information resources as the CDC's *Morbidity and Mortality Weekly Reports* represent more than a century's worth of

effort to develop accurate state and local data on death, disease, and injury. Efforts to devise summary measures that provide an overall view of individual or population health are more recent but still date back several decades. Such measures include both the preference-based measures described above and other ways of profiling health such as the SF-36.

Descriptive data on health status are often an important first step in identifying health problems as part of a process for improving community health, deploying resources, and overseeing the performance of those responsible in some way for individual or population health (IOM, 1997). For these purposes, comparisons are needed on both cross-sectional (single point in time) and longitudinal bases. Then, within populations, summary comparisons broken down—when data permit—by age, sex, ethnic, income, and other categories and by disease condition contribute to further analyses that seek to identify and understand differences in health status among different segments of the population.

A particularly important type of descriptive use of summary measures involves comparisons that identify inequalities, in health and well-being suggest testable hypotheses about the sources of these inequalities and lead to strategies for remedying them. These strategies may involve public health programs, personal health services, interventions beyond the health care system (e.g., income subsidies or antipollution programs), or some combination of these. To assess the relative impact of such strategies, it is desirable that data collection and measurement approaches be comparable across public health and personal health care systems and, if possible, beyond health systems. Comparable in this context means that the summary measures are constructed in the same nominal units (e.g., QALYs) and derived from data collected using a consistent set of conceptual and operational definitions.

The existence of summary measures of population health and the availability of techniques such as cost-effectiveness analysis not, of course, guarantee that they will be used—or used appropriately—in policymaking. For example, whether or not the Health Care Financing Administration should formally include cost-effectiveness as a criterion for making Medicare coverage decisions has generated controversy for nearly two decades. Similarly, when Oregon sought to set priorities for coverage of different health services, an initial ranking of health services developed using cost-effectiveness analysis was substantially modified by policymakers to better reflect "public values" and to rely on assessments of net benefit rather than cost-effectiveness in developing rankings (Eddy, 1991, Hadorn, 1991; Klevit et al., 1991; Patrick and Erickson, 1993). In contrast, hospital or health plan managers attempting to control costs may more readily consider cost-effectiveness in making less visible decisions (e.g., the choice of which drug to use first in managing an infection or treating depression) (Luce and Brown, 1995).

Such controversies about the use of summary measures need to be, first, understood (e.g., problems with the technical calculation or presentation of the measures should be distinguished from discomfort with the ethical implications of the measures) and, second, evaluated for possible methodological, educational, or other responses. The committee assumes neither that all barriers to use are surmountable nor that perfect measures are achievable. If the value judgments underlying a measure conflict with those of policymakers and the society that they serve, then it is reasonable that policymakers would look to other bases for making decisions. It is, however, important that these other, often informal bases be subject to similar critical analysis. The focus of the committee's work was on identifying directions that could strengthen the credibility and utility of summary measures of population health rather than specifying directions for developing alternatives.

Given both the technical and the ethical issues that need further exploration, there was considerable but not complete agreement among workshop participants that summary measures of population health were, at this time, best suited for use in descriptive comparisons of populations. There was not clear agreement on how large a role summary measures—improvements notwithstanding—should ultimately play in resource allocation decisions or comparisons of the performance of health care organizations. The committee believes that the kinds of developmental and analytic work recommended in this report will aid future discussions on the appropriate use of summary measures to guide resource decisions and performance comparisons.

The similarities and differences among summary measures of population health need further examination as part of a strategy for assessing how well particular measures and measurement strategies may serve different local, national, and international purposes.

Recommendations about roles for summary measures of population health depend, in part, on clearer understanding of (1) the uses of such measures, (2) the ways different partial and summary measures behave in depicting health, and (3) the extent to which users of the measures appreciate the characteristics or limitations of the measures and identify what additional information they may need. However, despite controversy over various issues, the ways in which different measurement strategies may shape the information and analyses provided to decisionmakers in different context has not been systematically described, evaluated, or compared to other decisionmaking tools and approaches. Some differences are clearly more consequential than others. For example, the differing strategies employed by measures of health status described earlier mean that some (those incorporating preference weights) can be combined with a measure of life expectancy to create a single integrative measure of population health. Topics that require attention include what aspects or domains of health to measure, to what extent the component domains of a summary measure can be used in conjunction with the overall measure to provide a fuller understanding of population health, and whether and how to assign weights or values to different health states and life durations. As discussed further below, these topics have important ethical as well as technical dimensions that deserve fuller exploration and consideration.

Another issue for examination is whether to treat the relationship between the valuation of health states and the duration of the health state as dependent or independent—that is, whether the time spent in a particular state might affect the perception or rating of this state. Assuming independence is methodologically convenient (Fryback, 1997) but may not be consistent with evidence suggesting that people with certain disabilities adjust and become more positive about their health state over time (Sackett and Torrance, 1978).

Further directions for investigation are described in following sections of this report. The next conclusion focuses on those that would contribute to a process of building consensus on standard measures and appropriate uses for such measures.

Although methodological innovation in population metrics has strengthened the analytical base for health decisions, the lack of accepted standard measures often creates confusion and caution among potential users.

Over the last 30-plus years, researchers have created an array of population metrics that have attracted the attention and, in some cases, the acceptance of policymakers who see these measures as a source of critical information for decisionmaking. At the same time, however, this innovation and diversity create some problems. For example, when the use of different measures produces noncomparable or even conflicting findings, the result may be confusion, distrust of quantitative analysis, and missed opportunities to building cumulative knowledge for decisionmaking. Wariness among users or potential users of summary measures may be a particular peril when decisionmakers realize that they have only a limited grasp of the technical differences among measures and their ethical and policy implications.

The committee does not expect that any single measure will necessarily (1) fit all purposes for summary measures of population health, (2) be without technical or other limitations, and (3) be appropriate as the sole measure for use in any decision. Nonetheless, it would be useful to consider steps that would increase consensus—both nationally and internationally—on which measures to use for which purposes. Some of the components of a consensus-building process include

- identifying parties to be involved such as users and consumers of health data and services as well as developers of tools and policymakers;
- clarifying and attempting to get agreement on standard terminology (i.e., nomenclature and definitions);
- creating mechanisms for process management, consultation, and clarification and for the resolution of disagreements at different stages in the process;
- defining target applications of measures;
- establishing important characteristics of measures for specific applications, including specification of minimum acceptable features insofar as possible;
- systematically evaluating and comparing existing measures along these dimensions; and
- setting priorities for refining measures to minimize technical and ethical problems.

A process of building consensus on standards and measures does not assume that summary measures are necessarily perfectible for all purposes or that policymakers will necessarily accept the value judgments on which they may be based. It does assume that the process will help clarify the imperfections of different measures, encourage resolution of correctable deficiencies in measures, and set guidelines for the appropriate application of such measures for different purposes (notwithstanding their imperfections).

A standard-setting process is not incompatible with continued methodological innovation. Indeed, such a process should take full advantage of the experience of the developers and users of current methods, with an eye toward identifying the strongest features of each measure as a basis for refining existing measures or creating new measures. This suggests the desirability of simultaneous testing of alternative methods to assess their strengths and limitations and to compare the results they generate for policymakers.

The creation of a process for building consensus on population metrics would be consistent with a recent IOM report on community-based performance monitoring. It recommended the development and voluntary use of a set of standard measures for profiling the health of communities as part of a process of community health improvement (IOM, 1997). Canada is already pursuing national consensus on health data needs and measures including both

highly detailed, longitudinal data on health status and health care utilization at the individual level and a family of compatible summary measures of population health (Wolfson, 1997). For this family of measures, a key idea is to construct an overall summary measure in such a way that it can be readily disaggregated to show distributions of health status for population subgroups. Moreover, work is under way on a statistical and analytical framework that provides connections with underlying disease processes (e.g., heart disease), risk factors (e.g., smoking), and social factors (e.g., family income). One objective of this framework is to allow the burdens of disease and the "attributable fractions" of risk factors to be expressed in terms of their incremental contribution to overall population health.

Beyond the issues of "how," "what," and "when" to value, workshop participants noted—from a pragmatic perspective—the utility of collecting population health information in ways that allow easy disaggregation on the basis of social, demographic, and economic risk factors such as age, occupation, or income. Linking information about population health (including both summary statistics and more detailed data about specific health problems and conditions) to information about risk factors can provide important epidemiologic insights for planning public health strategies and shaping health services delivery systems.

Starting points already exist for a process of building consensus on standards for summary measures of population health and agreement on adoption of particular measures for specific purposes. For example, the National Commission on Vital and Health Statistics serves as a forum for a wide variety of interested parties to collaborate in developing common data standards for different information systems and users, and the Organization for Economic Cooperation and Development has taken similar steps.

Building on well-established principles of measurement and analysis and a number of related exercises undertaken by the Institute of Medicine (1992) and the Medical Outcomes Trust (1995), the committee identified several desirable characteristics or attributes for summary measures of population health. These attributes include the following:

- Reliability or reproducibility—a measure is reliable if repeated use under identical circumstances by the same or different users produces the same results.
- Validity—a measure is valid if it measures the properties, qualities, or characteristics it is intended to measure.
- Sensitivity or responsiveness—a measure is sensitive/responsive if it can detect differences or changes in population characteristics that are of interest to users of the measure.
- Acceptability—a measure is acceptable if its intended users (and the constituencies upon which they depend) find the results of its application (e.g., a summary statistic) understandable, credible, and useful for their purposes.
- Feasibility or burdensomeness—a measure is feasible if users can collect the necessary data and perform the required analyses without imposing excessive administrative, economic, or other burdens on those whose participation or cooperation is needed.
- Universality or flexibility—a measure is universal/flexible if it is adaptable to the variability of problems, populations, settings, or purposes that face potential users.
- Documentation—a measure is documented when the methods, criteria, assumptions, and data employed in deriving or calculating the measure are clearly identified and publicly available.

For different purposes, more specific criteria may be needed. For example, the recent Panel on Cost-Effectiveness in Health and Medicine (1996, pp. 120–121) argued for the development of a standard catalog of weights for use in cost-effectiveness analysis. It suggested an approach with these characteristics: "(1) derivation from a theory-based method on which empirical data have been collected; (2) availability of weights from a representative community-based sample of the U.S. population; (3) low burden of administration in clinical and population-based settings; and (4) ability to furnish weights for health states, as well as for illnesses and conditions." The panel went on, however, to acknowledge that none of the systems presented conforms to all of these characteristics and that any specific system may not include information relevant for a particular analysis.

Any process of standard setting and standardization will have ethical and political as well as methodological or technical dimensions. The committee's next conclusion emphasizes the value component of measurement.

All measures of population health involve choices and value judgments in both their construction and their application.

Summary measures of population health have descriptive as well as evaluative uses at both the individual and the population levels. As indicated earlier, the most controversial uses will be those involving resource allocation decisions at the population level. To see why summary measures are controversial, one must examine the way in which value choices enter into both the construction and the application of different measures of population health.

Although the use of measures that adjust years of life by the quality of that life present special ethical issues, even the use of well-accepted mortality measures involves value judgments. In particular, when analysts or policymakers rely entirely on mortality measures in considering alternative health policies, they neglect the disease burden—both physical and mental—of nonfatal disabling conditions. Saving or extending lives is not the only morally important function of health care. Thus, reliance on mortality alone can involve an ethical error of omission.

Summary measures intended to correct this error of omission by combining mortality and morbidity can, however, lead to ethical errors of commission because they require a number of controversial value judgments. As described in the background papers by Daniels and Brock, basic choices include what aspects of health to measure and how to assign weights to different health states.

One central set of weighting choices involves whose evaluations of health states are to be used in these measures. On the one hand, health professionals may have, or may be perceived to have, systematic biases related to their training, social status, and work experiences. On the other hand, members of the general community may have biases related to their experience or lack of experience with illness, injury, or disability. For example, people with a long-standing disability have had time to adapt to their condition and as a result may value these states more favorably than those who have not experienced them. Whose evaluations of disabilities should be used? Different ethnic and other population subgroups may also evaluate health states differently. Careful comparison of different summary measures is needed to determine just how much variation in summary measures is caused by different choices about weighting health states. This will help in developing advice about whether and how to use particular measures.

Workshop participants suggested that for clarity, discussions of the ethical status of summary measures of population health have to distinguish between the preferences (weights assigned to different health states as determined in sample surveys) that are used to construct a measure and the cultural, moral, and other values that guide policymakers in making decisions. Preference weights, of course, involve value judgments, and the methods for incorporating preferences into summary measures are likewise value-laden. These values need to be made explicit.

In addition to choices about assigning weights to different health states, other value judgments also characterize different summary measures. Though most QALY-type measures value a year of healthy life at each age equally, the DALY, as currently constructed, has weighted the value of health at different ages differently, for example, applying lower weights for the very young and the very old. This feature is not an essential element of the general DALY approach, but it is a highly controversial choice that may affect the public acceptability of the measure. Similarly, as Brock notes (1997), value judgments influence choices about which life expectancies to use as a baseline in constructing DALYs because these expectancies differ by gender and other population subgroups, as well as across nations. In any case, the value judgments embedded in the construction of specific summary measures of population health must be made evident if these measures are to be responsibly used or revised for such purposes as resource allocation.

Additional ethical controversies about both the design and the application of summary measures arise when they are employed in cost-effectiveness analyses that are intended to guide resource allocation at the population level. These controversies involve a particularly difficult set of distributive issues that are intrinsic to decisions about resource allocation regardless of the data and measures used to inform the decisions. When should resources be allocated to produce “best outcomes” and when resources should be divided to give people fair chances at some benefit? How much priority should the sickest or worst-off patients have? When should the prospect of modest benefits to many people outweigh the delivery of significant benefits to fewer people?

The straightforward use of cost-effectiveness analysis favors specific, yet contested, answers to these questions (Harris, 1987). That is, it would give no priority to the sickest patients, would permit any aggregation that maximized health benefit per dollar spent, and would always support the best outcomes. The contested ethical assumptions behind this approach to cost-effectiveness analysis are that a unit of a summary measure—be it a QALY or a DALY—has the same moral value wherever it is distributed, that a benefit to one always compensates for a loss to another, and that it is always morally desirable to maximize benefit in the aggregate or at the margin. Thus, for example, a loss of one quality-adjusted life year for a single person can be offset by a gain of a twentieth of a quality-adjusted life year for 20 different people. These assumptions, as Rawls (1971) argues, ignore the “separateness” of persons (i.e., that the losers and the gainers are different people with different experiences not reflected in theoretical assumptions).

Some of these controversies do have implications for the construction—not just the application—of summary measures. One approach to addressing certain controversies about distributional choices is to survey the population to elicit the moral weights people place on resource allocation trade-offs. That is, people are not asked to assess how much different impairments would affect their own quality of life but rather how they would weigh the use of resources to prevent or relieve different impairments for different groups. For example, how would they weigh the prevention of one death against some number of cases of paraplegia prevented? Some propose that these weights could be used to construct an ethically more sensitive summary measure (Nord, 1994).

An alternative view expressed during the December workshop was that summary measures should not attempt to incorporate "equity weights." Rather, such weights, if they were to be developed and accepted, should be applied in a separate, explicit analytic step that makes the distributive analysis and judgments more explicit and visible. In any case, when analysts use an individual's assessments of what risks (e.g., earlier death) he or she would be willing to trade for what benefits (e.g., better quality of life) to measure preferences, they may be inappropriately treating these as statements about acceptable social (rather than personal) trade-offs involving risks and benefits to others.

Yet another perspective is that quality weights should not be used at all in measures and analyses employed in making decisions about different priorities for saving lives (Kamm, 1993). That is, in making decisions, for example, about whether to invest more in treating one disease rather than another, differences in the number of lives saved should be considered but not differences in the quality of life the or level of disability experienced by those whose lives would be saved. This proposal, which has met with mixed reactions, is one way of recognizing the concern in the disability rights movement that policy analyses should not make distinctions among lives that are still worth living.

In his background paper, Daniels suggests that these and other distributive questions form a family of unsolved rationing problems and that persistent controversy involving them may be traced to systematic moral views that differ in their distributive implications. In the face of such divergent views, Daniels claims, it is not possible to elicit meaningful weights from survey techniques, and more attention has to be paid to designing publicly accountable, fair deliberative procedures to resolve disputes about these distributive questions.

In sum, the ethical issues raised by summary measures of population health include both questions about measurement strategies and questions about the principles for rationing or distributing resources. Progress in resolving the former questions does not imply there will be agreement about the latter.

RECOMMENDATIONS

The committee's recommendations involve three basic areas: (1) strengthening understanding of the ethical and policy implications of different measurement strategies and applications; (2) developing agreement on standards for summary measures; and (3) educating policymakers, public health officials, and clinicians about the uses and limitations of summary measures of population health. Although it may seem unusual for work on ethical implications to precede consensus on standards for measures, the committee believes that better ethical understandings should inform standard setting and education. The recommendations for future work related to summary measures of population health focus on broad directions rather than specific technical issues. The objective is to promote the design of summary measures that are understandable and helpful and to encourage the use of these measures in ways that are accountable and credible. The recommendations are directed primarily at the Department of Health and Human Services, the sponsor of this study.

The Department of Health and Human Services should initiate a process of analysis and public discussion to clarify the ethical assumptions and value

judgments embedded in the summary measures and to assess the critical ethical and policy implications of differing designs, implementation approaches, and uses of these measures.

Because controversies about the ethical bases and implications of different measures may so greatly affect the acceptability and application of summary measures in policymaking, those encouraging the use of summary measures and those who would actually use them should have a better understanding of the ethical underpinnings and implications of different measures. Potential users of these measures would then be better informed about the measures' limitations and less anxious that controversial value judgments may lie buried in obscure, complicated statistical methodologies. As a consequence, they would be better prepared to make choices about whether and how to use summary measures.

To support better understanding of summary measures, the committee suggests several lines of research. One would continue deployment and testing of several different measures, at least for monitoring purposes. Here the goal would be to develop a body of empirical work on the distributive implications of different measures. Such research would highlight the extent to which methodological and value choices embedded in the measure shape the picture presented to decisionmakers.

A second line of research would aim to develop more sensitive and sophisticated techniques for examining public attitudes and reasoning processes related to resource allocations and valued health states. Here it would be crucial to dig beneath the initial preferences people might express to uncover the principles and rationales that guide their moral reasoning about these issues. A related area of investigation would involve a combination of empirical and normative work aimed at clarifying issues of public accountability for specific applications of summary measures in decisionmaking.

A third—and more sweeping—line of investigation would involve further philosophical work on distributive issues. It would seek ways to increase public agreement about distributive policies. Given the divisive potential of such policies and the diversity of philosophic and political positions in the United States, the committee recognizes that such agreement may continue to be elusive.

The Department of Health and Human Services should create a process to establish standards for population health metrics and assess the feasibility and practicality of a compatible set of health status measures that could be used for different descriptive and decisionmaking purposes.

Several agencies and units within DHHS would have to cooperate in establishing a standard-setting process and in integrating it with broader international efforts to secure agreement on standard summary measures to use for specific applications. The goals for the standard-setting process should include

- providing a coherent framework for the testing and application of summary measures of population health, including clear terminology and criteria for assessing when measures may not be suitable for specific uses;

• clarifying the values and assumptions underlying different measures and their application for different purposes;

• identifying similarities and difference between existing measures of individual and population health and proposed standard measures;

• coordinating with the development, application, and standard-setting efforts of other national and international bodies; and

• providing a long-term strategy to encourage compatibility between individual health outcome measures and summary measures of population health.

It was beyond the scope of this project to propose specific strategies for setting standards and developing national and international consensus on the suitability of different summary measures for different purposes. Certainly, the National Center for Health Statistics, the Center for Chronic Disease Prevention and Health Promotion, the National Committee on Vital and Health Statistics, the Agency for Health Care Policy and Research, and other parts of DHHS have led or otherwise participated in similar types of activities to develop shared standards for collecting and reporting many kinds of health information.

It is the consensus of the committee that linking public health and medical care measurement strategies would be of fundamental importance in assessing the performance of each system individually and in relation to one another. Accordingly, measurement strategies that afford this flexibility should have higher priority for further investigation and refinement. Other priority issues for exploration concern which domains of health are most important to include within measures designed for multiple uses; what ways of valuing health states are most valid and understandable; to what extent variations in measurement techniques result in markedly different estimates of health status or disease burden; and which measures best permit the identification of links between risk factors and health status.

The committee does not underestimate the difficulties to be faced in reaching agreement on a credible, useful set of health metrics. Some of the differences are political, for example, resistance from those already committed to particular measures. Others are philosophical, for example, disputes over the ethical implications of different measurement strategies and assumptions. Other difficulties are practical or methodological, for example, adopting new or different data collection procedures that impose new administrative costs, complicate comparisons over time, or both. Similarly, efforts to gain agreement on a common summary measure to span all international settings should consider the availability—or nonavailability—of the required data in both more and less developed nations. Given the cost and complexity of establishing reliable, valid data collection systems, it is desirable to make a given data source serve as many uses as possible.

The committee also recognizes that efforts to encourage the adoption of a single summary measure of population health have both pluses and minuses. The pluses are comparability in different analyses and greater simplicity in educating users in the interpretation of the measures. The minuses are the risk of adopting a single measure that is ill-suited for some purposes and, possibly, the premature adoption of a measure without a solid understanding of its practical and ethical implications compared to other measures whose development may be truncated. Thus, the committee agrees with the workshop participants that “side-by-side” comparisons of different measures are essential. for movement toward adoption of a standard, comparable set of measures.

Such comparisons should identify when and why measures produce different depictions of health when applied simultaneously to the same populations using broadly collected data.

> **The Department of Health and Human Services should cooperate with other public and private organizations to educate and train public health, medical, and other relevant professionals to promote their understanding of the interpretation and appropriate use of summary measures of population health.**

This recommendation is directed primarily at public health professionals and clinicians but also be extends to policymakers, health plan managers, and others. Public health professionals have traditionally focused on population health, but clinicians are increasingly participating in managed care plans and large medical groups that are being held responsible for the health of populations and not just for taking care of individual patients. To reach clinicians—and benefit from their special understanding of patients and their health problems—educational efforts must involve the professional organizations to which practitioners look for guidance. Such efforts can build on existing activities such as the medicine and public health initiative in which the American Medical Association and the American Public Health Association are collaborating.

To conclude, the committee believes that summary measures of health status can make important contributions to decisions about improving population health and well-being. To fully and successfully realize these contributions, continued work is needed to refine these measures and to compare the pictures of health status that they generate. It is likewise important to examine—empirically and normatively—how the use of summary measures and the differences among measures can shape, improve, or distort policy decisions. No measure will be perfect or complete for all purposes, so policymakers have to be aware of the limitations of different measures for different uses. They will continue to need many types and levels of information about individual and population health.

Moving toward better summary measures will involve difficult work along many fronts including ethical analysis, measurement of individual and social preferences, and data collection. It will also require funding from national governments and international organizations and cooperation among agencies and investigators who are sometimes competitors. A variety of initiatives, including activities described in this report, provide encouraging evidence of such support and collaboration to improve summary measures of population health.

REFERENCES

Bergner M, Bobbitt RA, Carter WB, et al. The sickness impact profile: development and final revision of a health status measure. *Medical Care* 19:787–805, 1981.

Brock DW. Ethical Issues in the Development of Summary Measures of Population Health Status. Paper prepared for Institute of Medicine Workshop on Summary Measures of Population Health Status, Washington, DC, December 12, 1997.

Brook RH, Ware JE, Davies-Avery A, Stewart AL, Donald CA, Rogers WH, Williams KN, Johnston SA. Overview of adult health status measures fielded in Rand's health insurance study. *Medical Care* 17(Suppl. 7):iii–x, 1–131, 1979.

Brown LD. The national politics of Oregon's rationing plan. *Health Affairs* 10:29–51, 1991.

Daniels N. Distributive Justice and the Use of Summary Measures of Population Health Status. Paper prepared for Institute of Medicine Workshop on Summary Measures of Population Health Status, Washington, DC, December 12, 1997.

Eddy, D. Oregon's methods: did cost-effectiveness analysis fail? *Journal of the American Medical Association* 266:2135-2141, 1991.

EuroQol Group. EuroQol—a new facility for the measurement of health-related quality of life. *Health Policy* 16:199–208, 1990.

Evans, R, et al. *Why Are Some People Healthy and Others Not: Determinants of Population Health.* New York: Aldine Van Guyter, 1994.

Fansel S, Bush JW. A health-status index and its application to health-services outcomes. *Operations Research* 18:1021–1066, 1970.

Fried B. Presentation for the Institute of Medicine Workshop on Summary Measures of Population Health Status, Washington, DC, December 12, 1997.

Fryback DG. Health-Related Quality of Life for Population Health Measures: A Brief Overview of the HALY Family of Measures. Background Paper for Institute of Medicine Workshop on Summary Measures of Population Health Status, Washington, DC, December 12, 1997.

Gunnig-Schepers L. Presentation for the Institute of Medicine Workshop on Summary Measures of Population Health Status, Washington, DC, December 12, 1997.

Hadorn, DC. Setting Health Care priorities in Oregon: Cost-Effectiveness Meets the Rule of Rescue. *Journal of the American Medical Association* 265:2218-2225, 1991.

Harris J. QALYfying the value of life. *Journal of Medical Ethics*. 13:117–123, 1987.

Institute of Medicine. Durch J, Bailey L, Stoto M, Eds. *Improving Health in the Community*. Washington, DC: National Academy Press, 1997.

Institute of Medicine. Field MJ, Lohr K, Eds. *Guidelines for Clinical Practice: From Development to Use*. Washington, DC: National Academy Press, 1995.

Jamison D, Ed. *Investing in Health: World Development Indicators.* World Development Report, The International Bank for Reconstruction and Development. New York: Oxford University Press, 1993.

Kamm FM. *Morality/Mortality*, Vol. 1: *Death and Whom to Save from It*. Oxford: Oxford University Press, 1993.

Kaplan RM, Bush, JW, Berry, CC. The reliability, stability, and generalizability of a health status index. Pp. 704–709 in *Proceedings of the American Statistical Association*, Washington, DC Social Statistics Section, 1978.

Kindig D. *Purchasing Population Health: Paying for Results*. Ann Arbor, MI: University of Michigan Press, 1997.

Klevit HD, Bates AC, Castanares T, Kirk EP, Sipes-Metzler PR, Wopat R. Prioritization of health care services: a progress report by the Oregon Health Services Commission. *Archives of Internal Medicine* 151:912–916, 1991.

Lohr KN, Ed. Advances in health status assessment: conference proceedings. *Medical Care* 27(3 Suppl.):1–293, 1992.

Lohr KN, Ed. Advances in health status assessment. proceedings of a conference. *Medical Care* 27(3, Suppl.):S1-S294, 1989.

Lohr KN, Brook RH, Kamberg CJ, et al. Use of medical care in the Rand health insurance experiment: diagnosis- and service-specific analysis in a randomized controlled trial. *Medical Care* 24(Suppl.):S1–S87, 1986.

Luce BR, Brown R. The use of technology assessment by hospitals, health maintenance organizations, and third party payers in the United States. *International Journal of Technology Assessment in Health Care* 11:79–92, 1995.

Marazzo D. Presentation for the Institute of Medicine Workshop on Summary Measures of Population Health Status, Washington, DC, December 12, 1997.

Medical Outcomes Trust. Instrument review criteria. *Medical Outcomes Trust Bulletin* 3(4):1–4, 1995.

Moriyama IM. Problems in the measurement of health status. Pp. 573–600 in Sheldon EB, Moore WE, Eds, *Indicators of Social Change*. New York: Russell Sage Foundation, 1968.

Murray CJL, Lopez AD, Eds. *The Global Burden of Disease*. Cambridge, MA: Harvard University Press, 1996.

Nord E. The Person Trade-off Approach to Valuing Health Care Programs. Working Paper 38, National Center for Health Program Evaluation, Fairfield Hospital, Victoria, Australia, 1994.

Panel on Cost-Effectiveness in Health and Medicine (PCEHM). Gold, MR, Siegel JE, Russell LB, Weinstein MC, Eds., *Cost-effectiveness in Health and Medicine*. New York: Oxford University Press, 1996.

Patrick DL, Erickson P. *Health Status and Health Policy: Quality of Life in Health Care Evaluation and Resource Allocation*. New York: Oxford University Press, 1993.

Poullier JP. Presentation for the Institute of Medicine Workshop on Summary Measures of Population Health Status, Washington, DC, December 12, 1997.

Rawls J. *A Theory of Justice*. Cambridge, MA: Belknap Press of Harvard University Press, 1971.

Sackett DL, Torrance GW. The utility of different health states as perceived by the general public. *Journal of Chronic Diseases* 31:697–704, 1978.

Sondik E. Presentation for the Institute of Medicine Workshop on Summary Measures of Population Health Status, Washington, DC, December 12, 1997.

Stewart AL, Greenfield S, Hayes RD, Wells K, Rogers WH, Berry SD, McGlynn EA, Ware JE Jr. Functional status and well-being of patients with chronic conditions: results from the Medical Outcomes Study. *Journal of the American Medical Association* 262(7):907–913, 1989.

Sullivan DF. A single index of mortality and morbidity. *HMSHA Health Reports* 86(4):347–355, 1971.

Torrance GW, Thomas WH, Sackett DL. A utility maximization model for evaluation of health care programs. *Health Services Research* 7:118–133, 1972.

U.S. Department of Health and Human Services, Public Health Service. *Healthy People 2000: National Health Promotion and Disease Prevention Objectives*. Washington DC: U.S. Government Printing Office, 1990.

Williams A. The measurement and valuation of health: a chronicle. Discussion Paper 136. York University (Great Britain) Center for Health Economics, 1995.

Wolfson MC. Health-adjusted life expectancy. *Health Reports* (Statistics Canada) 8(1):41–46, 1996.

Wolfson MC. Presentation for the Institute of Medicine Workshop on Summary Measures of Population Health Status, Washington, DC, December 12, 1997.

GLOSSARY

Disability-adjusted life years (DALYs). A summary measure of population health status created for use by the World Health Organization (Murray and Lopez, 1996) to represent the global burden of disease in the form of lost years of healthy life due to either disability or premature death. The measure is constructed using expert-derived weights for the value of additional years of life at each age and weights for more than 100 categories of health deficits resulting from different diseases or injuries.

EQ-5D: A five-dimension measure of health status developed by a consortium of European researchers using a mailed survey to collect information about health and functional states being experienced by individuals (EuroQol Group, 1990). Preference weights have been developed for the various health states described by the EQ-5D, making the measure suitable for use as quality adjustments to compute QALE.

Health-adjusted life expectancy (HALE): A summary measure of population health that integrates morbidity and mortality by summing the products of expected years of life at each age by a numerical weight representing average health status at that age (Fryback, 1997). The units of HALE are health-adjusted life years (HALYs). The only established types of HALE measures use some direct or proxy quality-of-life adjustment that represents people's preferences for different health states (see QALE).

Health-related quality of life (HRQL): A class of measures of an individual's overall health at a particular time that produces a single numerical summary of health-related quality of life on a scale anchored by 0 (state of being dead) and 1 (state of optimal health), with the property that the scale values specifically are constructed to represent degree or strength of preference for one health state over another. These weighting scales can form the morbidity component of a QALE measure.

Health status profile: A multidimensional characterization of an individual's health at a particular time that describes how healthy the person is relative to each dimension (e.g., having excellent perceived health). Examples include the **Sickness Impact Profile** (Bergner et al., 1981) and the **SF-36** (the 36-item short-form of the Medical Outcomes Study survey created from the Rand General Health Survey (Brook et al., 1979). Such profiles are not preference-weighted and do not meet the requirements for health utilities measures but are widely used in observational and experimental studies.

Health Utilities Index (HUI): Three health status indices (HUI:1, HUI:2, HUI:3) each an elaboration of its predecessor, developed by Canadian researchers (Wolfson, 1996). Weights for the HUIs were derived from population surveys using multiattribute utility theory preference elicitation techniques and have been used by researchers at Statistics Canada to compute population estimates of QALE. (The developers termed their resulting summary a "health-adjusted life expectancy," but under the definitions here, the index qualifies as one of the subset of QALE measures.)

Mortality-based population health measure. A measure of population health computed from information on deaths in a population without reference to morbidity. Examples include **mortality rates** (the number of deaths per 100,000 population, usually adjusted to a reference population age distribution) and **population (average) life expectancy** (expected duration of life averaged over all individuals and adjusted to a reference population age distribution).

Quality-adjusted life expectancy (QALE): A category of HALE measures that represents people's preferences for different health states, usually based on one of a specific class of measures of health-related quality of life (see HRQL). The units of a QALE are **quality-adjusted life years (QALYs)**. For some purposes, including cost-effectiveness analysis, measures of health-related quality of life must be based on utilities or preferences for health states that meet the conditions of welfare economics, which assumes that individuals seek to maximize **utilities** (preferences for particular outcomes) and that overall societal welfare is some function of these individual utilities.

Quality of Well-Being (QWB) scale: A well-established measure of health-related quality of life initially developed in the 1960s and 1970s for resource allocation and health planning purposes. It combines 23 general complexes of symptoms and problems with three domains of function (Kaplan et al., 1978). Preference weights for the different health states in this combined classification scheme were derived from a community-based psychometric survey. QWB weights have been used in the health policy literature and in the cost–utility analysis literature as the morbidity component of a QALE computation.

Summary measure of population health: An integrative measure of population health that combines both mortality and morbidity data to represent overall population health in a single number.

Years of healthy life (YHL). A measure created by the U.S. National Center for Health Statistics for use in the U.S. government report *Healthy People 2000*. YHL is computed using health state weights for 35 health status categories (combining self-rated health and level of activities of daily living function) and data from the National Health Interview Survey to adjust tables of U.S. life expectancy. The weights for health states were constructed to approximate health-related quality-of-life weights in the absence of primary data and explicit measures of health-related quality of life. The YHL is sometimes considered a quality-adjusted measure of population health status, but it does not meet the welfare economic assumptions described for health utility measures.

APPENDIX A

Workshop Agenda, Particpants, and Questions for the Working Groups

WORKSHOP ON SUMMARY MEASURES OF POPULATION HEALTH STATUS

AGENDA

Plenary Session
2101 Constitution Avenue, N.W., Washington, D.C

FRIDAY, DECEMBER 12 Lecture Room, National Academy of Sciences

7:30 a.m. Continental Breakfast and Workshop Registration

8:30 **SETTING THE STAGE**

Welcome, Introductions, and Overview
Kenneth Shine, M.D.
President, Institute of Medicine

Harvey Fineberg, M.D., IOM Workshop Chair
Provost, Harvard University

Workshop Overview
Marilyn Field, Ph.D.
Deputy Director, Health Care Services, IOM

Marthe Gold, M.D., M.P.H., Project Consultant
City University of New York Medical School

9:00 **APPLICATIONS OF MEASURES**
Examples of efforts to apply measures to real populations: Objectives, problems encountered, lessons learned, future directions

Michael Wolfson, Ph.D.
Director, General Institutions and Social Statistics Branch
Statistics Canada
Edward Sondik, Ph.D.

Director, U.S. National Center for Health Statistics

Paige Sipes-Metzler, D.P.A.
Manager, Oregon Blue Cross and Blue Shield
Former Executive Director, Oregon Health Commission

Louise Gunning-Schepers, Ph.D.
Professor, Department of Social Medicine, Amsterdam, Netherlands

James Marks, M.D., M.P.H.
Director, U.S. National Center for Chronic Disease Prevention and Health Promotion

Howard Seymour, M.D.
Director, Health Care Development Unit, United Kingdom

11:00 **METHODOLOGICAL ISSUES**
QALYs, DALYs, and other measures: Purposes, concepts, assumptions, data requirements, issues, problems and possible solutions

Dennis Fryback, Ph.D.
Professor of Preventive Medicine, University of Wisconsin

Christopher Murray, Ph.D.
Harvard School of Public Health

1:15 **METHODOLOGICAL QUESTIONS** (continued)

2:00 **ETHICAL ISSUES**
Principles for evaluating measures, preferences and perspectives, role of ethics

Norman Daniels, Ph.D.
Professor of Philosophy, Tufts University

Daniel Brock, Ph.D.
Professor of Philosophy, Brown University

3:50 **PERSPECTIVES OF DECISIONMAKERS AND POTENTIAL USERS**

Bruce Fried, Ph.D., and Jeffrey Kang, Ph.D.
U.S. Health Care Financing Agency

Barbara DeBuono, M.D.

Commissioner of Health, State of New York

Steven Safyer, M.D.
Vice President, Medical Affairs, Montefiore Medical Center

Jean-Pierre Poullier, Ph.D.
Organization for Economic Cooperation and Development

John Eisenberg, M.D.
Administrator, U.S. Agency for Health Care Policy and Research

5:30 **CONCLUDING REMARKS**

Working Groups
2001 Wisconsin Avenue, N.W., Washington, D.C.

SATURDAY, DECEMBER 13

8:00 a.m. Continental Breakfast

8:30 **GENERAL SESSION—CHARGE TO WORKING GROUPS**

9:00 **WORKING GROUPS**

12:00 noon **Lunch in Working Groups**

1:00 p.m. **REPORTS FROM WORKING GROUPS**

4:00 p.m. Adjourn

PARTICIPANTS

Sudhir Anand, M.D.
Acting Director, Harvard School of Public Health
Harvard Center for Population and Development Studies

Linda A. Bailey, J.D., M.H.S.
Senior Advisor for Health Policy
U.S. Department of Health and Human Services

Judith Bale, Ph.D.
Director, Board on International Health
Institute of Medicine

John Bartlett, M.D., M.P.H.
Atlanta, GA

Bobbie Berkowitz, Ph.D., R.N.
Deputy Director, Turning Point Program, School of Public Health and Community Medicine
University of Washington

Stanley Berman, Ph.D.
Economist, National Institutes of Health

Jo Ivey Boufford, M.D.
Dean, Robert F. Wagner Graduate School of Public Services
New York University

Peter Bouxsein, J.D.
Acting Director, Office of Clinical Standards and Quality
Health Care Financing Administration

A. David Brandling-Bennett, MD
Deputy Director, Pan American Health Organization

Dan W. Brock, Ph.D.
University Professor and Director, Center for Biomedical Ethics, Brown University

Claire V. Broome, MD
Deputy Director, Centers for Disease Control and Prevention

Ronald H. Carlson
Director, Office of Planning, Evaluation, and Legislation
Health Resources and Services Administration

Cheryl Austein Casnoff, M.P.H.
Director, Public Health Policy, Office of the Assistant Secretary for Planning and Evaluation

Nelba Chavez, M.D.
Administrator, Substance Abuse and Mental Health Services Administration

Carolyn Clancy, M.D.
Director, Center for Outcomes and Effectiveness Research
Agency for Health Care Policy and Research

Barbara A. DeBuono, M.D., M.P.H.
Commissioner of Health, State of New York

John Eisenberg, M.D.
Administrator, Agency for Health Care Policy and Research
U.S. Department of Health and Human Services

Pennifer Erickson, Ph.D.
Departments of Health and Human Development and Health Evaluation Sciences
Pennsylvania State University

Claude Earl Fox, M.D., M.P.H.
Acting Administrator, Health Resources and Services Administration

Spencer Foreman, M.D.
President, Motefiore Medical Center
New York City

Julio Frenk, M.D., M.P.H., Ph.D.
Executive Vice President, Mexican Health Foundation, MEXICO

Bruce M. Fried, Ph.D.
Director, Center for Health Plans and Providers
Health Care Financing Administration

Kristine Gebbie, R.N., Dr.P.H., F.A.A.N.
Associate Professor, Columbia University School of Nursing

Louise J. Gunning-Schepers, Ph.D.
Department of Social Medicine
Amsterdam Medical Center, THE NETHERLANDS

Jeffrey R. Harris, M.D., M.P.H.
Director, Division of Prevention Research and Analytic Methods
Centers for Disease Control and Prevention

Frances Kamm, Ph.D.
Department of Philosophy
New York University

Jeffrey Kang, M.D.
Chief Medical Officer, Center for Health Plans and Providers
Health Care Financing Administration

Robert Kaplan, Ph.D.
Department of Family and Preventive Medicine
University of California San Diego

Paul Kind
Senior Researcher, University of York
UNITED KINGDOM

David Kindig, M.D. Ph.D.
Professor of Preventive Medicine
Director, WI Network for Health Policy Research
University of Wisconsin-Madison

Richard J. Klein
Chief, Data Monitoring and Analysis Branch
Division of Health Promotion Statistics
National Center for Health Statistics

Jeffrey P. Koplan, M.D.
President, The Prudential Center for Healthcare Research

Samuel P. Korper, Ph.D., M.P.H.
Associate Director of National Intstitute on Aging/NIH
Sr. Advisor, Substance Abuse and Mental Health Services Administration

Donald P. Marazzo, M.D.
Pittsburgh, PA

James S. Marks, M.D., M.P.H.
Director, National Center for Chronic Disease Prevention and Health Promotion
Centers for Disease Control and Prevention

Kristine McCoy
Alexandria, VA

Michael McGinnis, M.D.
Scholar-in-Residence, Commission on Behavioral and Social Sciences and Education
National Research Council

Matthew McKenna, M.D., M.P.H.
National Center for Chronic Disease Prevention and Health Promotion
Centers for Disease Control and Prevention

Paul Menzel, Ph.D.
Provost, Pacific Lutheran University

David Moriarty
Centers for Disease Control and Prevention

Richard H. Morrow, M.D.
Director of Health Systems
Johns Hopkins University

Gregory Pappas, M.D., Ph.D.
Senior Public Health and Population Adviser
Office of International and Refugee Health

Jean-Pierre Poullier, Ph.D.
Office for Economic Co-operation and Development
Office of the Secretariat
FRANCE

Steven M. Safyer, M.D.
Vice President, Medical Affairs and Chief Medical Officer
Montefiore Medical Center

Howard Seymour, M.D.
Director, Health Care Development Unit
UNITED KINGDOM

Joanna Siegel, Sc.D.
Arlington Health Foundation

Clay Simpson, Ph.D.
Deputy Assistant Secretary for Minority Health
Office of Minority Health

Edward J. Sondik, Ph.D.
Director, National Center for Health Statistics

Dixie Snider, M.D., M.P.H.
Associate Director for Science
Centers for Disease Control and Prevention

Susanne A. Stoiber
Deputy Assistant Secretary for Program Systems
U.S. Department of Health and Human Services

Richard Surles, Ph.D.
President, Advanced Clinical Delivery
Merit Behavioral Care

Steven Teutsch, M.D.
Senior Research Scientist, Outcomes Research and Management
Merck and Co., Inc.

Joseph Thompson
National Committee for Quality Assurance

Martin Tobias, M.D.
Ministry of Health
NEW ZEALAND

Reed Tuckson, M.D.
Group Vice President for Professional Standards
American Medical Association

Peter Ubel, MD
Center for Bioethics
University of Pennsylvania

Diane K. Wagener, Ph.D.
Acting Director, Division of Health Promotion Statistics, Office of Analysis, Epidemiology and Health Promotion
National Center for Health Statistics

Daniel Wikler, Ph.D.
Professor, Medical School
University of Wisconsin

Alan Williams, Ph.D.
Professor of Economics
University of York, UNITED KINGDOM

Michael Wolfson, Ph.D.
Director, General Institutions and Social Statistics Branch
Statistics Canada, CANADA

QUESTIONS FOR THE WORKING GROUPS

1. Summary measures of population health status have been devised for various purposes. For the settings listed below, how useful would a summary measure be in providing information or insights not provided by other measures of health? What uses should have the highest priority in future efforts to develop and apply summary measures?

a. For public health surveillance/burden-of-disease monitoring at national/regional levels.
b. For comparing population health between/among countries/states/regions.
c. For resource allocation at national/regional levels.
d. For monitoring and comparing performance in managed care or similar settings.
e. For devising risk-adjusted capitation payments for use with managed care plans or other health care providers paid in whole or in part on a per-person basis.
f. For measuring health-related quality of life in clinical trials and contributing to cost-effectiveness analyses of alternative clinical interventions.
g. Other (discuss).

2. The practicality, credibility, and understandability of summary measures in different settings or for different constituencies is an important issue.

a. What problems of understanding or credibility do you see as deterrents to the use of summary measures in the settings of most concern to you?
b. Would these measures be comprehensible and credible to the constituencies that you serve or for whom you make decisions?
c. What efforts (e.g., education, survey of value systems, media coverage) would contribute to their comprehensibility or credibility?

3. Societal values (e.g., equity) may figure in decision making in different ways. For example, they may be among specific criteria used to evaluate measurement data and assess different options. Alternatively, value judgments can be incorporated in the construction of measures themselves (e.g., by assigning elements of a measure different weights).

a. To what extent would measures be more useful if they took explicit account of distributive/equity issues (e.g., related to disability or chronic disease burdens across different sociodemographic groups based on such variables as age, race, gender, or income)?
b. What social value or ethical issues should be emphasized in future efforts to develop and apply summary measures?

4. Certain summary measures incorporate information about preferences for living with different kinds of disabilities or illnesses.
a. For the uses listed in question 1, how relevant or central to credibility is the inclusion of such preferences into the measures themselves? Are there decision-making contexts where preferences are not necessary?

b. Where preferences are important to incorporate, whose preferences (e.g., patient, community) are most relevant in the different decision-making contexts?

c. What issues related to preferences should be priorities for attention in future work to develop and apply summary measures?

How important is it that a single summary measure be suited to a range of different uses?

a. What different goals or aspects of health might warrant different measures?

b. What are the trade-offs to consider in using single versus multiple measures?

APPENDIX B

Overview: Workshop on Summary Measures of Population Health Status

Marthe Gold, M.D., M.P.H.
City University of New York Medical School

INTRODUCTION

Summary measures of health include an array of descriptors that are intended to create an understanding of the well-being of populations. The groups they describe may range from patients in clinical trials to representative samples of communities or nations. The descriptors can be as rough as mortality rates, or may be more finely tuned, describing most aspects of commonly shared understandings of the components of health. For example, there is wide consensus that physical functioning, mental and emotional well-being, social and role functioning, general health perceptions, pain, energy, and vitality as a core set of concepts are central in conceptualizing health.[1] At this meeting, presentations and discussions will focus on the types of summary indicators that are variously referred to as "health status," "health-related quality of life," and "functional status" measures. They share the common purpose of informing a wide range of decision making in the public health and medical systems both in the United States and abroad.

More specifically, it has been proposed that health status measures are suited to measuring the efficiency or effectiveness of medical interventions, assessing quality of care, estimating the needs of populations, improving clinical decisions, and understanding the causes and consequences of differences in health.[2] Framing their uses in a slightly different way, Bergner and Rothman[3] have suggested that health status assessment measures serve four different functions, including examination of the health of general populations, clinical interventions and their effects, changes in the health care delivery system, and health promotion activities and their effects.

In addition to capturing a picture of health at a point in time, some summary measures are designed to allow the incorporation of information about projected life expectancy, thereby allowing estimates of both the quality and the quantity of life that is associated with health care interventions. Health-Adjusted Life Expectancies (HALEs), Quality-Adjusted Life Years (QALYs), Years of Healthy Life (YHLs), and Disability-Adjusted Life Years (DALYs) integrate health status with survival. All of these measures attach a single number, ranging from 0 (death) to 1.0 (optimal health), to a complex of social and personal attributes that represent health status. That number is then linked to life expectancy and combined into a one-dimensional measure of overall health. Although they have been generated for use in slightly different contexts, (for example, the QALY has been used primarily to study the cost-effectiveness of medical

treatments,[4] the DALY to measure the burden of disease worldwide,[5] and the YHL to track the health of the U.S. population,[6] in measuring the aggregate health of defined populations, HALEs, QALYs, DALYs, and YHLs are all suited to making comparative assessments of the health needs of populations. In addition, they share the potential for evaluating the effects and performance of different types of health programs.

As we shall hear in more detail during this meeting, there is growing interest at international, federal, and state levels in using these measures to guide prioritization of health care investments. Within the U.S. health care policy arena, QALYs were used in the Oregon Medicaid Demonstration project to develop a list of priority services for which full coverage would be available to all Oregonians.[7] DALYs were initially applied in the context of developing priorities for resource investments in health within the developing world.[8] YHLs, HALEs, and a variation on DALYs have been used, respectively, at the federal level within the United States, in Canada, and in the Netherlands to describe the overall health of nations. The Health Care Financing Administration (HCFA) of the U.S. Department of Health and Human Services (DHHS) is exploring the longitudinal use of one health status measure, the SF36,[9] to track outcomes at the clinical care level. Although the SF 36 is not designed to be combined with survival information, the intent of the tracking parallels other efforts being conducted using combined measures.

There are a number of arenas where summary measures are being used to assess the health of populations. In general, however, these metrics are not in wide use at the policy level. Despite a 25-year history of development and fine-tuning, summary measures are most fre-quently used and cited by the research community that has created them. Discussion of the merits and demerits of summary measures has been conducted primarily among philosophers and meth-odologists. Application at the delivery and program evaluation side of health care has lagged substantially behind the general public health tracking function.

The reasons for the disconnect between the measurement and the policy communities are many. Most obviously, busy policy makers often are not aware of these techniques. A wider effort to inform decision makers about the range of uses of summary measures is required and we hope that this meeting and the committee's report will contribute to this effort and help familiar-ize different interests with the potential opportunities created by the measures. It is also true, however, that many decision makers are at least glancingly aware of the opportunities that summary measures create in informing policy, but are reluctant to use them because methodological and ethical concerns they feel have been inadequately addressed to date.

For example, on the methods side, summary measures capture a continuum of components of health which may not have credibility to all constituencies. Some measures harvest information directly from patients regarding their mental health, symptoms, and physical, social, and role function; they are designed to provide information at the clinical care level. Others capture the judgments of health experts about the health states associated with particular diseases or conditions with the goal of capturing snapshots of disease burden in large populations. This type of summary measure may not always capture what is most salient to decision makers and their constituencies as they consider priorities for resource allocation.

Other concerns may arise from the differential sensitivities that measures have in recording decrements in particular aspects of health. For example, a measure that includes information only on pain and physical function may paint a far harsher picture of arthritis than of some types of cancer, thereby giving pause to those who see this as a type of bias built into the measurement strategy. It is well known that QALYs are constructed from many different kinds of summary

measures of health-related quality of life.[10] These measures comprise different domains of health and use different scoring strategies to value the health states, resulting in wide variations in scores for similar conditions. Decision makers may be legitimately concerned about the results from different evaluations using measures that do not readily translate to one another.

On the ethics side, decision makers may be aware of, and not fully comfortable with, the utilitarian," "greatest good for the greatest number" perspective that is fundamental to the assessment of gain used in the current QALY model. It is the sum of the QALYs produced by an intervention that measures the desirability of the intervention; it does not matter to whom the QALYs accrue. Life-saving programs or therapies for persons with relatively low health-related quality of life, or for older persons, may produce far fewer QALYs than would interventions that are primarily health status improving. Therapies with a small health status enhancing impact on large numbers of persons, for example a cure for the common cold, can confer far more QALYs than those that save the lives of a few people with a rare disease. Concerns about how one accounts for distributive justice—helping those in most need first, or placing priority on life-saving rather than quality-enhancing procedures—and whether these measures are in fact discriminatory, are likely areas of discomfort within the policy community. Finally, many of the measures rely on assessments of peoples' values for differing levels of health. There is a good deal of controversy about whose values are most appropriate to capture in judgments about health states. For example, there are known differences between the values people who have experience with a disease bring to the assessment of that disease compared to a general population. In addition, there are gaps in information as to what impact sociodemographic variables such as social class, culture and ethnicity, and age have on assessments of health status.

The methodological and ethical issues related to the use of summary measures may be more or less important or controversial depending on the function a measure is intended to serve. For example, the Years of Healthy Life measure was intended to track trends in the U.S. population for purposes of providing an overview of the nation's health. As such, it parallels economic indicators such as the Gross Domestic Product (GDP) which provides a snapshot of economic functioning without implying a particular direction for action. Snapshot indicators increase or decrease or remain the same; this overview can provide useful information to a nation or a region about trends in health. Although analysts can attempt to identify the specific variables responsible for changes in the summary indicators, the indicators themselves do not suggest specific policies or actions. On the other hand, measures intended to capture the impact and performance of particular interventions (e.g., coronary artery bypass grafting), or particular health systems, (e.g., a managed care plan), would reasonably be held to higher levels of performance. The outcomes charted by these measures might suggest to planners and decision makers a need to expand or abolish access to therapies or programs which have importance to different constituencies.

WORKSHOP STRUCTURE

This workshop will focus on summary measures of health status and health-related qual-ity of life that have been created to serve a number of informational and decision making needs in both medicine and public health. The extent to which these measures accomplish performance and tracking objectives in health and the areas where they need further refinement are important issues for health care decision makers. Workshop participants are methodologists, ethicists, and decision

makers from the public and private sectors, in the United States and abroad. The public health and the medical care delivery systems are both represented.

Day 1

Members of the first panel will focus on applications of these measures in policy-relevant settings. Presenters will describe the nature of their evaluation requirements and the measurement strategy they have used to accomplish them. Work in tracking the health of the U.S. population will be presented by Ed Sondik, Director of the National Center for Health Statistics. Louise Gunnig, from the Netherlands, and Michael Wolfson, from Statistics Canada, will describe their respective efforts in developing these measures to track and prioritize population health in their countries. Paige Sipes Metzler, who worked with the Oregon Health Services Commission, will describe use of QALYs in state-level decision making on the components of a basic health care benefit package. Jim Mark, Director of the Center for Chronic Diseases at the Centers for Disease Control and Prevention, will discuss CDC's initiative to assess burden of disease in a manner that creates opportunities for exploring prevention priorities for the United States. Howard Seymour will provide an account of the use of these measures at a local health planning level in the United Kingdom.

The next series of presentations will look more closely at the methods used to build these measures. The intent here is to provide all workshop attendees with a working understanding of the different aspects of health and disability that are captured by these measures, the manner in which these aspects are aggregated (or not) to provide a composite number that represents a state of health, and the different techniques for integrating the health status numbers with life expectancy. Dennis Fryback will provide an overview of health status measurement, describing differences between measures that incorporate peoples' values for health states and those that do not and how different measures can be used to fashion health-adjusted and quality-adjusted life expectancies. Chris Murray will describe a particular subset of quality-adjusted life expectancy, the Disability-Adjusted Life Year, commenting on its intended uses and data requirements.

These papers will be followed by presentations from Dan Brock and Norman Daniels who will focus on a number of the ethical issues that have been raised by the use of QALYs and DALYs. Dan Brock's comments will be directed toward exploring some of the ethical assumptions that are contained within the actual methodology. Norman Daniels will explore the issues of distributive justice that arise within resource allocation decision making, and the degree to which these measures both solve and create problems for decision makers.

Central objectives of this conference are to gather careful input from the policy community regarding directions for further developmental work in methods and ethics, and to explore the potential for broader use of summary measures in decision making. A series of questions, noted below, have been developed to focus small group discussions on the second day of the conference. The final session of the first day of the workshop will include comments from decision makers who have begun to examine how they can best use summary measures in their work settings. They will reflect on a number of the questions in the context of their own planning and policy needs. Bruce Fried and Jeff Kang from the Center for Health Plans and Providers will discuss HCFA's interest in, and use of, the Health of Seniors' measure for plan accountability with regard to health outcomes. Barbara DeBuono, Commissioner of Health for New York State,

will comment on the relevance of these measures at the state level, in an array of programs ranging from managed care to public health. Jean-Pierre Poullier, from the Organization for Economic Cooperation and Development, will describe international discussions under way to achieve consensus concerning the use of summary measures for shared economic evaluations. Stephen Safyer will reflect on how a large, multifaceted health care delivery system providing a continuum of care from primary care to tertiary care, to rehab and nursing home, views the use of these measures. Finally, John Eisenberg, director of the Agency for Health Care Policy and Research (AHCPR) in the U.S. Department of Health and Human Services, will describe research and applied uses for these measures including cost-effectiveness analyses of new technologies and therapies and evaluation of plan performance and potential for use in risk-adjustment of premiums for enrollees in managed care organizations.

Day 2

The second day of the meeting will involve deliberations in small, pre-assigned working groups. Group assignments reflect the experiences and needs of conference participants in the use of these measures in diverse decision-making contexts. In addition, to technically clarify some of the issues under discussion, each group has representation from the methodology and philosophy communities. A facilitator and a rapporteur, who is a member of the IOM committee, are assigned to each of the working groups. The rapporteur will bring the groups' discussions back to the larger meeting, and to the deliberations of the committee.

The questions below will form the nidus for the discussion in the morning small group sessions. We hope that early review of and reflection on these questions both prior to the meeting and during the presentations of the first day will aid members of the working groups in providing guidance to the committee in a number of different areas including direction for:

- development of the methodology,
- further exploration of ethical issues, and
- possible applications of summary measures.

On the afternoon of the second day, the full group will reconvene for presentation and discussion of the summarized deliberations of each work group.

REFERENCES

1. Lohr, KN. Applications of Health Status Assessment Measures in Clinical Practice: Overview of the Third Conference on Advances in Health Status Assessment. *Medical Care* 30:MS1–MS14; 1992.
2. Ware, JE, Brook, RH, Davies, AR, and Lohr, KN. Choosing measures of Health Status for Individuals in General Populations. *American Journal of Public Health* 71:620–625, 1981.
3. Bergner, M, and Rothman, ML. Health Status Measures: An Overview and Guide for Selection. *Annual Review of Public Health* 8:191–210, 1987.
4. Weinstein, MC, and Stason, WB. Foundations of Cost-Effectiveness Analysis for Health and Medical Practices. *New England Journal of Medicine* 296:716–721, 1977.
5. Murray, CJL. Quantifying the Burden of Disease: The Technical Basis for Disability-Adjusted Life Years. *Bulletin of the World Health Organization* 72:429–445, 1994.
6. U.S. Department of Health and Human Services. *Healthy People 2000: National Health Promotion and Disease Prevention Objectives*. Washington D.C.: Government Printing Office, 1990.
7. Klevit, HD, Bates, AC, Castanares, MD, Kirk, EP, Sipes-Metzler, PR, and Wopat R. Prioritization of Health Care Services: A Progress Report by the Oregon Health Services Commission. *Archives of Internal Medicine* 151:912–916, 1991.
8. World Bank. *World Development Report: Investing in Health*. New York: Oxford University Press, 1993.
9. Ware, JE, and Sherbourne, DC. The MOS 36-Item Short-Form Health Survey. *Medical Care* 30:473–483, 1992.
10. Gold, MR, Siegel, JE, Russell, LB, and Weinstein, MC, eds. *Cost-effectiveness in Health and Medicine.* New York. Oxford University Press, 1996.

APPENDIX C

Methodological Issues in Measuring Health Status and Health-Related Quality of Life for Population Health Measures: A Brief Overview of the "HALY" Family of Measures

Dennis G. Fryback, Ph.D.
University of Wisconsin-Madison

INTRODUCTION

A "Simple" Concept

There is little dispute that the well-being of individuals involves two main conceptual parts, one dealing with longevity (or threats to longevity), and the other dealing with morbidity or the nonlethal aspects of daily function and pain and suffering that affect peoples' lives. Any single summary measure of health and well-being of individuals and of populations will need to account for both these aspects.

This overview deals with a seemingly simple and powerful idea about numerical representation of health of individuals and its extrapolation to populations. Suppose we have a numerical measure of degree of morbidity (or lack of it) experienced by an individual at any given time. This numerical measure can be used to weight each passing moment of an individual's life, and when the weighted moments are cumulated across the person's life, the numerical aggregate denominated in "health-adjusted life years" (HALYs) can be used as a summary statistic accounting for both longevity and degree of morbidity experienced.

In retrospect, the HALYs accrued by an individual can be one numerical summary of that person's lifetime health experience. This summary is the conceptual generalization of just counting the number of years an individual lived; we can say, for example, that a person lived to the age of 83 years (accumulated 83 life years) or we can say that individual accumulated 72.5 HALYs. Applied prospectively, the health-adjusted life expectancy (HALE) might be computed to index an individual's lifetime and health prospects. For example, the estimated life expectancy of a 50-year-old male in the United States in 1981 was 25.0 years; in 1993 the estimated life ex-pectancy of a 50-year-old male in the United States was 29.2 years, and the increase of 4.2 years in life expectancy over the 12-year period gives us information about changes in one of the two aspects of health and well-being in the United States during that interval. If we could prospec-tively weight life years with expected degree of morbidity during those years, we might be able to report the equivalent health-adjusted life expectancies for 50-year-old males. Of course, we may wish to

compute these summary numbers for any subgroup of the population, or for a population as a whole.

This, then, is the simple concept: if we have suitable measures of morbidity experience and of life expectancy, we can compute a health-weighted life-year measure—HALYs, or HALE—to summarize population health. Although this seems a simple idea, it can be complex in application. This paper presents the major methodological issues surrounding construction of such measures.

I use the terms HALYs and HALE to refer generically to summary numerical representations that are an accumulation of health-weighted life years. In the course of this overview subclasses of HALYs—e.g., quality-adjusted life years (QALYs)—are discussed as well as some differentiations within subclasses.

Brief Historical Background: Descriptive and Decision Making Roots of HALYs

The groundwork for the population-based HALY representation[a] was laid in the 1960s and early 1970s in publications from the U.S. Department of Health, Education, and Welfare, in the operations research literature reporting work funded by HEW. The problem was to describe the overall health of the population. For years this had been indicated by mortality rates, the decline in which began to level off some in the 1950s and 1960s. During the 1960s and 1970s researchers were looking for methods to bring into the descriptive summaries information about morbidity as well as mortality.

The earliest paper I have seen is by Sanders, who introduced the mathematical combination of a measure of functional capacity and a measure of time to make a combined measure of "effective life years."[1] Sullivan[2] amplified on this concept as did Moriyama,[3] and later Sullivan used a stationary life table technique to compute age-specific disability-free life expectancy from census data about mortality and the National Health Interview Survey (NHIS) data about disability.[4] Sullivan multiplied the number of people in each age range's stationary population (from the census' life table) by the proportion of the population that was disability-free in that age range in the NHIS data. The averaged result was mathematically an early progenitor of health-adjusted life expectancy, where the health weighting was a 0 for health states involving disability and 1 otherwise. This same approach has been used more recently but with different continuous measures of overall health to compute "Years of Healthy Life" (YHL) for the U.S. population[5] and HALE for Canadians.[6]

As a descriptive statistic the HALE for a population draws meaning from its broad conceptual foundations and experience we gain with it over time. A good descriptive indicator should behave about like one expects it to given the conceptual grounds it springs from, and with experience over time we will learn its quirks and associate its behavior retrospectively and prospectively with other indicators of interest. This is not unlike the fashion in which some summary economic indicators such as Gross National Product (now evolved to Gross Domestic Product), or various stock market indexes around the world have acquired meaning.[b]

As an input for decision making, however, a HALY measure may have more demands placed on it. In particular, one type of HALY measure called the "quality-adjusted life year" (QALY) has a history associated with welfare economics and a principle for decision making under uncertainty termed "maximizing expected utility" or the EU principle. This line of reasoning begins by assuming that individuals have well-defined preferences; these preferences are termed

"utilities" and individuals are presumed to make decisions so as to maximize their overall expected utility. Welfare economics concerned with how to devise a social utility function to guide social decision making so that aggregate social outcomes are related in desirable ways to the preferences of the in-dividuals making up the society. This background is discussed at length in a book on cost-effectiveness analysis in health and medicine commissioned by the U.S. Public Health Service.[9]

While the HALE concept was being developed as a descriptive index for health gains in populations, economists, operations researchers, and psychologists were developing the QALY concept for making decisions about which medical treatment or which health system intervention should be undertaken by virtue of its being the most cost-effective.[10,11] The seminal work in this regard is by Fanshel and Bush,[12] who developed a preference-based measure for measuring health in populations, tied it to the foundations of EU decision making for public policy, and then demonstrated its application.[c] Less than a decade later, the QALY as an outcome measure for cost-effectiveness analysis appeared in the major medical literature.[13]

Technically, as a measure of decision outcomes for expected utility decision making, the QALY measure must satisfy some stringent mathematical and psychological properties, as dis-cussed in following sections. These properties were first analyzed and discussed in the operations research literature,[14] and very recently reanalyzed with surprisingly simplified result.[15] Although the use of QALYs for EU decision making and cost-effectiveness analysis has been predominant in the medical decision making literature, EU decision making is by no means the only theory for how to make decisions. The study of the mathematical properties of QALY-like measures for other de-cision theories has been given modest attention[16] probably because non-EU theories of decision making, while possibly descriptive of how people actually do make decisions, are rarely advocated as theories of how decisions under uncertainty should be made.

COMPONENTS, CONSTRUCTION, AND PROPERTIES OF HALY MEASURES

Viewpoints

Confusion about what is and is not required of a HALY measure often stems from the use of such measures for individual decision making versus use for societal decisions. In the dis-cussion following, it is important to distinguish the use of a HALY measure to assist in decisions about health and medical care by individuals versus decisions made in a social policy setting.

In constructing measures to assist individuals, we must take into account the psychology of how individuals make decisions in complex situations of uncertainty. Construction and use of HALY measures in individual decision making is not the subject of this paper.

Societal decision making applications may be more deliberative, and we may get a chance to ask how we (as society) wish to make the decisions. Summary measures of population health may be only one of several decision considerations. Although a social measure of population health needs to deal with being representative of how individuals in the society value health, it also needs to deal with the problem of how health may be distributed in a population as a consequence of decisions. And, the societal viewpoint must also deal with problems of data availability on a population scale.

In the discussion following, I assume that we are intending the HALY measure to aid us to describe the health of populations and to describe the aggregate health consequences of social decisions that affect health of populations, for the purpose of policy-level decision making.

Separation of "H" from "LY"

A health state is defined as "the health of an individual at any particular point in time (p. 399)."[17] Gold and colleagues[18] characterized lifetime health paths as lifetime movement through different health states:

> In the most general case, each individual is born and lives out a lifetime that consists of moving through different health states over time until death. Each individual has a different path through these health states that terminates at a different time of death. For example, consider a perinatal intervention designed to improve the health of a newborn. Without the intervention the average newborn faces a probability distribution of possible paths through life. With the intervention, the person faces a different, hopefully improved, probability distribution of paths. . . .
>
> Two special paths represent the extremes of possible health outcomes and they bound the range of effectiveness scores for possible health paths. One is the path that consists of immediate death at birth. The other is the path that consists of 'optimal' health for a 'full' lifetime. [pp. 89–90]

The term "health outcome" is used throughout the literature; unfortunately it is used in two different ways, one referring to what is denoted here as a "health state" and the other to a "health path." Readers often have to determine from context which of these usages is meant.

It is possible, of course, to speak also of conditional lifetime health paths such as age-specific and condition-specific health paths that are the health paths starting from a particular age with a particular health condition. The problem of measurement of health paths is the problem of assigning a number to each possible lifetime health path (or conditional health path) so that their numerical ordering represents better or worse health paths, and so that we can do meaningful arithmetic (such as taking averages) and the numerical results (e.g., averages) also are meaningful in the same way in that the numerical order of the averages represent better or worse aggregated or prospective health paths.

The most fundamental assumption in the construction of HALY measures is that the part of the measure dealing with weighting health states can be obtained separately from the life-years, or time duration part of the measure. This assumption implies that the relative degree of health represented by a particular health state can be rated separately from knowing the duration of that health state.

There are data showing this assumption is not uniformly true about how individuals think about their health. The relative weight that individuals give to some health states has been shown to vary with how long those states are endured.[19] There are other data showing for other health states such as angina the assumption is approximately true.[20] Some acute states may be relatively more tolerable than if they were chronic states. Alternatively, individuals may accommodate to a

health state over time, thus giving higher weights to certain conditions endured for a long time versus a brief time. At the level of individuals' ratings of their health this is a complicated issue.

This assumption is made for population-based measures largely for practical reasons. It immensely simplifies data collection and computations of average HALYs. It means we can collect data about relative weighting of health states in surveys that are separated from collecting data about longevity and later put these two sets of data together. And it simplifies understanding and communicating the aggregated HALY measure. There are ways to relax this assumption to a degree, for example by constructing health state descriptions that include time in the state for states where duration of the health state is known, to systematically and seriously affect the weight given to the state.

Decision makers guiding construction of HALY measures for population-focused decision making will need to weigh simplicity against fidelity of the measure with respect to the assumed degree of separability of health state from duration. "High-fidelity" health state description systems, able to differentiate thousands or millions of health states that might affect humans, may need to be combined with time measures as if they were fully separable. High-fidelity HALY measures, not relying on the assumption of separability, may well be too complex to allow data collection at a population level.

For the subsequent sections of this paper, it will be assumed that the H part of the measure is deemed to be separable from the LY part, and we will focus on different systems for measuring H.

MEASURING "H"

Systems for Measuring Health Status and Health-Related Quality of Life

The problem of deriving weights for health states is generally divided into three exercises which are discussed below:

- First, we must decide what aspects of health will be the foundation of the classification scheme—i.e., what are the important aspects of health we wish to enumerate as health states.
- Next, we must devise a system by which a real individual's health can be mapped into our discrete set of health states—i.e., we must operationalize the classification system.
- Third, we must devise a system for assigning numbers to each of the health states so that these numbers can be used as weights in a HALY computation.

In the discussion following, I use six different systems as examples of health measures:

- QWB: The Quality of Well-Being Scale[21] is the direct descendant of the early work by Fanshel and Bush cited earlier. The QWB has been used as a general health measure in many clinical studies and in policy research.
- HUI: Torrance and colleagues[22,23] created the Health Utilitities Index from work conducted only slightly later than the Fanshel and Bush scale. This index has gone through several augmentations over the years; the successive indexes are the HUI-Mark I, HUI-Mark II, and

HUI-Mark III. The HUI is used for population health indexing in the province of Ontario and preliminarily throughout Canada by Statistics Canada.

• YHL: Years of Healthy Life is a measure of population health computed experimentally by the U.S. National Center for Health Statistics from data collected in the National Health Interview Survey.[5] I will use "YHL" to refer elliptically to the health state weighting component of this measure.

• DALY: Disability-Adjusted Life Years is a measure created by Murray and colleagues to index global burden of disease for the World Health Organization.[24,25] For purposes of this section I will use DALY to refer indirectly to the disability state weights that are one component of the DALY measure.

• EQ-5D: The EuroQol collaboration is a European multinational collaboration on measuring health-related quality of life in Europe.[26] The EQ-5D is the main health state data collection instrument used in this collaboration.

• SF-36: The Medical Outcomes Study Short Form-36 (known as the "SF-36")[27] is a descendant of the Rand General Health Survey questionnaire. It is a health profile that is widely used in the United States and other countries. It is being tested as an instrument for monitoring population health change in the Medicare population (the "Seniors Project").

While these six do not exhaust the possibilities appearing in the literature (e.g., see McDowell and Newell[28]) for a survey of many different health state description systems; McHorney[29] reviewed this literature; and Gold et al.[18] discussed systems that might be used in cost-effectiveness analyses), a more comprehensive review is beyond the scope of this overview.

Descriptive systems for health states. The universe of health states that humans experience is immense. The problem faced by any classification system is to represent the complex reality of this universe in a manageable fashion. The simplest of classifications is to classify all health into two states: "alive" and "dead." Policy decisions based only on mortality rates and on life expectancy ultimately rest on this simple binary classification without further differentiation. Once we decide that there are subclasses in the "alive" state that we wish to differentiate, the trick is how to do this in a meaningful way without being overwhelmed by complexity.

What constitutes "health"? Any specific answer to this question is bound to be controversial. But examination of different classification systems that have been proposed results in a list of concepts and concerns that seems to encompass most. These are shown in Table 1.

Constructs as listed in the left-hand column of Table 1 are referred to by various terms used approximately synonymously: "concepts," "constructs," "domains," "dimensions," "attributes," and "factors." The different aspects of health are not assumed to be independent in a statistical sense; for instance, the occurrence of depression may well be influenced by presence or absence of acute physical function restrictions or pain. Statistical independence of health dimensions is not a necessary condition of health state measurement systems. Most measurement schemes do require another sort of independence: that the weight accorded to different levels of one of these dimensions does not depend on other dimensions or its particular etiology. I return to this in the next section.

TABLE 1 Typical Concepts and Concerns Used in Measures for Health States

Concepts and Domains	Indicators
Health perceptions	Self-rating of health; health concern, health worry
Social Function	
Social relations	Interaction with others; participation in the community
Usual social role	Acute or chronic limitations in usual social role (major activities) of child, student, worker
Intimacy/sexual function	Perceived feelings of closeness; sexual activity and/or problems
Communication/speech	Acute or chronic limitations in communication/speech
Psychological Function	
Cognitive function	Alertness; disorientation; problems in reasoning
Emotional function	Psychological attitudes and behaviors
Mood/feelings	Anxiety; depression; happiness; worries
Physical Function	
Mobility	Acute or chronic reduction in mobility
Physical activity	Acute or chronic reduction in physical activity
Self-care	Acute or chronic reduction in self-care
Impairment	
Sensory function/loss	Vision; hearing
Symptoms/impairments	Reports of physical and psychological symptoms, sensations, pain, health problems or feelings not directly observable; or observable evidence of defect or abnormality

SOURCE: This table appears in Gold et al. (1996) as adapted from Patrick and Erickson (1993).

Figure 1 provides a representation of the concepts in Table 1. I have placed the constructs in a hierarchy. Each branching in the hierarchy can be construed as a parsing of the construct from which the branches emanate into subconstructs. It is not always easy to tell whether two different measures include the same constructs or dimensions of health. Not all concepts in Table 1 or Figure 1 appear explicitly in every measure of health. Alternatively, the same concept appearing in two different measurement systems may be operationally specified in quite different ways. Finally, health measures may differ in how they parse a construct at any given level in a conceptual hierarchy of health, and in how far down the tree they will continue parsing constructs into subconstructs.

For the most part, these domains cover the descriptive schemes that attempt to generalize across health experiences. Measures that deal only with manifestations of a particular disease or condition may be much more detailed in the aspects of health affected by that disease to the exclusion of aspects not generally affected (for example, measures intended only to describe health states of persons with arthritis may be very detailed about pain and physical function, but generally do not include dimensions dealing with visual ability).

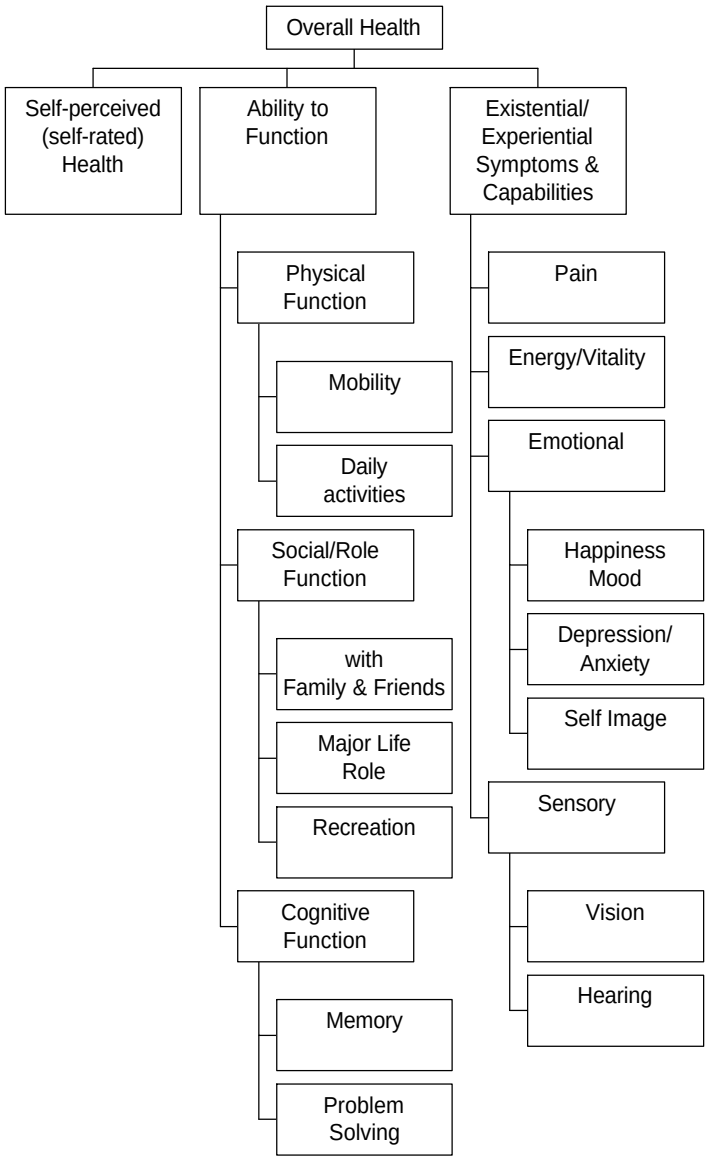


Figure 1 A possible hierarchical view of health dimensions.

For purposes of public policy, where decisions may well be made that involve many different diseases and conditions, it is generally agreed we need a measurement scheme that covers the wide front of human experience and not just one disease or one aspect of health. But common health indexes devised to be "generic" in fact differ in which aspects are included and which are not. A major point of difference comes at the top level of the hierarchy in Figure 1. Some measurement systems include self-perceived health, and some do not. The Years of Healthy Life (YHL) measure from the U.S. National Center for Health Statistics[5] and the Medical Outcomes

Study Short Form (SF-36)[27] both include self-perceived health. The Quality of Well-Being Scale (QWB),[21] the Health Utilities Index (HUI),[22,23] and the scale of disability used in the Disability-Adjusted Life Years (DALY) measure,[24,25] and the EQ-5D (the instrument being used by a multi-national collaboration on measuring health-related quality of life in Europe)[26] do not use this dimension as part of the classification scheme.

A second major difference is inclusion/exclusion of the dimension I have labeled "existential/experiential symptoms and abilities." Some developers of health measures have taken the view that the attributes down this branch are important aspects of health, but only insofar as they affect functioning. In this view, these aspects of health are valued for how they do or do not limit ability to conduct oneself in the broad world physically and socially. The QWB gives some of these weight as symptoms or problems in its "symptom/problem complex" attribute. The HUI, the EQ-5D, and the SF-36 detail degree of functioning on a number of these. However in DALYs and in the YHL these are not measured except indirectly through how they affect functioning in various activities.

Operationalizing the classification system. There are methodological differences associated with the manner in which health classification systems are operationalized. The problem of operationalization is that of "mapping" the health state of a given individual into the set of classes that is created by a measurement system. It is perhaps easiest to describe these by example.

Many classification systems are operationalized according to attribute. The system developers have listed the dimensions to be included in the measure (dimensions such as those in Table 1 and Figure 1), and have worked out a system for locating where along each dimension a given individual resides whose health is to be rated. Each dimension is categorized into a discrete set of levels, ranging from best function or best health on that dimension to worst. Table 2 shows attribute names and number of categories created for each attribute for the four systems here that are operationalized in this fashion.

The number of different health states that potentially can be distinguished by a classification system is the product of the number of levels of each dimension since a health state is defined by picking one level from each dimension. So the QWB can distinguish $3 \times 3 \times 5 \times 26 = 1{,}170$ different health states (plus the state "dead," for a total of 1,171 states). The HUI-Mark III can distinguish up to 972,000 health states. The EQ-5D can distinguish 243. The YHL measure classifies 30 health states. And the disability classification for DALYs has 7 classes (including no disability). In fact, many potential combinations of attributes in the first three systems are probably impossible combinations (e.g., "in a coma" and yet driving around the community!), so these simple computations represent upper bounds. And by no means are the various combinations considered to be equally likely—so much of the population will probably congregate in a relatively few cells of the classification schemes having many categories. The sheer number of health states does not indicate good or bad on the part of a measure; it is only an indicator of potential for differentiating states of health.

TABLE 2 The Attributes and Their Number of Levels for Four Health Measurement Systems

System	Attribute Name	Number of Levels
Quality of Well-Being Scale	Mobility Scale	3
	Physical Activity Scale	3
	Social Activity Scale	5
	Symptom/Problem complex	26
Health Utility Index-Mark III	Vision	6
	Hearing	6
	Speech	5
	Ambulation	6
	Dexterity	6
	Emotion	5
	Cognition	6
	Pain	5
EQ-5D (EuroQol)	Mobility	3
	Self-Care	3
	Usual Activities	3
	Pain/Discomfort	3
	Anxiety/Depression	3
Years of Healthy Life	Activity Limitation	6
	Self-Rated Health	5
Health Scale for DALYs	Disability	7

The EQ-5D and the HUI are designed so that individuals are asked to place themselves on each of the dimensions in the instrument. The EQ-5D divides each attribute into three lev-els—no problem, some or moderate problems, and inability or extreme problems with the con-struct. The individual respondent is required to interpret these adjectival phrases, so "some problems" may have a different meaning for different individuals. In this sense, the EQ-5D in-corporates individuals' perceptions of degree of effect that their condition has on the different dimensions of health. The HUI also is designed to have individuals pick the category in which they fit on each attribute. The categories, however, are defined more extensively and somewhat less subjectively than in the EQ-5D—for example, the middle category of ambulation is "Able to walk around the neighborhood with walking equipment, but without the help of another person."

An individual is classified on the QWB dimensions as a result of either direct self-classification on the symptom/problem complex, or slightly less directly on the other dimensions. For the symptom/problem complex, the subject is shown or read a list of possible health conditions (e.g., "general tiredness, weakness, or weight loss," or "trouble learning, remembering, or thinking clearly"). The subject (or a proxy) identifies all conditions that they have been affected by in the past 6 days. For the other dimensions, questions are asked that help classify the per-son's functional ability. For example, for "mobility," which is used to indicate how able the per-son is to get around the community, the person is asked if he or she has a driver's license, and if not, is this because of their health. If the person does have a license, he or she is asked if he or she drove a car in the past day; if not, was this for health reasons. If the person did not drive, he or she is asked whether he or she used public transit, and if so, was more help than usual needed, and if not, was this because of his or her health. These questions are to determine the degree of mobility

and mobility restrictions caused by health. This type of questioning is an alternative to directly the person to self-classify the degree of limitation.

The YHL health classification system was developed to use data from the U.S. National Health Interview Survey (NHIS). One question asks each respondent to self-classify their health as excellent, very good, good, fair, or poor. A pre-defined logic is used to classify the person's degree of limitations in daily activities based on a number of questions about which activities he or she can engage in. The respondent is not asked to self-classify into limitation categories.

The QWB, HUI, and EQ-5D are designed to be assessed directly as primary data. The health classification of the YHL is a secondary analysis of pre-existing NHIS data, although in principle it could be administered as the primary purpose of data collection.

The disability classification used for computing DALYs relies on secondary data analysis, rather than primary data collection. A panel of experts has classified diseases and disabling conditions by how likely they are to produce varying levels of disability in recreation, education, procreation, occupational activities, or in activities of daily living. These judgments are used in secondary analyses of existing data sets to infer prevalence and degree of limitations in different populations (and to relate these inferences to specific health conditions).

The SF-36 classifies individuals using an entirely different process than the other five measures. This instrument, and many similar instruments, use multiple questions about functions, abilities, feelings, and attitudes. Developers of instruments in this genre of health measures start by thinking of the various dimensions of health they wish to measure. But rather than going through a process of enumerating a categorical scale for each concept in the measure, they create many specific questions relating to the concept.

For instance, if a scale of physical function is to be created, a pool of questions about specific physical activities will be created. This pool of questions, or "items" in the jargon, will cover items ranging from activities that all but the most impaired may be able to do, such as turning over, or toileting, or dressing or bathing oneself, to items about activities only the most physically adept might be able to do such as running long distances or playing vigorous sports. Ideally, if "physical functioning" is a unidimensional concept, all the items can be arranged so that they form a ladder of physical function—a person of a given ability would be able to perform all tasks from the low end of this ladder up to a certain point, then would not be able to perform any tasks on the ladder above this point.[d]

Unfortunately, it is difficult to make such a perfect list as there are different aspects of physical functioning, such as fine motor control, flexibility, cardiovascular fitness, strength, for example. A person with arthritis affecting the finger joints may not be able to sew, for example, but may walk long distances.

Items to form a scale are tested during instrument development and selected to ensure they sample a wide spectrum of activities thought to be associated with the underlying construct. The more items that are used in construction of the scale, the more precision there is to measure an individual's performance on that dimension. Developers of such scales must balance the desire to have greater precision with the problem of getting subjects to answer long questionnaires. So a minimum number of items is selected for the instrument so that they collectively have good test-retest reliability and can be shown to correlate well with other measures and indicators of the construct.

The SF-36 uses 36 items spread across 8 constructs—physical function, role functioning as limited by physical problems, bodily pain, social function, mental health, role function as lim-

ited by emotional problems, vitality, and general perceived health. The measurement philosophy, using multiple items regarded as each being a sample of abilities and attitudes, constitutes a different class of measurement approach from the attribute categorization of the other health measurement systems discussed above.

Assigning numerical weights to health state categories. Once a classification system is set up, the final step is to assign numbers—the health state weights—to the different health states that the classification system can distinguish. There are both fundamental and real differences in how different systems derive the numbers used to weight the states, and there are apparent differences that are in fact mostly superficial.

Let us deal with the superficial aspects first so that they can be put aside. The first of these is orientation—which end of the scale is "up." The QWB and the HUI are scaled from 0 to 1, where higher numbers mean better health states. The DALY disability weights range from 0 to 1, but the lower end of this scale indicates less disability, so lower numbers indicate better states. Orientation is a matter of cosmetics and perhaps convenience and not fundamental differences.

Another matter of cosmetics is the numerical range of the scale. In the same way that temperature can be measured in degrees Kelvin, Celsius, or Fahrenheit, we should not be concerned with scales which are simple transformations of each other. Scales that simply use different numerical endpoints are not necessarily fundamentally different measures.

So what is a fundamental difference in assigning numbers to categories? The most basic difference is whether the numbers reflect preferences—i.e., whether they are derived from a human judgment about the relative desirability of being in one state or another—or are derived in a manner not directly related to preferences.

The eight scales of the SF-36 are each computed from a simple scoring scheme that is not preference-based. For example, the physical function scale is formed by 10 items asking about degree of limitation in performing various physical tasks. The respondent can answer "limited a lot," "limited a little," or "not limited at all" for each of the 10 items. These responses are scored numerically as 1, 2, or 3, and the responses across the 10 items are added to yield a total score that can range from 10 (limited a lot on all 10 items) to 30 (not limited at all on all 10 items). This sum is then rescaled by simple transformation to range from 0 to 100, where 0 is minimum physical function on all 10 items and 100 is maximum function on all 10 items. The physical function scale of the SF-36, formed by this score ranging from 0 to 100 is a reliable indicator of physical function ability and has been shown to correlate well with many other indicators of physical function and to discriminate well between different groups of patients known to differ in physical abilities.[27]

While we generally can agree that a higher score on this scale is likely to be a better health state than a lower score, the scale is not constructed so that a 10-point difference, for example, is the same amount of increase in desirability of the health state all along the scale. A change from 40 to 50 on the scale may or may not be equivalent to a change in desirability when increasing from 80 to 90. So while this scale may be a reliable indicator of changes in physical function, there is no psychometric guarantee that equal numerical changes are equally desirable changes.[e]

Alternatively, the QWB, HUI, and DALY health measures were directly developed from judgments of desirability of the health states. Apparently the EQ-5D is being scaled in this manner from data that has been collected. The assignment of weights for the YHL was done in a manner allowing it to indirectly approximate the HUI, and to reflect preference judgments even if not developed from primary judgment data.

Although the terminology has not been standardized formally, there is a recent trend in the scientific literature to refer to measures based on preferences as health-related quality of life (HRQL) measures, to distinguish them as a subclass of all health status measures. By this convention, all the measures are health status measures, the QWB, HUI, DALY, and EQ-5D are health-related quality of life measures, and the YHL health measure is a proxy HRQL measure. By extension, the generic class of all adjusted life-year measures is HALYs and the subclass employing HRQL measures for the health component are QALYs.

The U.S. Public Health Service Panel on Cost-effectiveness in Health and Medicine, after due deliberation, concluded that a QALY measure is required for generalizable cost-effectiveness calculations.[18] However, this does not preclude the use of a non-QALY HALY measure for descriptive summary purposes.

There is much dispute about whose preferences should be used, and how to collect preference judgments. We'll first discuss whose preferences to collect. For QALYs to be used for individual decision making there is no question that it is that individual's own preferences that we wish to use. But for public policy purposes the answer is not so clear. It seems desirable to weight health states in such a way as to reflect public opinion. But there are many health states with which the general public is unfamiliar. If a survey of public opinion is conducted, these states may receive more or less weight than they would from persons who have suffered them. This may be because of an inability to imagine validly living with the health state, or because of fear of the health state, or other reasons. For instance, it is commonly held that the state of being blind is weighted substantially lower by people who have not been blind than by many who have been blind for an extended time.

The QWB and the HUI (and thereby the YHL) have been developed from broad community surveys. Members of the community were asked to rate health states and their answers were pooled and analyzed to develop a scoring system that predicts the community-assigned score for each particular health state. There was no special effort made to ensure that persons with a particular condition were assigned to rate that health state. So people with different conditions are theoretically represented in the sample roughly in proportion to the prevalence of the health condition in the community.[f] The DALY weights were derived as considered judgments by a panel of experts.[24]

The EQ-5D uses peoples' ratings of their own general health state to derive weights for levels of the five dimensions. In a large population survey with each respondent rates himself or herself on each of the five dimensions, and then makes an overall numerical rating of his or her general health from worst imaginable health state (0) to best imaginable health state (1). Statistical modeling of these data can then be used to develop weights that relate observations on the dimensions to the overall rating of health.

How to collect preferences is hotly debated. In the theory of decision making based on the principle that we should maximize expected utility of decision outcomes, there are many demands placed on the numerical utility scale used to evaluate decision outcomes. Because cost-effectiveness analysis comes from these roots, there is strong theoretical reason to demand that our eventual QALY measure should be a utility measure. And to do this, the health state weights have to be a utility measure. To explain exactly what this means mathematically is beyond the scope of this paper, and those interested can find technical discussions in Pliskin, Shepard, and Weinstein (1980); Miyamoto and Eraker (1985); Torrence, Boyle, and Horwood (1982); Torrande, Thomas, and Sackett (1972);[30] Torrance (1988),[31] and Gold, Patrick, et al. (1996). In-

depth technical references are books by Keeney and Raiffa (1976)[32] and by von Winterfeldt and Edwards (1986).[33]

The idea is to tie the numerical judgment scale to decisions by people about health states. The Standard Gamble (SG) method for doing this incorporates uncertainty as a part of the decision environment. Respondents are shown a scenario describing a particular health state and are asked to imagine the certain prospect of living in that state for their remaining lifetimes and then dying. Alternatively, they can choose an action where they have a 60 percent chance of immediate death or a 40 percent chance of living their remaining lifetimes in excellent health. If they choose the first alternative, the gamble is made more attractive by increasing the chance of excellent health; if they choose the gamble, it is made less attractive by decreasing the probability of excellent health. Then they are offered the two choices again, the sure thing with the specified health state or the gamble. This continues until they cannot choose between the two alternatives. At this point of indifference, it is possible to derive a utility scale number for the health scenario in the sure thing alternative that is a function of the probability of excellent health in the gamble.

Repeated use of the SG assessment method will obtain health state utilities for any specified health states. Or, this technique can be used with the separate dimensions to develop an entire scoring algorithm relating the components of the health state classification system to scores for health states.

A second method for assessing weights for health states is the Time Trade-Off (TTO). Using this method the person is asked to consider living his or her remaining lifetime (usually specified as a fixed time roughly commensurate with age) with the health state described in the scenario, (same as the first alternative in the SG technique). The alternative to this is to live for a fraction of the fixed remaining lifetime, for example, 50 percent, but in excellent health. If the person prefers the first alternative, longer life but in worse health, the amount of time in excellent health is increased to make the second alternative more attractive. If the preference was the other way, then the amount of time in excellent health is decreased. Proceeding in this manner it is possible to find the indifference point, where remaining lifetime in the worse state of health is equivalent to less time in excellent health. This equivalence yields an equation by which a numerical weight can be derived for any health state.

Although there are undoubtedly psychometric problems with these two assessment techniques, the SG and the TTO are considered to be the "gold standard" methods to elicit utility-based health state weights. There are computer programs under development that can be used to conduct the elicitation so that no interviewer bias intrudes. The HUI was explicitly developed using these elicitation techniques.

What alternative is there to these methods? We can always simply ask a person to rate the relative desirability of a health state on a scale relative to 0 (dead) and 1 (excellent health). Vis-ual aids such as a drawing of a scale marked in divisions of tenths can aid this direct rating. These methods, often utilized, are called the direct rating (DR) and the visual analog (VA) meth-ods. The QWB was derived from data collected using the DR method. The EQ-5D is being scaled from VA-elicited ratings.

The weights derived by the four methods do differ. Generally, a given state is weighted highest with the SG method and the TTO method. DR and VA weights range more throughout the 0–1 scale. Which are more valid? It is not clear, especially for purposes of public policy decision making. Although the SG and TTO methods are in principle tied to the theory of individual decision making, many consider the questions stilted and too hypothetical to elicit meaningful

responses. Others seem to have little problem eliciting apparently meaningful weights with these methods.

The Panel on Cost-Effectiveness in Health and Medicine preferred these assessments by a slim margin. Where SG and TTO are strongly supported by theory, people are discomfited by their cognitive awkwardness. Where DR and VA are easily understood and may be consistently used by respondents, it is not clear that their specific numerical results give us the scale properties needed to do precise arithmetic for health care cost-effectiveness analysis.

I do not see the definitive experiment on the horizon to decide this issue. I expect that public policy decision makers will have to choose the measurement technique that prima facie produces the numerical results most meaningful to their decisions, then gain experience with how the weights derived with that technique behave when tracked in the form of QALYs or QALE over time in a population.

My opinion is that most health state weighting schemes will weight states in pretty much the same overall order, at least on a gross scale. For instance, my colleagues and I have shown that about 50 percent of the population variation in QWB scores can be predicted at a gross level using SF-36 data.[34] From a policy standpoint, it is important that the measures on which decisions will be based include the constructs of interest in describing health. The numbers should be meaningful to the policy makers—this will come in part from the logic of the construction and derivation, but also from experience examining data in a variety of contexts and time frames.

PUTTING IT ALL TOGETHER

We have explored a number of issues concerning the "H" part of HALY measures. These include:

- choice of constructs to be included in the measure of health states
- how the health state conceptualization is operationalized
- how the specific weights are obtained for health states
  - — preference versus nonpreference scaling
  - — how to collect preference weights: SG, TTO, DR, VA

We have established a terminology that helps to distinguish methodological choices:

- A Health Status measure is a system for weighting health states. An example is the SF-36.
- Preference-based health status measures are health-related quality of life measures. Examples are the HUI, QWB, DALY, EQ-5D, and maybe the YHL measure.
- A HALY is a health-adjusted life year summary computed with a health status measure
- A QALY is a health-adjusted life year summary computed using a health-related quality of life measure.

QALYs appearing in the literature have been based on the HUI and on the QWB. These are community-based indexes of health-related quality of life. QALE for a population has been computed using HUI in Canada[6] and approximated in the United States using YHL.[5] The QWB

has been used to make a similar computation for a community population based on community data.[35,36]

All these computationsare done with QALYs in a positive orientation, that is, more is better. The computations are meant to describe the health-adjusted longevity of a population based on cross-sectional or brief longitudinal observations of quality of life and longevity.

The DALY is also a form of QALY, however it has a negative orientation as a loss of health rather than a gain in health. It is a representation of deficit in adjusted life years from full-health life expectancy in a population. QALYs computed with HUI or QWB are a "glass half full" measure and DALYs are simply QALYs as a "glass half empty" measure. In the positive orientation we have a measure of health achieved compared to accruing no life years at all. In the negative orientation we have a measure of burden of illness in the sense of deficit from full achievement. The latter calculation will depend on what standard is used to represent "full health"; this is discussed at length by Murray.[25] As implemented by Murray, the DALY also has one other important property: the life years are also weighted. In every other computation of QALY or HALY that I am familiar with, a life year is counted as a life year regardless of the age of the person living that life year. In the DALY framework, not only are health states weighted, but there is a weighting of life years that accentuates the concept of dependency of the very young and the very old, in a sense giving more weight to years accumulated in the population during productive adulthood. Although this possibility of weighting life years exists for any QALY measure, the DALY implementation is the only one to date that has incorporated it.

I have not seen QALYs computed using EQ-5D. One study computing clinical cost-effectiveness made a comparison HALY computation using the physical function scale from the SF-36,[37] but did not rely on this for the final results. I have recommended that researchers can use the regression equation I have reported[34] to convert full SF-36 profiles into pseudo-QWB scores and from these to compute approximate QALYs. I continue to believe this would be acceptable at a population level but not at an individual decision making level.

What are policy makers to do with all of these choices? First, examine the assumptions underlying the HALY computation. Beyond assuming that "a HALY is a HALY," the assumption of separability of health state weights from duration of the health state is an important one—it greatly simplifies the data required, and this simplification needs to be weighed against the meaningfulness of the results. Policy makers should make sure that the health measure used to make the computation in fact responds to the aspects of health they feel are important. They also should decide whether it is important to have a scale for health states that is weighted in a manner responsive to preferences or not.

Although we can compute HALYs or QALYs (or DALYs) by many different means, the answers may not be the same, so a choice among measures is a real choice. It may be important to adopt one or two measures provisionally and gain experience with them over time in a wide variety of settings, to understand where they correspond with intuition and where they do not, where they are responsive and where they are not, and where they are collectable and where they are not.

NOTES

[a] What I recount here is a North American history of the construct. I am unfamiliar with the early European literature in this regard.

[b] In fact, a director of the U.S. National Center for Health Statistics proposed a use of an index representing shortfall in HALE from what could be expected ideally as a "gross national health index" (Linder [1966][7] cited in Patrick and Erickson [1993].)[8]

[c] This was an idea whose time had come. It was being invented in several literatures about this time and reviewers of different disciplinary backgrounds will no doubt trace its roots differently.

[d] A continuum of items with this property is called a "Guttman Scale," named after the psychologist who first identified it as a scale type. A physical model for this is the machine for sorting oranges by size. Oranges are rolled down a trough with holes of successively larger diameter cut in the bottom. An orange will roll down the trough crossing the holes until it comes to the hole just big enough for it to fall through. We know that it is bigger than all the holes it roled across and smaller than holes farther down the trough from where it fell through. In the other instruments, YHL, HUI, QWB, DALY, and EQ-5D, each single dimension is divided into categories that at least at a gross level are defined to form Guttman scales. For example, the five categories of the HUI's speech dimension are: (1) able to be understood completely when speaking with strangers; (2) able to be understood partially when speaking with strangers, but to be understood completely when talking with someone who knows me well; (3) able to be understood partially when speaking with strangers or people who know me well; (4) unable to be understood when speaking with strangers but able to be understood partially by people who know me well; (5) unable to be understood when speaking to other people or unable to speak at all.

[e] An easily understood analogy is that of the fever thermometer. It is a reliable instrument for indicating change in fever status of a patient, and we can probably find that in relevant ranges higher temperatures indicate more uncomfortable health states in most people. But a change from 99°F to 100°F may not at all be the same in subjective discomfort as a change from 102°F to 103°F.

[f] In fact, as empirical collection of community survey data usually require exclusion of the very ill or those mentally unable to respond, these states may be under represented

REFERENCES

1. Sanders, B. Measuring community health levels. *American Journal of Public Health* 54(7):1063–1070, 1964.
2. Sullivan, DF. *Conceptual Problems in Developing an Index of Health*. Washington, D.C.: U.S. Department of Health, Education, and Welfare, 1966.
3. Moriyama, IM. Problems in the measurement of health status. In Sheldon, EB, Moore, WE, eds. *Indicators of Social Change*. New York: Russell Sage Foundation, 573–600, 1968.
4. Sullivan, DF. A single index of mortality and morbidity. *HMSHA Health Reports* 86(4):347–355, 1971.
5. Erickson, P, Wilson, R, Shannon, I. Years of healthy life. *Healthy People 2000 Statistical Notes*. Hyattsville, MD: National Center for Health Statistics/CDC/DHHS, vol. April, 1995.
6. Wolfson, MC. Health-adjusted life expectancy. *Health Reports (Statistics Canada)*. 8(1):41–46, 1996.

7. Linder, FE. The health of the American people. *Scientific American* 214(6):21–29, 1966.
8. Patrick, DL, and Erickson, P. *Health Status and Health Policy. Quality of Life in Health Care Evaluation and Resource Allocation*. New York: Oxford University Press, 1993.
9. Garber, AM, Weinstein MC, Torrance GW, and Kamlet MS. Theoretical foundations of cost-effectiveness analysis. In Gold MR, Siegel JE, Russell LB, and Weinstein MC, eds. *Cost-Effectiveness in Health and Medicine*. New York: Oxford University Press, 25–53, 1996.
10. Klarman, HE, Francis, JO, and Rosenthal, GD. Cost-effectiveness analysis applied to the treatment of chronic renal disease. *Medical Care* 6:48–54, 1968.
11. Packer, AH. Applying cost-effectiveness concepts to the community health system. *Operations Research* 16:227–253, 1968.
12. Fanshel, S, and Bush, JW. A health-status index and its application to health-services outcomes. *Operations Research* 18:1021–1066, 1970.
13. Weinstein, MC, and Stason, WB. Foundations of cost-effectiveness analysis for health and medical practices. *New England Journal of Medicine* 296:716–721, 1977.
14. Pliskin, JS, Shepard, DS, and Weinstein, MC. Utility functions for life years and health status. *Operations Research* 28:206–224, 1980.
15. Bleichrodt, H, Wakker, P, and Johannesson, M. Characterizing QALYs by risk neutrality. *Journal of Risk and Uncertainty* 15:107–114, 1997.
16. Bleichrodt, H, and Quiggin, J. Characterizing QALYs under a general rank dependent utility model. *Journal of Risk and Uncertainty* 15:151–165, 1997.
17. Gold, MR, Siegel, JE, Russell, LB, and Weinstein, MC. *Cost-Effectiveness in Health and Medicine*. New York: Oxford University Press, 1996.
18. Gold, MR, Patrick, DL, Torrance, GW, et al. Identifying and valuing outcomes. In Gold, MR, Russell, LB, Weinstein, MC, ed. *Cost-Effectiveness in Health and Medicine*. New York: Oxford University Press, 82–134, 1996.
19. Sackett, DL, Torrance, GW. The utility of different health states as perceived by the general public. *Journal of Chronic Diseases* 31:697–704, 1978.
20. Miyamoto, JM, and Eraker, SA. Parameter estimates for a QALY utility model. *Medical Decision Making* 5(2):191–213, 1985.
21. Kaplan, RM, Anderson, JP, Wu, AW, Mathews, WC, Kozin, F, and Orenstein, D. The Quality of Well-Being Scale. Applications in AIDS, cystic fibrosis, and arthritis. *Medical Care* 27(3 Suppl):S27–43, 1989.
22. Torrance, GW, Boyle, MH, and Horwood, SP. Application of multi-attribute utility theory to measure social preferences for health states. *Operations Research* 30(6):1043–1069, 1982.
23. Boylem, MH, Furlong, W, Feeny, D, Torrance, GW, and Hatcher, J. Reliability of the Health Utilities Index-Mark III used in the 1991 cycle 6 Canadian general social survey health questionnaire. *Quality of Life Research* 4(3):249–257, 1995.
24. Murray, CJL. Quantifying the burden of disease: The technical basis for disability-adjusted life years. *Bulletin of the World Health Organization* 72(3):429–445, 1994.
25. Murray, CJL, Lopez AD, eds. The global burden of disease : A comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020. Cambridge, MA: Harvard School of Public Health on behalf of the World Health Organization and the World Bank; distributed by Harvard University Press, 1996.
26. Group, TE. EuroQol—a new facility for the measurement of health-related quality of life. *Health Policy* 16:199–208, 1990.

27. Ware, J, Jr., and Sherbourne, CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Medical Care* 30(6):473–483, 1992.
28. McDowell, I, and Newell, C. Measuring Health: A guide to rating scales and questionnaires. (2nd ed.) New York: Oxford University Press, 1996.
29. McHorney, CA. Generic health measurement: Past accomplishments and a measurement paradigm for the 21st century. *Annals of Internal Medicine* 127(8):743–750, 1997.
30. Torrance, GW, Thomas, WH, and Sackett, DL. A utility maximization model for evaluation of health care programs. *Health Services Research* 7:118–133, 1997.
31. Torrance, GW. Measurement of health state utilities for economic appraisal: A review. *Journal of Health Economics* 5:1–30, 1988.
32. Keeney, RL, and Raiffa, H. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. New York: John Wiley & Sons, 1976.
33. Von Winterfeldt, D, Edwards, W. *Decision Analysis and Behavioral Research*. New York: Cambridge University Press, 1986.
34. Fryback, DG, Lawrence, WF, Martin, PA, Klein, R, and Klein, BEK. Predicting Quality of Well-Being scores from the SF-36: Results from the Beaver Dam Health Outcomes Study. *Medical Decision Making* 17(1):1–9, 1997.
35. Lawrence, WF, Fryback, DG, Klein, R, and Klein, BEK. Community-based quality-adjusted life expectancy: Results from the Beaver Dam Health Outcomes Study. Medical Decision Making 16:454, Oct.–Dec. 1996 [Abstract].
36. Rosenberg, MA, Fryback, DG, and Lawrence, WF. Population-based estimates of health-adjusted life expectancy: Comparison of alternative methods for computation. *Medical Decision Making* 17(4):527, 1997 [Abstract].

APPENDIX D

Distributive Justice and the Use of Summary Measures of Population Health Status

Norman Daniels, Ph.D.
Tufts University

OVERVIEW OF A MORAL CONTROVERSY ABOUT METHOD

In his contribution to this workshop, Dan Brock has focused on ethical issues that arise in the construction of different summary measures of health status and the benefits of health interventions. The questions he raises about possible age bias or bias against people with disabilities are important whether one is simply surveying the health status of different populations or actually making resource allocation decisions. I focus more narrowly on issues of distributive justice that arise when methods using these measures, such as cost-effectiveness analysis (CEA), are deployed to help us make resource allocation decisions. Obviously, how these measures are constructed has distributive implications in this central use, as does the underlying utilitarian framework of the method itself. These implications mean that the division of labor between us cannot be a line in the sand.

My ultimate conclusion is about both the construction and use of these measures. In contexts where we use CEA—and thus these measures—to make resource allocation decisions, we face a particularly difficult set of distributive issues. How much priority should we give to the sickest or worst-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? These questions form a family of "unsolved rationing problems." We have no principled solutions to them (though we may eventually discover some), and there is considerable moral controversy focused on them.

The straightforward use of CEA would, however, push us toward specific, yet contested, answers to these questions (Harris, 1987). The use of these measures in CEA would give no priority to the sickest patients, would permit any aggregation that maximized health benefit per dollar spent, and would always support best outcomes. What is contested—and I believe unacceptable—is the underlying utilitarian thrust to these answers. The central utilitarian assumptions are that a benefit to one always compensates for a loss to others, and that it is always morally desirable to maximize in the aggregate or at the margin. These assumptions, as Rawls (1971) argued, ignore the "separateness" of persons.

To get around this criticism, some might propose building into summary population measures "moral weights" that reflect our preferences or values when we take a stand on these distributive problems. Their goal, perhaps, might be to develop an ethically sensitive tool or decision procedure. Planners could then use it with fewer moral qualms in different resource allocation contexts, perhaps substituting "objective" calculation for "subjective" moral deliberation.

I believe that the unsolved nature of these morally contested rationing problems poses a serious obstacle to this strategy. For reasons I develop, we must view these measures and the methodologies that use them—or even highly improved versions of them—as inputs into a fair and deliberative decision making process. Our goal must be better informed and ethically sensitive deliberators making decisions, not methodologies that substitute for them.

To that end I urge a two-pronged research program. One prong explores social attitudes toward the distributive problems that must be addressed in making these resource allocation decisions. For example, Eric Nord's (1994) "person trade-off" approach explicitly asks people how many health outcomes of one kind (e.g., moving patients from one health state to another) they consider equal in social value to outcomes of a different kind. This approach avoids inferring an answer to this question from the very different question people are asked in standard summary measures, where they assign a personal utility to them of being in one state rather than the other. If developed further in directions Nord suggested, the person trade-off approach could help us learn more about how our society, or subgroups in it, reason about these issues. Serious obstacles confront this approach, however. Nord himself recognized some, and I note others. Still, properly pursued, it might provide an important body of information that could assist decision makers who must allocate resources.

The second component of the research program explores the requirements of a fair decision making process in the different contexts in which resource allocation decisions must be made. I make some preliminary suggestions about some elements of such a process, but much more work needs to be done. Decision makers constrained by such a fair process could then use information from summary health measures, CEA, and information about the attitudes and reasoning people use to think about these distributive issues to arrive at decisions others should view as legitimate and fair.

This argument has a brief history. Five years ago, at an international bioethics conference in Amsterdam, I claimed that the absence of principled solutions to these rationing problems means that we need to develop an account of fair procedures for resolving them (Daniels, 1993). Nord, speaking in the same session, replied that his person trade-off method can tell us how the public solves these problems and gives us a way to produce an instrument that incorporates the values underlying these solutions (Nord, 1993, 1994). I objected then that his method—which Brock and I believe asks the right questions—could not substitute for moral deliberation for various reasons. I continue that line of reply here, but I embrace the effort he makes to find out more about our beliefs about these distributive issues.

Early in 1993, the Public Health Service Panel on Cost-Effectiveness Analysis in Health and Medicine began its deliberations about the role and limits of CEA, and the argument I offer here was one of the considerations that led it to recommend that CEA should be viewed as an input to decision making and not a decision making procedure (Gold, Siegel, et al., 1996; Russell, Gold, et al., 1996). Unfortunately, simply making that recommendation without providing more assistance in helping us make these controversial, distributive decisions risks letting people give too much weight to the distributive implications of CEA. Even imperfect or distributively

insensitive measures could still act as an important input into a fair deliberative process that was highly sensitized to distributive controversies and our beliefs about them.

WHEN QUESTIONS OF DISTRIBUTIVE JUSTICE ARISE

The unsolved problems of distributive justice that concern us arise in some decision making contexts in which summary measures may be used but not in others, as Brock has suggested in his companion paper. Consider first a context in which these issues are not raised. In managed care organizations, coverage decisions for new technologies (drugs, devices, or procedures) are now generally made on a noncomparative basis (Daniels and Sabin 1997). In effect, each new technology is compared only to the existing or standard ways of treating the same group of patients with the same condition, and not to other technologies used for treating quite different groups of patients. For the most part, the decision is made to introduce a new technology if it produces a net benefit to that group of patients. To manage its costs and assure its quality, a coverage decision is then usually coupled with decisions that limit who may get it and who may perform it—that is, a "mini practice guideline" is developed.

Usually, this coverage decision is uninformed by formal or even informal CEA, except in the case of pharmaceuticals and, in some instances, devices. (Cost-effectiveness studies are generally not available in time to make these coverage decisions.) Suppose, however, a MCO used a formal CEA—with QALYs or DALYs or some other summary population health measure—to compare the new technology to the standard one. If the CEA shows they get the same or greater health benefit per dollar spent with the new technology, the MCO can make a completely noncontroversial distributive decision. In this special case the new technology benefits targeted patients, other enrollees, and the organization itself.

Suppose, however, we find that the new technology, when compared to the standard alternative treatment, produces a modest increment in benefit but at a significant increase in cost, yielding a less favorable cost-effectiveness ratio. The classic case in the literature is streptokinase versus TPA. A new one involves clopidogrel, a blood thinning agent much more expensive than aspirin but some 8 percent more effective in reducing infarcts among those at high risk. A MCO refusal to cover the less cost-effective technology rests on the judgment that the additional resources it would use can be put to better use elsewhere.

This decision does raise distributive questions, but they are most likely to be left unanswered, because the "better use" of those resources will not be specified. (The MCO would be unlikely to specify whether the "savings" would be used solely to deliver more cost-effective treatments or directed to nonmedical organizational uses (for remarks on "closed" or budgeted systems, see Daniels [1986], Daniels and Sabin [1998a]). Because the alternative uses are not specified and the distributive implications not determined, the judgment may seem noncontroversial. In this context we cannot see what might be problematic about the seemingly uncontroversial (especially among economists and economically trained planners) judgment that "we should act so as to get the most health benefit for every dollar we spend."

A more typical use of CEA to make a similar kind of judgment in an MCO (or in a task force or professional association) would be in evaluating alternative screening protocols. A CEA might compare screening protocols with different frequencies for mammograms, colonoscopies, or (to take the classic case) stool guaiacs. A recommendation to adopt one protocol because it is

appropriately cost-effective, and to reject an alternative because it reduces extra risks at too high a cost, has distributive implications. It imposes those extra risks on that screened population on the promise that there are better ways to use the extra resources that would be involved in the protocol with the worse CE ratio.

Here, too, the "better" uses of the extra resources are unspecified, and so we cannot say exactly which issue of distributive justice is raised. Again, it might seem that getting the most benefit for our marginal health dollar ought to be completely noncontroversial, but the response of advocacy groups to recommendations about screening protocols, sometimes dismissed as uninformed special pleading, should give us some pause. Those who face the identified extra risks may feel that reducing them is more important than assisting the unidentified others who may benefit from other unspecified uses of the extra resources. In effect, they portray themselves as "identified victims" and the others as completely indeterminate "statistical ones."

Now, however, suppose that MCOs make coverage decisions for new technologies in a directly comparative way, adopting a strict budget for the introduction of all new treatments and thus only able to adopt some of those competing for adoption. (As I noted earlier, there is no such budgeting now; whether continued downward pressure on premiums and the failure to squeeze savings out of traditional sources forces such budgeting remains to be seen.) They must compare as competitors, for example, a new treatment for people facing a life-threatening condition, a new procedure for the rehabilitation of patients with a debilitating injury, a surgical regimen for improving the quality of life of patients who have a chronic, disabling disease, and a pharmaceutical that reduces modest depression for a broad class of patients. If a MCO now decides to use CEA to chose among the new treatments, it faces various kinds of distributive questions. (Alternatively, to see the same issues, we could recast this as the problem of a federal or state board [e.g., Oregon's Health Services Commission] deciding which treatments should be included in a public health insurance benefit package. It is also the problem facing the World Bank when it evaluates which health care investments it should make among alternatives proposed for a developing country, or the problem faced by a country considering whether to accept a World Bank loan that uses CEA to impose certain priorities.)

Some of these distributive questions, as Brock suggests in his paper, result from the fact that different summary measures incorporate distinctive assumptions that have recognized distributive effects on different population groups. For example, if one of the new treatments involves primarily middle-age patients and another involves very young or very old ones, using QALYs or DALYs might affect their CE ranking. We must then decide whether it is morally desirable to value a life year the same at every age or to give it age-adjusted weights.

The distributive questions I want to focus on arise independently of some of these questions Brock has addressed. We can bring out their force by asking this question: is a QALY (or DALY) worth the same wherever it is distributed? Or should we ascribe a different value or moral importance to it depending on who gets it? Should we give moral priority, for example, to distributing the QALY to patients who are worst off to start with? Should we be neutral between distributing a thousand QALYs to 1000 people, each of whom improves by one QALY, or distributing those thousand QALYs to 50 people, each of whom gains 20 QALYs? Challenging the assumption that a QALY is always equal (in moral value) to a QALY, or a DALY to a DALY, goes to the heart of the distributive questions that concern us in what follows.

SOME UNSOLVED RATIONING PROBLEMS

The Priorities Problem

To illustrate the moral controversy that surrounds these unsolved rationing problems, consider first the priorities problem: How much priority should we give to treating the sickest or most disabled patients?

Imagine two extreme positions. The "Maximin" position ("maximize the minimum") says that we should give complete priority to treating the worst-off patients. The "Maximize" position says that we should give priority to whatever treatment produces the greatest net health benefit (or greatest net health benefit per dollar spent) regardless of which patients we treat. Suppose comparable resources could be invested in technology A or in B, but the resources are "lumpy" (we cannot introduce some A and some B) and we can only afford one of A or B in our MCO budget. The Maximin position would settle the matter by determining whether patients treated by A are worse off before treatment than patients treated by B. If so, we introduce A; if patients treated by B are worse off, we introduce B. If the two sets of patients are equally badly off, we can break the tie by considering to whom we can provide the most benefit. The Maximize position chooses between A and B solely by reference to which produces the greatest net benefit.

In practice, most people are likely to reject both extreme positions. If the benefits A and B produce are nearly equal, but patients needing A start off much worse than patients needing B, most people seem to believe we should introduce A. They prefer to provide A even if they know we could produce somewhat more net health benefit by introducing B. But if the net benefit produced by A is very small, or if B produces significantly more net benefit, then most people will overcome their concern to give priority to the worst off and will prefer to introduce B. Some people who would give priority to patients needing A temper their preference if those patients end up faring much better than patients needing B. In all situations where groups of students or health professionals have been informally polled on these cases, there is considerable disagreement: a definite but very small minority are inclined to be "maximizers" and a definite but very small minority are inclined to be "maximiners". Most people fall somewhere in between, and they vary considerably in how much benefit they are willing to sacrifice to give priority to worse-off patients.

Those who argue over these hypotheticals are quite willing to back their conclusions with reasons. Some will say, for example, "although patients needing B are being asked to forgo a significant benefit, I simply cannot turn my back on patients needing A, since they are so badly off." In response, someone else will say, "I hate to abandon A, but I simply cannot expect B to sacrifice a much greater benefit just because A starts off so poorly."

We might hope, faced with this kind of complexity, that a very careful examination of hypothetical cases might reveal some convergence on a complex set of underlying principles. This hope, however, may be unrealistic. The weightings that different people give to different moral concerns, such as helping the worst-off versus not sacrificing achievable medical benefits, probably depend on how these moral concerns fit within wider moral conceptions people hold. If so, there is good reason to think these disagreements will be a persistent feature of the situation. Indeed, some of the kinds of theoretical devices we might appeal to, such as forcing people to choose an allocation scheme from behind a veil of ignorance, are themselves the focus of considerable dispute. Is it reasonable, for example, for such people to gamble on their likelihood of

being one type of patient or the other, or must they somehow identify with each category of patients and refuse to gamble (see Daniels [1993], Kamm [1993], Scanlon [1982])?

The Aggregation Problem

When should we allow an aggregation of modest benefits to larger numbers of people to outweigh more significant benefits to fewer people?

In June 1990, the Oregon Health Services Commission (OHSC) released a list of treatment/condition pairs ranked by a cost/benefit calculation. Critics were quick to seize on rankings that seemed completely counter intuitive (other critics argue the problem arose because the OHSC used crude numbers). For example, as Hadorn (1991) noted, tooth capping was ranked higher than appendectomy. The reason was simple: an appendectomy cost about $4,000, many times the cost of capping a tooth. Simply aggregating the net medical benefit of many capped teeth yielded a net benefit greater than that produced by one appendectomy.

As Eddy (1991) pointed out, our intuitions in these cases are based largely on comparing treatment/condition pairs for their importance on a one to one basis. One appendectomy is more important than one tooth capping because it saves a life rather than merely reduces pain and preserves dental function. Our intuitions are much less developed when it comes to making one to many comparisons.

Kamm (1993) explored hypothetical cases, showing that we are not straightforward aggregators of all benefits, though we do permit some forms of aggregation. Nevertheless, our moral views are both complex and difficult to explicate in terms of well-ordered principles. Kamm argued quite plausibly that some minor goods are irrelevant to aggregation and should not be weighed at all against significant benefits. She also suggested that benefits we cannot be expected to sacrifice to help someone else may count as significant enough to be aggregated even against saving life. Thus, since someone is not required to sacrifice an arm to save someone else's life, it may be possible to aggregate the saving of some number of arms as opposed to saving a life. Kamm's discussion thus provides some important structure to the aggregation problem. Nevertheless, the principles that emerged do not constitute an adequate framework for addressing many real aggregation questions. The principles also may have emerged only because Kamm ignored variation in responses to some of her cases, as I suggested previously.

The Best Outcomes/Fair Chances Problem

How much should we favor producing the best outcome with our limited resources as opposed to giving people some fair chance at deriving some benefit from them? For example, which of several equally needy individuals should get a scarce resource, such as a heart transplant? Suppose that Alice and Betty are the same age, have waited the same time, and that each will live only 1 week without a transplant. With the transplant, however, Alice is expected to live 2 years and Betty 20. Who should get the transplant? Giving priority to producing best outcomes, as in some point systems for awarding organs, would mean that Betty gets the organ and Alice dies (assuming persistent scarcity of organs). But Alice might complain, "Why should I give up my only chance at survival—and 2 years of survival is not insignificant—just because Betty has a chance to live longer?" Alice demands a lottery that gives her an equal chance with Betty.

To see the problem in its macro-allocation version, suppose our health care budget allows us to introduce one of two treatments, A or B, which can be given to comparable but different groups. Because A restores patients to a higher level of functioning than B, it has a higher net benefit. We could produce the best outcomes by putting all our resources into A; then patients treatable by B might, like Alice, complain that they are being asked to forgo any chance of a significant benefit. One variation on this scenario raises the question of discrimination against people with disabilities: suppose the disease treatable by A can strike anyone, but the disease treatable by B tends to be associated with people suffering from some other significant disability. Then favoring "best outcomes" by putting our resources into A would clearly discriminate against people with disabilities who need B.

The problem currently has no satisfactory solution at either the intuitive or theoretical level. Brock (1988) proposed breaking this deadlock by giving Alice and Betty chances proportional to the benefits they can get (e.g., by assigning Alice one side of a 10 sided die). Kamm (1993) proposes a more complex assignment of multiplicative weights. Both suggestions have the advantage of taking what people view as relevant reasons into account. Such weighted lotteries might, for example, be justifiable as the result of deliberation among people about what weights they can agree to assign to these reasons. Viewed in this way, they seem less arbitrary or ad hoc than if they (the weightings) are taken to capture some underlying precision in our moral intuitions.

In the Absence of Consensus on Principles

The best outcomes/fair chances, priorities, and aggregation problems may turn out to have principled solutions on which consensus can be obtained. My claim here is not that they are unsolvable but only that they are unsolved now and that we have no real prospect of arriving at solutions that would be publicly acceptable in the foreseeable future.

My skepticism about any rapid solution rests in part on the fact that these are all problems on which there is moral disagreement. Such disagreement emerges quickly in class or group settings where hypothetical cases are discussed in detail. Some people give more weight to helping the worst off than others, or permit some forms of aggregating that others think objectionable, or give more weight than others to best outcomes rather than fair chances. Nord found that different subgroups of the Norwegian population tend to have systematically different responses to these problems. When the Swedish government set up a commission to establish principles for establishing priorities in its health care system, it gave great weight to helping the sickest or most disabled individuals, probably much more weight than other societies considering the same question would give, and more weight than many of my students polled on the issue were willing to give.

Even if there are principled solutions that philosophical investigation may eventually uncover, there is considerable disagreement now and among different groups about how to solve these problems. Typically, for example, a minority will be willing to give significant priority to the sickest patients, trading away much more significant benefits to those who are less sick to obtain some benefits for the sickest (as was the Swedish commission), but the majority is not. These commitments support two distinct criteria for ranking (or rationing) various kinds of treatments that might be applied to the very sick or the mildly sick. Each group is willing to provide reasons for its belief about the correct solution.

How should we decide among policies when there is this kind of disagreement underlying moral commitment? I suggest in Section 5 that we need to retreat to a form of procedural justice by giving an account of a fair decision making procedure.

On the Generality of These Problems

Though I have discussed these unsolved rationing problems primarily in health care contexts, I want to note that they are quite general. Similar issues would arise if we were allocating legal aid services, special education services, or even income support services. One extremely important context in which they arise is economic growth or development policy.

THE DEMOCRACY PROBLEM: ISSUES FACING PERSON TRADE-OFFS AND VOTING

Brock and I believe that Nord's (1994) person trade-off approach to valuing alternative health care programs addresses explicitly the distributive questions that need answers. By directly surveying people's attitudes toward trade-offs between allocating resources to groups that differ in their initial health state and ultimate health outcomes, Nord hoped to uncover the structure of our moral concern or our values regarding how much priority to give sickest patients. Nord (1994,201) noted some methodological problems his "demanding" approach faces, and I add some philosophical worries. Nevertheless, I urge we pursue a research strategy using such approaches to see if we can develop meaningful information about how society or subgroups in it reason about these distributive issues.

Nord's approach faces questions about reliability. He noted that surveys of small populations using his approach show considerable variance in answers; large populations would be needed to eliminate random errors. Reliability of the tool is also suspect because there are significant "starting point biases." People are asked how many of one kind of outcome they would sacrifice to achieve N outcomes of another kind. Higher initial values trigger higher responses. Other kinds of "framing" issues are raised by the attempt to elicit consideration of specific reasons or rationales for making the trades. To control for framing effects, Nord suggested subjects have to be taken through various steps in which they are exposed to different arguments that might be relevant to the exercise. These efforts by Nord are the most promising features of his approach because they may tap into beliefs that function as true reasons.

Based on my experience raising hypotheticals like these in classroom discussions (Nord himself uses seminars!) I speculate that the reliability problems Nord found may reflect subgroup disagreements about the weights to give different factors involved in these trades. These disagreements may reflect underlying differences in comprehensive moral views. Nord himself has presented some evidence for differences that vary with Norwegian political party affiliations. Taking means or ranges as a way of addressing this reason for divergence may be begging the very question raised by moral disagreement. Nevertheless, knowing the magnitude of these differences could help decision makers think about how much weight to give disagreements they encounter.

Suppose we find considerable convergence in a population (or subgroup) on the magnitudes involved in trade-offs. What should we make of it? Should we view it as a prevalent "taste" or "preference," the equivalent of a predominant taste for chocolate over vanilla ice cream? Or are

the responses the result of a deliberative or reflective process in which people weigh various reasons, proposed principles, and intuitions about particular cases and arrive at a coherent set of moral beliefs that for them are justified?

We risk having uncovered only "tastes" if we carry out a straightforward survey of attitudes toward trade-offs. Nord suggests, however, that we develop and deploy more complex methods that lead subjects in these surveys through a series of questions that import arguments and reasons that might be the basis for making these trades. This complex technique—which is quite demanding—begins to approximate the sort of philosophical exploration that is involved when students are led through the complexities of these issues by posing various hypothetical cases (cf. Kamm, 1993). Such an approach is more likely to uncover some evidence about what reasons people give weight to, and not simply their unconsidered tastes.

To see why it is important to seek reasons and not merely tastes, consider an analogy to democratic procedures. Surveys and voting are both ways of consulting people's opinions. What gives majority (or plurality) rule its legitimacy as a procedure for resolving moral disputes about public policy and the design of institutions? One prominent answer, which Cohen (1996) refered to as the "aggregative" conception of democracy, holds that the procedure is fair and acquires legitimacy simply because it counts everyone's interests equally in the voting process: each counts for one, not more or less. The same might be true of a survey of a representative sampling of people.

Something important seems to be left out of this proceduralist view of the virtues of aggregation through voting. It allows us to compel people to abide by majority rule, even where there are matters of fundamental moral disagreement, simply by aggregating the preferences of the voters, whatever they happen to be. If we had the option of buying only one flavor of ice cream, vanilla or chocolate, for a large group, we might settle the matter by voting. In this case, aggregating preferences through the mechanism of voting is a way to achieve the greatest net satisfaction of preferences, since the frustration of the vanilla lovers is offset by the greater pleasure of the chocolate lovers.

Abiding by a majority decision that compels people to act in ways that are counter to their fundamental beliefs about what is morally right is not, however, simply like frustrating a taste for vanilla ice cream. A strong craving for vanilla is not a moral conviction. Settling moral disputes simply by aggregating preferences seems to ignore fundamental differences between the nature of values and commitments to them and tastes or preferences. I have the same worry if a person-trade survey reveals only preferences.

The aggregative conception seems insensitive to how we ideally would like to resolve moral disputes, namely through argument and deliberation. We expect people to offer reasons and arguments for their moral views, and we hope that the better arguments will prove persuasive. We want to be shown what is right by appeal to reasons that we consider convincing. If a good moral argument persuades us that our original belief about what is right is in fact incorrect, we may be chagrined, but we are (or should be) grateful as well. We have been spared doing what is wrong. It is more important to end up knowing what is right and doing it, given our motivation to act in ways that we can justify morally, than it is to get our way.

These points help to explain why we are not satisfied in cases of moral disagreement simply to be told, "a majority of people think otherwise." The problem is not that the majority will simply keep us from getting our way, as it would be if we preferred vanilla but ended up with chocolate, but that majorities can be morally wrong and may make us do the wrong thing. In

addition, they may be moved by reasons that minorities cannot even accept as relevant to resolving the dispute.

The aggregative account fails as an account of the legitimacy of a democratic procedure because it ignores the way reasons play a role in our deliberations about what is right. An alternative account of how a procedure such as majority rule acquires legitimacy depends on emphasizing the deliberative process that may conclude in a vote. Specifically, it imposes some constraints on the kinds of reasons that can play a role in that deliberation. Not just any preferences will do. Reasons must reflect the fact that all parties to a decision are viewed as seeking terms of fair cooperation that all can accept as reasonable. Where their well-being or fundamental liberties or other matters of fundamental value are involved and at risk, people should not be expected to accept binding terms of cooperation that rest on reasons they cannot view as acceptable types of reasons. For example, reasons that rest on matters of religious faith will not meet this condition. Reasonable people differ in their religious, philosophical, and moral views, and yet we must seek terms of fair cooperation that rest on justifications acceptable to all.

Suppose that a deliberation appeals only to reasons that all can recognize as acceptable or relevant kinds of reasons, but that consensus about an outcome is still not achieved. To settle the practical matter, we rely on a majority vote. What can be said in favor of reliance on this voting procedure that could not be said on the purely proceduralist view? The minority is not being compelled to do something for reasons it thinks irrelevant or inappropriate—even if it does not accept the weight or balance given to various considerations by the majority. On the aggregative view, the minority has to accept that it loses only because more people prefer an alternative, for whatever reasons. On the deliberative democracy view, the minority can at least assure itself that the preference of the majority rests on the kind of reason that even the minority must acknowledge appropriately plays a role in the deliberation. The majority does not exercise brute power of preference, but is constrained by having to seek reasons for its view that are justifiable to all who seek mutually justifiable terms of cooperation.

A research strategy building on Nord's approach could reveal information about the types of reasons and arguments people give for their distributive choices and the weights they attach to them. This knowledge could inform a further deliberative process in which resource allocations must be made. Still, it is not a substitute for that process.

FAIR, DELIBERATIVE PROCEDURES: THE EXAMPLE OF MCOS

Elsewhere (Daniels and Sabin, 1997) I have discussed a partial account of fair procedures intended to address the kind of moral controversy that pervades decision making about health care resource allocation in MCOs. I can only sketch the approach here. The basic intuition behind it is that institutions making decisions about resource allocation—as MCOs do when they make coverage decisions—should meet several conditions that impose what I call "public accountability for reasonableness." These conditions connect deliberations about how to address distributive issues made by private organizations (or public agencies) to a broader social deliberation that involves broader democratic processes. For the sake of specificity, I state these as conditions that must be met for a highly visible and controversial area of decision making, coverage for new technologies, though they can be generalized to cover other forms of limit-setting.

1. Publicity condition: Decisions regarding coverage for new technologies (and other limit-setting decisions) and their rationales must be publicly accessible.

2. Relevance condition: These rationales must rest on evidence, reasons, and principles that all parties—managers, clinicians, patients and consumers in general—can agree are relevant to deciding how to meet those the diverse needs of a covered population under necessary resource constraints.

3. Appeals condition: There is a mechanism for challenge and dispute resolution regarding limit-setting decisions, including the opportunity for revising decisions in light of further evidence or arguments.

4. Enforcement condition: There is either voluntary or public regulation of the process to ensure that conditions 1–3 are met.

The guiding idea behind the four conditions is to convert private MCO solutions to problems of limit-setting and resource allocation—where highly controversial moral issues are at stake—into part of a larger public deliberation about a major, unsolved public policy problem, namely, how to use limited resources to protect fairly the health of a population with varied needs, a problem made progressively more difficult by the successes of medical science and technology. If met, these conditions help these private institutions to enable a more focused public deliberation that involves broader democratic institutions.

The publicity condition thus provides a public record of the commitments to which the plan adheres in making these kinds of decisions. A case law record such as this improves fairness in decision making because it provides a basis for judging the coherence and consistency of decisions made over time. It gives those affected by decisions—often when they have no real choice to seek alternatives—a way of knowing why they face the restrictions they do. The publicity condition thus satisfies what many believe is a fundamental requirement of justice: the grounds for decisions that fundamentally affect our well-being must be publicly available to us.

The relevance condition imposes important constraints on the kinds of reasons that should play a role in rationales for coverage decisions. The basic idea is that all parties in a MCO pursue a common goal or common good: they enter into a plan that aims to meet their diverse needs under necessary resource constraints. Since hard choices will have to be made about how to meet those needs fairly, the grounds for those decisions must be ones that all can agree are relevant to that kind of decision. A justification must be based on reasons all accept as relevant.

The relevance condition does not mean that all parties will agree with the specific decisions made. They may agree that reasons are relevant but still give different weight or importance to them. As long as all parties who make and are affected by the decision can accept that the grounds for it are relevant, then even those who do not like or agree with the specific outcome of the decision cannot complain that it is unreasonable.

CONCLUSION

Standard measures of population status and the benefits of health intervention, coupled with methods like CEA, are themselves insensitive to important questions of distributive justice. These ethical questions must be faced head-on. Unfortunately, they include a family of “unsolved rationing problems.” In the absence of prior consensus on principled moral solutions, we must

develop procedurally fair decision making processes and rely on them to give us legitimate and fair outcomes.

I urge a two-pronged research strategy to address this problem. One component involves empirical research on the distributive problems. It aims at uncovering the kinds of reasons people employ in making choices about these issues, the weights they give to these reasons, and the magnitude and patterns of divergence on their solutions. This is a very demanding type of survey research, mimicking in some ways highly "qualitative" philosophical exploration of these issues. A better understanding of a population's beliefs about these questions could then inform deliberators who face allocative choices. A further ethical issue is just how this information about beliefs should be used by decision makers.

A second component involves research—both empirical and ethical—into the design of fair, deliberative procedures for making these decisions in the various contexts and institutions where they must be made. My suggestions here about some constraints on that process in MCOs making decisions about coverage for new technologies are intended to illustrate the solutions that must be sought. The research is empirical, because it must reflect facts about the institutions in which decisions are made and because we may also want to test proposed procedures to see how feasible and effective they are at achieving legitimacy. It is also ethical because procedural fairness itself raises complex issues in ethics and political philosophy.

ACKNOWLEDGMENT

I wish to thank my research assistant, Roxanne Fay, for her extensive help in preparation of this paper.

NOTES

1. Brock asked the same question to show the age-weighting of DALYs versus QALYs; I ask it more generally here.

2. The claim is based on observations over several years of how audiences of students and medical personnel vote on hypothetical cases of this sort. Nord (1993) reported variations in attitudes toward priorities of this sort between different groups of students and professionals. There is some cross-national evidence that people are not straight maximizers in Nord, Richardson, Street, Kuhse, and Singer (1995).

3. A distinct minority of students and health professionals would argue as follows: if helping the better-off patient B actually returns B to a level of functioning that permits her to work and carry out other social functions, whereas helping the sicker patient A does not accomplish this outcome, then it is more important to help B. Some holding this view give as a reason that B returns more to society than A, but others justify their view by saying B's returning to closer to normal functioning will make B happier that A is likely to be; this reason focuses on the well-being of B and A, not on social contribution.

4. Frances Kamm (1993) suggested this may be true. Her brilliant discussion of cases often points to less disagreement than I find in thinking about them myself or with students or public audiences. For some concerns about her methods, see Daniels (1998a).

5. I had once criticized the proposal for being ad hoc and arbitrarily importing precision in this way. See Daniels (1993).

6. personal communication and presentation at the Stockholm Conference on Priorities in Health Care, October 17, 1996.

7. Suppose an antipoverty program in Bangladesh that has fixed resources might be targeted at the very poorest (VP) segment of a population or on the next poorest (but still poor) subgroup (P). Using the resources on P leads to more people being moved out of poverty and becoming producers capable of contributing in the future to further anti-poverty measures. But putting the resources into helping P leaves those in the most dire straits unaided. Is it fair to favor helping P over VP? Here the extra complexity of the problem is that it is more plausible to think not only of benefits to those helped but of the future social contribution of those who are helped. The ethical dimensions of these problems in these different contexts has been largely unexplored and warrant s considerable interdisciplinary effort.

REFERENCES

Brock, D. 1988. "Ethical Issues in Recipient Selection for Organ Transplantation." In D. Mathieu (ed.), *Organ Substitution Technology: Ethical, Legal, and public Policy issues.* Boulder: Westview, pp. 86–99.

Cohen, J. 1996. "Procedure and Substance in Deliberative Democracy." In Seyla Benhabib, ed., *Democracy and Difference: Changing Boundaries of the Political.* Princeton, NJ: Princeton University Press.

Daniels, N. 1998. "Kamm's Moral Methods." Forthcoming in *Philosophy and Public Affairs* 26:4:303–350.

Daniels, N. 1993. "Rationing Fairly." *Bioethics* 7:2–3:224–233.

Daniels, N. 1986. "Why Saying No is so Hard in the U.S." *New England Journal of Medicine* 314:1381–1383.

Daniels, N., and Sabin, J.E. 1998. "Closure, Fair Procedures, and Setting Limits in managed Care Organizations." *Journal of American geriatrics Society.* (In press).

Daniels, N. and Sabin, J.E. 1997. "Limits to Health Care: Fair Procedures, Democratic Deliberation, and the Legitimacy Problem for Insurers." *Philosophy and Public Affairs* 26:4:303–350.

Eddy, D. 1991. "Oregon's Methods: Did Cost-Effectiveness Analysis Fail?" *Journal of the American Medical Association* 265:2218–2225.

Gold, M., Siegel, J., Russell, L., and Weinstein, M. 1996. *Cost-Effectiveness in Health and Medicine.* New York: Oxford University Press.

Hadorn, D. 1991. "Setting Health Care Priorities in Oregon: Cost-Effectiveness Meets the Rule of Rescue." *Journal of the American Medical Association* 265:2218–2225.

Harris, J. 1987. "QALYfying the Value of Life." *Journal of Medical Ethics* 13:117–123.

Kamm, F. 1993. *Morality and Mortality, Volume 1: Death and Whom to Save from It.* New York: Oxford University Press.

Kamm, F. 1989. "The Report of the U.S. Task Force on Organ Transplantation: Criticisms and Alternatives." *Mount Sinai Journal of Medicine* 56:207–220.

Nord, E. 1993. "The Relevance of Health State after Treatment in Prioritizing between Different Patients." *Journal of medical Ethics* 19:37–42.

Nord, E., Richardson, J., Street, A., Kuhse, H., and Singer, P. 1995. "Maximizing Health Benefits vs. Egalitarianism: An Australian Survey of Health Issues." *Social Science and Medicine* 41:10:1429–1437.

Russell, L.B., Gold, M.R., Siegel, J.E., Daniels, N., and Weinstein, M.C., 1996. "The Role of Cost-Effectiveness Analysis in health and Medicine." *Journal of the American Medical Association* 276:14:1172–1177.

Scanlon, T.M. 1982. "Contractualism and Utilitarianism." In A. Sen and B. Williams, *Utilitarianism and Beyond.* Cambridge, England: Cambridge University Press, pp. 103–128.

APPENDIX E

Ethical Issues in the Development of Summary Measures of Population Health Status

Dan W. Brock
Brown University

INTRODUCTION

This paper will discuss briefly some of the main ethical issues in the development or construction of summary measures of population health status and of the health benefits of interventions designed to improve the health status of a population.[1] In a companion paper Norman Daniels will address issues of equity and distributive justice in the use of such measures for prioritization of health resources. Typical measures of the health status of a population at a point in time include the Health Utilities Index (HUI),[2] and the Quality of Well-Being Scale (QWB).[3] Some such measure will be needed as well to calculate the burdens of various diseases as well as the benefits of interventions to reduce the burdens of disease. Typical summary measures of the benefits over time of health interventions include Quality-Adjusted Life Years (QALYs) and Disability-Adjusted Life Years (DALYs), each of which will have to make use of some point-in-time measure of health status like the HUI. Both QALYs and DALYs are intended to be comprehensive measures of the overall benefits of health interventions because they each combine changes in length of life and changes in HRQL, the two general forms of benefits of health interventions.[4]

Both QALYs and DALYs are often employed in cost-effectiveness analyses (CEAs) that compare the aggregate health benefits secured from a given resource expenditure with different health interventions. The assumption underlying much of the use of CEAs is that limited resources (and, of course, as economists correctly remind us resources are always limited) should be used to maximize the aggregate health status of a population, or to minimize the burdens of disease for a population. Natural, even self-evident, as this assumption may appear to many health policy analysts and economists, both Daniels and I will argue that it assumes a utilitarian or consequentialist moral standard, or, more specifically, a standard of distributive justice, and that the utilitarian account of distributive justice is widely and correctly taken to be utilitarianism's most problematic feature.

FIRST ISSUE: HOW SHOULD STATES OF HEALTH AND DISABILITY BE EVALUATED?

Early summary measures of the health status of populations were often measures of a single variable which stood as a more or less crude measure of the health of a population, or as a surrogate for a measure of the health of a population, such as life expectancy or infant mortality. For some purposes, and in the absence of more fine-grained data about population health, these single variable measures can sometimes provide useful information. Health interventions can also be evaluated for the impact they have on increasing life expectancy or reducing infant mortality. Since virtually no one disagrees that it is desirable to reduce infant mortality rates, we can evaluate interventions for their effects in doing so without raising the problem of how to assign relative values to different health outcomes.

The usefulness of life expectancy or infant mortality rates is clearly very limited, however, because they give us information about only one of the aims of health interventions— extending life or preventing premature loss of life—and they provide only limited information about that aim. They give us no information about another aim, at least as important, that of health interventions to improve or protect the quality of life by treating or preventing suffering and disability. Multi-attribute measures like the Sickness Impact Profile[5] and the MOS 36[6] provide measures of different aspects of overall HRQL on which a particular population can be mapped, and an intervention assessed for its impact on these different components of health, or HRQL. Because these measures do not assign different relative value or importance to the different aspects or attributes of HRQL, they do not provide a single overall summary measure of HRQL. Thus, if one of two populations or health interventions scores higher in some respect(s) but lower in others, no conclusion can be drawn about whether the overall HRQL of one population, or from one intervention, is better than the other.

This limitation may not be serious in some contexts. For example, when evaluating some alternative interventions or pharmaceuticals in clinical trials, the impacts of the different interventions may be clustered in a limited domain of HRQL, and the different impacts in that domain of one intervention may be uniformly, or nearly uniformly, better than those of alternative interventions. Nevertheless, even in many clinical trials the outcomes of different interventions may be more multidimensional and conflicting in the sense that no one alternative dominates or is better than all the others in all of its effects on HRQL. For comparing overall population health between and among countries, states, or regions, and for resource prioritization or allocation at national, state, or regional levels, assignments of different relative value or importance to different health-related outcomes or effects are necessary.

Measures like QALYs and DALYs have the important advantage that they address both length of life and health-related quality of life, and they provide a basis for assigning relative value to length versus quality of life, as well as to different impacts on quality of life. The construction of any measure like the QALY or DALY requires a two-step process: first, different states of disability or conditions limiting HRQL are described; second, different relative values are assigned to those different conditions. Instruments like the HUI and the QWB have been developed to play this role. The determination of people's different health-related conditions both before and after a particular health intervention is an empirical question that should be answered by appeal to relevant data regarding the burden of a particular disease and the reduction in that burden that a particular health intervention can be expected to produce; the overall HRQL of a particular

population can likewise be determined by empirical data regarding that population's condition on the different attributes of a measure like the HUI. Needless to say, often the relevant data are highly imperfect, but that is a problem to be addressed largely by generating better data, not by ethical analysis.

The second step in developing measures like the HUI involves assigning relative values or utilities to the different conditions that reduce people's HRQL. The developers of the DALY used expert health professionals to make these value judgments. This may have been convenient, but value judgments about the degree to which various conditions reduce quality of life is not a matter to be settled by professional expertise. Moreover, health professionals may have systematic biases that skew their value judgments about quality of life compared to those of ordinary persons. Other measures like the HUI and QWB use the value judgments of a random group of ordinary citizens to evaluate different states of disability or limitations in function. The utilities for different attributes and their levels in the HUI are shown in Table 1.

Table 1 Health Utility Index—Mark II (Adult/ Children)

Vaccine preventable condition:
Disease scenario description:

Attribute	**Level**	**Description**	***utility fn***	***b***	
1. Sensory	1	Able to see, hear, and speak normally for age	1.00	0	
	2	Requires equipment to se or hear or speak	0.95	0	
	3	Sees, hears, or speaks with limitations, even with equipment	0.86	0	
	4	Blind, deaf,. or mute	0.61	0	
1. Sensory Total					**0**
2. Mobility	1	Able to walk, bend lift, jump, and run normally for age	1.00	0	
	2	Walks, bends, lifts, jumps, or runs with some limitations, but does not require help	0.97	0	
	3	Requires mechanical equipment (such as canes, crutches, braces, or wheelchair) to walk or get around independently	0.84	0	
	4	Requires the help of another person to walk or get around and requires mechanical equipment as well	0.73	0	
	5	Unable to control or use arms and legs	0.58	0	
2. Mobility Total					**0**
3. Emotion	1	Generally happy and free from worry	1.00	0	
	2	Occasionally fretful, angry, irritable, anxious, or depressed (or suffering night terrors—for children)	0.93	0	
	3	Often fretful, angry, irritable, anxious, or depressed (or suffering night terrors—for children)	0.81	0	
	4	Almost always fretful, angry, irritable, anxious, or depressed	0.70	0	
	5	Extremely fretful, angry, irritable, or depressed, usually requiring hospitalization or psychiatric institutional care	0.53	0	
3. Emotion Total					**0**
4. Cognitive	1	Learns and remembers normally for old age (e.g., schoolwork—for children)	1.00	0	
	2	Learns and remembers more slowly than normally for age—for adults; learns and remembers schoolwork more slowly than classmates, as judged by parents and/or teachers—for children)	0.95	0	
	3	Learns and remembers very slowly and usually requires special assistance in learning situations	0.88	0	
	4	Unable to learn and remember	0.65	0	
4. Cognitive Total					**0**

5. Self-care	1	Eats, bathes, dresses, and uses the toilet normally for age	1.00	0
	2	Eats, bathes, dresses, or uses the toilet independently with difficulty	0.97	0
	3	Requires mechanical equipment to eat, bathe, dress, or use the toilet independently	0.91	0
	4	Requires the help of another person to eat, bathe, dress, or use the toilet	0.80	0
5. Self-care Total				**0**
6. Pain	1	Free of pain and discomfort	1.00	0
	2	Occasional pain; discomfort relieved by nonprescription drugs or self-control activity without disruption of normal activities	0.97	0
	3	Frequent pain; discomfort relieved by oral medicines with occasional disruption of normal activities	0.85	0
	4	Frequent pain; frequent disruption of normal activities; discomfort requires prescription narcotics for relief	0.64	0
	5	Severe pain; pain not relieved by drugs and constantly disrupts normal activities	0.38	0
6. Pain Total				**0**
7. Fertility	1	Able to have children with a fertile spouse	1.00	0
	2	Difficulty in having children with a fertile spouse	0.97	0
	3	Unable to have children with a fertile spouse	0.88	0
7. Fertility Total				**0**
Health State: Utility fn		1.06*(b1*b2*b3*b4*b5*b6*b7)-.06		-0.06

One problem concerning which evaluations to use in examining different states of disability or functional limitation arises from the ability of individuals to adjust to their disabilities. This results in the disabled often reporting less distress and limitation of opportunity and quality of life from their disability than the nondisabled believe would occur. If the evaluations of disability states by the nondisabled are used for ranking different states of health and disability, then disabilities will be ranked as more serious health needs, but these rankings are open to the charge that they are biased by the ignorance of the evaluators of what it is like to live with the conditions in question. If the evaluations of the disabled themselves are used, however, the rankings are open to the charge of bias for the opposite reason—that is that the burden of disability has been unjustifiably underestimated because of the adjustment process that the disabled person has undergone. The dilemma here is in determining the appropriate evaluative standpoint, for ranking the importance of different disabilities, to avoid the potential for bias inherent in these differing perspectives.[7]

Since the preferences for different states of disability or HRQL used to determine their relative values should be informed preferences, it is natural to think that the preferences of those who actually experience the disabilities should be used. The disabled should have a more informed understanding of what it is actually like to live with the particular disability in question than a person who has never experienced the disability. Nondisabled persons will often have false beliefs about what it is like to live with particular disabilities, beliefs which should not influence the relative values assigned to the HRQL of living with those disabilities. One way of avoiding that influence is to use the preferences of persons who have the disabilities in question. But this is to miss the deeper nature of the problem caused by adaptation to disabilities.

Why should we expect nondisabled persons to assign a lower quality of life to living with various disabilities than do persons who have those disabilities, even after correcting for the prejudices and false beliefs of the nondisabled about the disabled? In evaluating how bad life

with a particular disability would be, for example, on a scale on which zero represents death and one represents a full quality of life with no impairments, nondisabled persons would ask how seriously that disability would impair their pursuit of their life plans, that is the projects and activities that give their lives value and meaning. People who have lived with that same disability for a significant period of time usually will have adjusted and adapted. They will have given up the activities and projects that are no longer possible because of their disability, and substituted other new activities and projects compatible with it. Typically, both their objective ability to pursue their plans of life, as well as their subjective satisfaction with their lives, will have been improved upon by adaptation to their disability from what nondisabled persons correctly estimate their quality of life would be with that disability.

Fundamental to understanding the difficulty of adjustment to disabilities for preference evaluation of HRQL with various disabilities is that neither the nondisabled nor the disabled need have made any mistake in their different evaluations of the quality of life with that disability. They arrive at different evaluations of the quality of life with that disability because they make those evaluations from the standpoint of different plans of life, plans of life which differ specifically in the degree to which they are limited by the disability in question. People who have, and have adapted to, the disability can look back and see that before they became disabled they too would have evaluated the quality of life with that disability as nondisabled people now do. But this provides no basis for concluding that their pre-disability evaluation of the quality of life with that disability was mistaken. The problem that I call the perspectives problem is that the nondisabled and the disabled evaluate the quality of life with the disability from two different evaluative perspectives, or plans of life. They are different because of the adaptive change in plan of life and evaluative perspective that the disabled have made to their disability, but neither evaluative perspective can be shown to be mistaken as a result of that adaptation.

When measures like the HUI or QWB are applied across different economic, ethnic, cultural, and social groups, the meaningful states of health and disability important in different groups may vary; different groups may place significantly different relative importance or value on the same states of health and disability. For example, in a setting in which most labor is manual, limitations in physical functioning will have greater importance than they would in a setting in which most individuals are engaged in non-physical, knowledge-based occupations, where certain cognitive disabilities are of greater importance. Different evaluations of health conditions and disabilities are necessary for groups with significantly different relative needs for different functional abilities, but then cross-group comparisons of health and disability, and of the relative value of health interventions in those different groups will not be possible. The health program benefits will have been measured on two different and apparently incommensurable valuational scales. These differences will be magnified when summary measures of population health are employed for international comparisons among very disparate countries.

The ethical evaluation of health differences is complicated further when great differences between and among groups or countries, such as in the level of economic development or the treatment of disadvantaged groups, are themselves unjust. For example, many measures of HRQL include some measure of subjective satisfaction or distress, which will be importantly influenced by people's expectations. In a society which has long practiced systematic discrimination against women, for example, women may not be dissatisfied with their unjustly disadvantaged state, including the health differences that result from that discrimination. The fact that victims are

sufficiently oppressed that they accept an injustice as natural should not make its effects less serious as measures of HRQL with a subjective satisfaction or distress component would imply.

A different problem with any use of QALYs or DALYs for assessing the benefits of health interventions is that they appear to discriminate against the disabled.[8] For example, an intervention that extends for 10 years the life of patients with a disability that reduces their HRQL to 0.75 would produce 7.5 QALYs for each patient receiving the intervention. By an intervention that extends for 10 years the lives of patients with an otherwise unimpaired HRQL of 1.0 would produce 10 QALYs for each patient receiving the intervention. The use of QALYs for evaluating and prioritizing life-saving interventions appears to discriminate against the disabled and place less value on their lives by assigning less value to extending their lives simply because of their disability. For this reason, some have argued that quality weights should not be used in the evaluation of saving lives, as long as the individuals in question consider their lives worth living.[9] The problem of whether QALYs unjustly discriminate against the disabled remains unresolved.

SECOND ISSUE: SHOULD THE EVALUATIVE PERSPECTIVE FOR DETERMINING RELATIVE HRQL BE INDIVIDUAL OR SOCIAL?

What does it mean to say that the evaluative perspective for determining relative quality of life of different states of disability is individual or social? Consider again the HUI scale in Table 1. It contains seven attributes, or areas of functioning, with from three to five levels of function within each attribute or area. How can the relative quality of life, or health utility, be determined for the different levels of function within the seven attributes? One natural approach is to ask individuals what their relative quality of life would be, on a scale in which one represented full, unimpaired function and zero represented death, for each level of function within each attribute. There are several different technical measures for obtaining the relative values, such as standard gambles and time trade-offs, but their details for the most part do not bear on the point of concern now.

Suppose individuals can tell us that their quality of life would be reduced from 1.00 to 0.86 at level three of attribute one if they were able to see, hear, or speak with limitations, even with equipment; they also give us utility levels for each of the other attributes and levels. Because different individuals are likely to assign somewhat different utilities or quality of life to the various attribute levels, let us suppose that we take the mean (or perhaps the median) numbers assigned by a randomly selected group of individuals. The actual health utility levels for the various attributes and levels of the HUI are shown in Table 1. Although this approach takes account of differences between individuals within a society in the relative quality of life they assign to different levels of impaired function, many of which will arise from their different social roles, the perspective for assigning relative quality of life weights or utilities remains individual in the sense with which I am concerned here. It is individual because it asks individuals how much an individual's (or their own) quality of life would be reduced if he or she suffered the various impairments of function.

This individual perspective is appropriate for a number of uses to which a summary measure of population health status might be put. It generally would be appropriate for monitoring the health status or overall burden of disease in a population, as well as across different populations, as long as the mean (or median) utility weights did not vary substantially in

the different populations. It would be appropriate also for comparing alternative health interventions for a given group of patients with a particular medical condition. More generally, the individual perspective is correct for evaluating alternative interventions, or for monitoring changes over time in health status or the burdens of disease, for the same person or group of persons.

In other contexts, the summary measure of health may be used for a quite different purpose—resource prioritization of alternative health interventions each of which would benefit not the same persons, but different persons; for example, a cost-effectiveness analysis of treatment interventions for different diseases. Employed in this context, here is an example of what I believe is one of the most ethically implausible implications of the individual perspective of the HUI. According to the HUI, the benefit of saving the life of a person who will live at a full quality of life level of 1.0 for a year, with no impairment of function in any of the seven attributes or areas of function, is equivalent in the aggregate benefit produced to keeping 20 different otherwise healthy individuals from having to use eyeglasses or a hearing aid to see or hear for a year (which has a utility level of 0.95); each produce 1 QALY. Here, the HUI is being employed in a cost-effectiveness analysis that weighs the trade-off between using limited resources to meet the different health needs of different groups of individuals. Do people who assign the utility level of 0.95 to requiring equipment to see or hear or speak mean by that assignment that saving one healthy person's life is of equal importance to keeping 20 persons from having to use eyeglasses or a hearing aid? It is highly doubtful that people are thinking of such trade-offs between or among different persons or groups when they assign utility levels to the different attribute levels in the HUI. They do not understand or intend their assignments to have those implications, and they reject in their explicit trade-offs these inferred trade-offs from their utility assignments in measures like the HUI.

This individual perspective of aggregating QALYs employing the HUI displays two distinct difficulties. The first is that it does not reflect the relative ethical importance people give to saving life in comparison to health benefits that improve the quality of life of other individuals or groups. As the example of eyeglasses and hearing aids shows, the HUI allows inferences about what people's trade-offs are between saving life and improving the quality of life, but the inferred trade-offs do not match people's explicit trade-offs between saving life and improving quality of life. QALYs calculated with measures like the HUI do not give the relative weight to saving life versus improving HRQL that people give when asked to make explicit trade-offs of that sort between different individuals or groups. I consider a different measure below that attempts to remedy this defect.

The second difficulty is that calculations of aggregate QALYs from different health interventions fail to reflect the ethical importance people place on the fact that health benefits to different individuals or groups are being traded off or prioritized. When summary measures of population health status like the HUI are used to evaluate alternative interventions that will benefit different individuals, as opposed to alternative interventions that will benefit the same individuals, issues of equity and distributive justice are raised. There are at least two broad approaches for taking account of this difference. One is to restrict the use of summary measures like QALYs calculated with instruments like the HUI to the evaluation of health interventions that serve the same individuals; for example, alternative treatments for the same patients. In this approach, we explicitly recognize that people do not believe that the social value or importance of different health interventions serving different people or patients can be determined simply by comparing the aggregate health benefits in QALYs to the different individuals served by the two

interventions. These determinations involve issues of distributive justice between persons in which it matters who gets what benefits, not just what the aggregate benefits of the different interventions are. We might use QALYs to determine the aggregate health benefits of interventions serving different individuals, but should do so only with the clear recognition that this will not tell us the overall social value people would place on these interventions. That can only be determined by also addressing concerns for equity, such as the unsolved rationing problems that Daniels discusses in his paper.

The second approach seeks to develop a measurement tool appropriate to the evaluation of interventions that serve different groups of individuals. The most prominent example is Erik Nord's "person trade-off" approach which explicitly asks people how many outcomes of one kind they consider equivalent in social value to "X" outcomes of another kind, where the outcomes can be for different groups of individuals. For example, people can be asked, for two diseases of equal initial severity, how many patients with the disease whose treatment is fully effective would be equivalent in social value to treating 100 patients with the other disease whose treatment is only partially effective and so results in a lesser health improvement for each patient treated; in this example, Nord found that people tended to give more weight to initial severity of illness than to the degree of health improvement, "saying that they would prefer the 'less effective' program even if it treated only one person more than the 'more effective' program."[10]

The person trade-off approach is designed to permit people to incorporate concerns for equity or distributive justice into their judgments about the social value of alternative health programs. There has been relatively little exploration and use of this methodology in comparison with the studies and methodological work on measures of aggregate QALYs, in part because many health policy analysts and health economists assume, often with little or no argument, that the social value of health programs is the sum of the individual utilities produced by the program. That utilitarian assumption is rejected by most philosophical work on distributive justice, as well as by the preferences ordinary people express for different health outcomes and programs. Because the main specific issues of equity, or "rationing problems," together with the possibility and difficulties of incorporating equity within the health evaluation measure, are explored in Daniels' paper for this workshop, I will not pursue the person trade-off approach further. But I do emphasize that for purposes of resource prioritization and allocation, the social approach is the proper perspective, whether by means of a methodology like the person trade-off that seeks to incorporate people's concern for equity within the measure of the social value of health programs, or by separate attention to issues of equity outside the summary measure of population health.

Within moral and political philosophy, as well as among ordinary people, debate and disagreement continue about utilitarian and nonutilitarian accounts of distributive justice. Individual approaches to the value of health programs, which measure QALYs using an instrument like the HUI, and a social approach using the person trade-off method differ in their sensitivity to whether people hold utilitarian or nonutilitarian views. The individual approach which assesses aggregate QALYs produced by different health interventions is insensitive to, and unable to reflect, people's concerns for equity and the distribution of benefits. The social approach using person trade-offs, by contrast, reflects whether different people's evaluations are utilitarian or nonutilitarian; utilitarians and nonutilitarians will make different person trade-offs, to which the person trade-off method will be sensitive.

THIRD ISSUE: DO ALL QALYS COUNT EQUALLY?

QALYs have been widely used in health care and other contexts to compare the outcomes of different resource allocations and health programs. QALYs can be used to measure the difference that a health care program makes in the expected years of life and in the quality of those years of life for the persons affected by the intervention. The QALY measure assumes that an additional year of life has the same value regardless of the age of the person who receives it, assuming that the different life years are of comparable quality. A year of life extension for an infant, a 30-year-old, and a 75-year-old all have the same value in QALYs, and, in turn, in a cost-effectiveness analysis using QALYs, assuming no difference in the quality of the year of life extension. This is compatible, of course, with using age-based quality adjustments to reflect differences in average quality of life of a population at different ages. For example, if average quality of life in a population of persons at age 85 is less than that of persons at age 45, a year of life extension for 45-year-olds would have greater value in QALYs than would a year of life extension for 85-year-olds.

In the World Bank Study, *World Development Report 1993; Investing in Health*,[11] the alternative DALY measure was developed to measure the burden of disease. Probably the most important ethical difference between QALYs and DALYs is that DALYs assign different value to a year of life extension at the same level of quality, depending on the age at which an individual receives it; in particular, life extension for individuals during their adult productive work years is given greater value than a similar period of life extension for infants and young children or the elderly. The principal justifications offered for this feature of DALYs were the different social roles that individuals typically occupy at different ages and the typical emotional, physical, and financial dependence of the very young and the elderly on individuals in their productive work years.[12]

I believe this justification of age-based differences in the value of life extension adopts an ethically problematic social (in a different sense of "social" than that used in the preceding section) perspective on the value of health care interventions that extend life, or maintain or restore functionó that is, an evaluation of the benefits *to others* of extending an individual's life, or maintaining or restoring his or her function, in addition to the benefit to that individual of doing so. This social perspective is in conflict with the usual focus in clinical contexts only on the benefits to the individuals who receive the health care interventions in question. Typical practice in health policy and public health contexts is more ambiguous on this point, since there benefits to others besides the direct recipient of the intervention are sometimes given substantial weight in the evaluation and justification of health programs. For example, with treatment programs for substance abuse, the benefits of reductions in lost workdays and in harmful effects on family members of the substance abuser. Using this social perspective is ethically problematic because it gives weight to differences between individuals in their social and economic value to others; it can be argued that this discriminates against persons with fewer dependencies and social ties and so is not ethically relevant in health care resource allocation. The social perspective justifying the DALY measure is ethically problematic, in a way the alternative QALY measure is not, if the value of health benefits for individuals should focus on the value to those individuals of the health benefits, not on the social value for others of those health benefits.

Placing different weight on life extension at different ages, however, might be justified ethically if done for different reasons. For example, Daniels has argued that because everyone can

expect to pass through the different stages of the life span, giving different value to a year of life extension at different stages in the life span need not unjustly discriminate against individuals in the way giving different weight to life extension for members of different racial, ethnic, or gender groups would unjustly discriminate.[13] Each individual can expect to pass through all the life stages in which life extension is given different value, while he or she is a member of only one race, ethnic group, and gender. Thus, use of DALYs does not constitute unjust age discrimination comparable to gender, ethnic, or racial discrimination.

Moreover, individuals and their society might choose to give lesser weight to a year of life extension beyond the normal life span than to a year of life extension before one has reached the end of the normal life span. People's plans and central long-term projects will typically be constructed to fit within the normal life span, and so the completion of these central projects will typically require reaching, but not living beyond, the normal life span.[14]

FOURTH ISSUE: WHAT LIFE EXPECTANCIES SHOULD BE USED FOR CALCULATING THE BENEFITS OF LIFE SAVING INTERVENTIONS?

There are significant differences in the life expectancies of different groups in American society, for example between genders and among racial and ethnic groups, as well as differences correlating with differences in socioeconomic status. In a broader international context the differences in life expectancies within and between different countries are much larger. Should these differences affect calculations of the life years gained by life-extending health care and public health interventions? An accurate estimation of the additional life years actually produced by specific health care or public health interventions should not ignore differences in life expectancies that are not caused by the particular condition that the health care intervention affects.

Yet the differences in life expectancy between and among different racial, ethnic, and socio-economic status groups, as well as the very large differences among life expectancies in economically developed and poor countries, are often principally a result of unjust conditions and deprivations suffered by those with lower life expectancies. It would seem only to compound those injustices to give less value to life-saving or life-extending interventions for groups with lower life expectancies caused by the unjust conditions and deprivations from which they suffer.

Differences in life expectancies between the genders, on the other hand, are believed to rest in significant part on biological differences, not on unjust social conditions. Whether the biologically based component of gender differences in life expectancies should be reflected in measures like QALYs or DALYs is more controversial. On the one hand, the lower life expectancy of men does not reflect an independent injustice, but, on the other hand, it is explicit public policy, required by law, not to take account of this gender-based difference in most calculations of pension benefits and annuity costs. The developers of the DALY explicitly chose to use a single uniform measure of life expectancy (except for the biologically based gender difference), specifically that observed in Japan which has the highest national life expectancy, to measure gains from life-saving interventions. They justified their choice in explicitly ethical terms as conforming to a principle of "treating like events as like."[15]

I note below even more briefly two other ethical issues which are important in cost-effectiveness analyses of health programs and in prioritization of health interventions and

programs; the first arises directly in constructing a summary measure of increases and decreases in population health. Space limitations preclude any exploration of the details of these two issues.

FIFTH ISSUE: SHOULD DISCOUNT RATES BE APPLIED TO HEALTH CARE BENEFITS?

It is both standard and recommended practice in cost-effectiveness analyses, in health care and elsewhere, to assume a time preference and thus to apply a discount rate to both the benefits and costs of different programs under evaluation.[16] It is important to be clear both about what precisely the ethical issue is in whether health benefits should be discounted, as well as why it is important for policy. It is not controversial that a discount rate should be applied to economic costs and economic benefits. The ethical issue is whether a discount rate should be applied directly to changes in well-being or health, and in particular to benefits in the form of increased well-being from health interventions. Is an improvement in well-being, such as a specific period of life extension, a reduction in suffering, or an improvement in function, extending, say, for one year of less value if it occurs 20 years from now than if it occurs next year?

Distant benefits are appropriately discounted when they are more uncertain than proximate benefits. Proximate benefits, such as restoration of an individual's function, also are of more value than distant benefits if they make possible a longer period of benefit by occurring sooner. But neither of these considerations require the use of a discount rate—they will be taken account of in the measurement of expected benefits of alternative interventions. The ethical question is whether an improvement in an individual's well-being is of lesser value if it occurs in the distant future than if it occurs in the immediate future, simply and only because it occurs later in time. This is a controversial issue in the literature on social discounting and my own view is that no adequate ethical justification has been offered for applying a discount rate directly to health and well-being. The avoidance of paradoxes that arise if a discount rate is applied to costs and the same discount rate is not applied to benefits, has influenced many economists to support use of the same discount rate for costs and benefits,[17] but I believe these are properly dealt with not through discounting, but instead through directly addressing the ethical issues of equity between different generations that are raised.

The policy importance of this issue is relatively straightforward in the prioritization of health care interventions. Use of a discount rate for evaluating alternative health care programs that take significantly different lengths of time to produce their benefits leads to giving an unwarranted priority to programs producing benefits more rapidly. Put differently, a program that produces benefits in health and well-being 20 years into the future will be given lower priority than an alternative health care program that produces substantially less overall improvement in health and well-being, but produces that improvement much sooner. Many public health and preventive interventions—for example, vaccination programs and changes in unhealthy behavioróreap their health benefits years into the future. If those benefits are inappropriately discounted, they will receive less priority and result in a health policy that produces fewer overall health benefits over time than could have been produced.

SIXTH ISSUE: WHAT COSTS AND BENEFITS SHOULD COUNT IN

COST-EFFECTIVENESS ANALYSES OF HEALTH PROGRAMS?

It is widely agreed that cost-effectiveness analyses in health should reflect the direct health benefits for individuals of their medical treatment, such as improving renal function or reducing joint swelling, and of public health programs, such as reducing the incidence of infectious diseases through vaccination programs. The direct costs of medical treatment and public health programs, such as the costs of health care professionals' time and of medical equipment and supplies, should also be reflected. But medical and public health interventions typically also have indirect benefits and costs. Some disease and illness principally affect adults during their working years, thereby incurring significant economic costs in lost workdays associated with the disease or illness, whereas other disease and illness principally affect either young children or the elderly who in each case are not typically employed and so do not incur lost wages or work time from illness. Should an indirect economic burden of disease of this sort be given weight in prioritizing among different health care interventions?

From a broad utilitarian perspective encompassing all effects of disease and of efforts to treat or prevent it, indirect benefits and costs are real benefits and costs, even if not direct health benefits and direct treatment costs. They should be reflected in the overall cost-effectiveness accounting of how to use scarce health resources to produce the maximum aggregate benefits. Is there ever a moral reason to ignore these indirect costs and benefits in health resource prioritization? Giving priority to the treatment of one group of patients over another because treating the first group would produce indirect benefits for others (for example, other family members who were dependent on these patients) or would reduce indirect economic costs to others (for example, the employers of these patients who incur less lost work time) could be held to fail to treat each group of patients with the equal moral concern and respect, and in particular the equal moral concern for their health care needs, that all people deserve. Instead, giving lower priority to the second group of patients simply because they are not a means to the indirect benefits produced or indirect costs saved by treating the first group of patients gives the second group of patients and their health care needs lower priority simply because they are not a means to these indirect benefits or cost savings to others. This Kantian reason for ignoring indirect benefits and costs could serve as a moral basis of the idea of "separate spheres," that is, that the purpose of health care and of public health is health and the reduction of disease, and so only these goals and effects should guide health care and public health programs.[18]

The six ethical issues discussed very briefly above are all issues involved in developing a summary measure of population health, and of changes in population health, that permits equitable evaluation of populations' health or health programs. In each case, there are important ethical and value choices to be made in constructing the measures; the choices are not merely technical, empirical, or economic; they are moral and value choices as well. There are further ethical issues in the use of these measures that Daniels takes up in his companion paper.

NOTES

1. "Considerations of Equity in Relation to Prioritization and Allocation of Health Care Resources," in *Ethics, Equity and Health For All*, eds. Z.Bankowski, J.II. Bryant, and J. Gallagher (Geneva: CIOMS, 1997).

2. G.W. Torrance, et. al., *Multi-Attribute Preference Functions for a Comprehensive Health Status Classification System*. Working Paper No. 92-18. (Hamilton, Ontario: McMaster University, Center for health Economics and Policy Analysis, 1992).
3. R.M. Kaplan and J.P. Anderson, "A General Health Policy Model: Update and Applications," *Health Services Research*. June 23 (1988) 203–235.
4. D.M. Brock. "Quality of Life Measures in Health Care and Medical Ethics." in *The Quality of Life*. eds. A. Sen and M. Nussbaum (Oxford: Oxford University Press, 1992).
5. M. Bergner, R.A. Bobbitt, W.B. Carter, and B.S. Gibson. "The Sickness Impact Profile: Development and Final Revision of a Health Status Measure." *Medical Care* 19 (1981) 787–805.
6. J.E. Ware and D.C. Sherbourne. "The MOS 36-Item Short Form Health Survey." *Medical Care* 30 (1992) 473–483.
7. D.M. Brock. "Justice and the ADA: Does Prioritizing and Rationing Health Care Discriminate Against the Disabled?" *Social Theory and Policy* 12 (1995) 159–184.
8. Brock, *ibid*.; D.C. Hadorn, "The Problem of Discrimination in Health Care Priority Setting," *Journal of the American Medical Association* 368 (1992) 1454–1459; D. Orentlicher, "Deconstructing Disability: Rationing of Health Care and Unfair Discrimination Against the Sick," *Harvard Civil Rights-Civil Liberties Law Review* 31 (1996) 49–87.
9. F.M. Kamm, *Morality/Mortality. Volume One. Death and Whom to Save from It* (Oxford: Oxford University Press, 1993).
10. E. Nord, "The Person Trade-Off Approach to Valuing Health Care Programs," Working Paper 38, National Centre for Health Program Evaluation (Fairfield Hospital, Fairfield Victoria, Australia) 7.
11. World Bank, *World Development Report 1993: Investing in Health* (Oxford: Oxford University Press, 1993).
12. C.J.L. Murray, "Qualifuing the Burden of Disease: The Technical Basis for Disability-Adjusted Life Years," in *Global comparative Assessments in the Health Sector: Disease Burden, Expenditures and Intervention Packages*, eds. C.J.L. Murray and A.D. Lopez (Geneva: World Health Organization, 1994).
13. N. Daniels, *Am I My Parents Keeper? An Essay on Justice Between the Young and the Old* (New York): Oxford University Press, 1988).
14. Daniels, op.cit: D.W. Brock, "Justice, Health Care, and the Elderly," *Philosophy and Public Affairs* 18, 3 (1989) 297–312.
15. C.J.L. Murray *op. cit.*, 7.
16. M.R. Golde, et. al. *Cost-Effectiveness in Health and Medicine* (New York: Oxford University Press, 1996), chap. 7.
17. E.B. Keeler and S. Cretin, "Discounting of Life-Saving and Other Nonmonetary Effects," *Management Science* 29 (1983) 300–306.
18. Kamm, *op.cit*., chap. 8.