RFF REPORT

MARCH 2004

Valuing Health Outcomes:

Policy Choices and Technical Issues

Executive Summary

Alan J. Krupnick

RESOURCES
FOR THE FUTURE

MISSION

Resources for the Future improves environmental and natural resource policymaking worldwide through objective social science research of the highest caliber.

As the premier independent institute dedicated exclusively to analyzing environmental, energy, and natural resource topics, Resources for the Future (RFF) gathers under one roof a unique community of scholars conducting impartial research to enable policymakers to make sound choices.

Through a half-century of scholarship, RFF has built a reputation for reasoned analysis of important problems and for developing innovative solutions to environmental challenges. RFF pioneered the research methods that allow for critical analysis of environmental and natural resource policies, enabling researchers to evaluate their true social costs and benefits.

RFF Reports address major issues of public policy in a manner designed to enrich public debate and meet the needs of policymakers for concise, impartial, and useful information and insights.

Core areas of knowledge at Resources for the Future include:

Energy, Electricity, and Climate Change
Environment and Development
Environmental Management
Food and Agriculture
Fundamental Research
Natural and Biological Resources
Public Health and the Environment
Technology and the Environment
Urban Complexities

RESOURCES FOR THE FUTURE
1616 P Street, Northwest · Washington, D.C. 20036-1400
Telephone: (202) 328-5000 · Fax: (202) 939-3460 · www.rff.org

Valuing Health Outcomes

Policy Choices and Technical Issues

Alan J. Krupnick

Executive Summary

Regulatory and nonregulatory activities by governments at the federal, state, and local levels often affect the economic and physical health of their constituents. The effectiveness, efficiency, and distributional implications of such activities are often compared through cost–benefit analysis and cost-effectiveness analysis. These analytical tools can also be used to help design policies. Their use requires measures that track and aggregate expected or realized health outcomes.

The government's choice of analytical tools varies. Some statutes (such as the Clean Air Act) are silent on the subject of cost–benefit analysis (which the Supreme Court has recently affirmed means that it cannot be used to make certain types of decisions), while others (such as the Safe Drinking Water Act) mandate its use. Nevertheless, all federal agencies are required to evaluate the costs and benefits of any regulation expected to cost $100 million or more.

The choice of health outcome measure is as varied—and as controversial. The Environmental Protection Agency (EPA) uses monetary measures of the value of a statistical life (VSL), but the EPA's specific value differs from that used by the Department of Transportation. In contrast, the National Institutes of Health and the Food and Drug Administration tend to use health indices because these measures help rationalize and prioritize health interventions under their purview.

The Office of Management and Budget (OMB) is charged with overseeing the use of these tools and in September 2003 issued new regulatory guidance in Circular A-4. This document boosts the standing of cost-effectiveness analysis in Regulatory Impact Analyses (RIAs) carried out by agencies and promises more widespread use of health indices and health effect measures in describing outcomes.

This report sorts out the assumptions underlying the alternative analytical tools and health valuation measures and informs practitioners about the implications of their choices on the effectiveness, efficiency, and equity of the policies they evaluate. The research grew out of a February 2003 conference and an April 2003 workshop, cosponsored by Resources for the Future and a host of federal agencies, at which these tools and methods were discussed and compared.

Analytical Tools

Cost–benefit analysis (CBA)

CBA is a normative accounting technique for capturing the advantages and disadvantages of a policy in monetary terms: subtracting costs from benefits yields the net benefits to society. These net benefits are measured according to the efficient allocation of society's scarce

resources and leave out by design other important but less quantifiable criteria, such as the distribution of these benefits across income classes. Policies that deliver negative net benefits are judged inferior to those that offer positive net benefits, and policies with the larger net benefits are considered superior to those with smaller net benefits, all other things being equal. While CBA focuses on aggregate measures, it can be used to track the distributional effects of a policy, such as by income group, sex, or race.

CBA offers transparency because the results can be clearly linked to the assumptions, theory, methods, and procedures used. It fosters revelation of uncertainty because its template permits the practitioner to see whether important information is missing. And it allows comparability because it attempts to capture in a single index all the features of a policy decision that affect the well-being of society, that is, in terms of the efficiency of resource allocation. In the context of this report, note that CBAs also routinely add valuation of nonhealth benefits to the analysis. This is important because what is being regulated, such as pollution, often affects more than human health.

In CBA, the monetary values of both benefits and costs are expressions of an aggregation of individual well-being, or utility. But a fundamental tenet of welfare-based valuation approaches is that one cannot make interpersonal utility comparisons and aggregate individual utility measurements. The solution embraced by CBA—the compensation criterion—evaluates policies on the basis of whether they could, in principle, lead to greater welfare: Can the winners from the policy fully compensate the losers and be at least as well off as before the policy? That such compensation hardly ever takes place, coupled with the difficulty of even identifying all winners and losers, leads to questions about using the compensation criterion to aggregate individuals' welfare into social welfare.

Applied to health policy, CBA attempts to capture preferences for alternative health states and put them in monetary terms so that they can be compared with other monetary estimates of the policy's effects. Some people, however, consider it unethical to place a value on human life. This view reflects a persistent misunderstanding about the valuation process. This process measures preferences of individuals, not researchers. These preferences are also for small changes in reducing death risks, some things people do every day—from deciding whether to have medical tests to deciding whether to cross the street against a red light to save a few seconds of waiting. However, because of technical difficulties and expense, some health endpoints are left unvalued. With benefits (and possibly costs) only partially estimated, the resulting net benefit estimates will likewise be incomplete.

Cost-effectiveness analysis (CEA)

CEA is a form of CBA in which the benefits are not monetized; therefore, net benefits cannot be calculated. Instead, one calculates regulatory costs per unit of an effectiveness measure, such as lives saved. Although CEA does not help determine whether a policy increases social welfare, like CBA, CEA can help identify the policy that achieves the specified goal with the smallest loss in social well-being, and it can help rank alternative policies—in this case, according to their cost-effectiveness. By avoiding monetization of health benefits, CEA may be less controversial than CBA, and, through the use of a single effectiveness measure, CEA can be simpler to conduct and communicate.

Those advantages come at a cost, however. The results can be misleading for social welfare, since the smallest welfare loss might not be associated with the smallest dollar cost. Also, CEA can unambiguously compare only policies that have a single outcome (such as lives saved), outcomes that move proportionally, or outcomes described by a health index. As a result, it addresses uncertainty less comprehensively and offers less comparability than CBA.

Cost-utility analysis (CUA)

Cost-utility analysis, as defined in this report, is the same as CEA except the effectiveness measure claims to represent (or be based on) utility, or individual welfare. It is used most often to consider the appropriateness of alternative medical interventions as well as to analyze health policy. Relative to CBA, the advantages and disadvantages of CUA are the same as for CEA except that CUA includes an effectiveness measure reflecting utility, albeit not in monetary terms.

Health Valuation Measures

Health indices, such as those using a quality-adjusted life year as the unit of account are based on multiplying the duration of a health state by a score reflecting the quality of the health state. Monetary measures include willingness to pay, which measures what individuals would be willing to give up to obtain health improvements, and cost-of-illness, which measures medical costs and forgone wages associated with health effects. A conversion factor to translate a health index outcome into a monetary measure, such as $/quality-adjusted life year, also appears in the literature and in practice.

Quality-adjusted life year (QALY)

The term "QALY index" was used to stand for all health indices. QALY indices are in use throughout the world, primarily to examine the effectiveness of medical interventions. Whether they are ready for and appropriate for use in any given policy setting are open questions.

The QALY approach uses the quality of a life year as the basic unit of account and aggregation. With dead represented by a score of zero and perfect health by a score of one, living five years longer would add five life years, subject to any adjustment for a less-than-perfect quality of life during those years. In general, numeric values are assigned to various health states to permit morbidity effects (such as severity and types of illness) to be combined with mortality effects (or likelihood of death) to develop an aggregated measure of health outcomes. For example, a year of extreme pain may be valued at 0.5. A basic assumption is that the QALY values are additive, so that a treatment eliminating extreme pain for one year for two individuals (2 x 0.5) is equivalent to a treatment that adds one healthy year of life. Life years are treated equally for all individuals, implying that a single healthy year is weighted the same regardless of age or income.

The health states are based on specific symptoms or general functionality, such as chronic limitations in one's motor functions. A crucial decision for eliciting scores is whose opinions will be sought—experts, health care professionals, affected groups, or the general population.

Willingness to pay (WTP)

The WTP approach is based on the trade-offs that individuals make (or think they would make) between health and wealth. Such trade-offs in daily life are easily recognized and as easily quantified: we may take a riskier job if higher pay will compensate us for the greater risk. WTP health valuation studies attempt to make such preferences explicit by either uncovering the trade-offs people actually make (revealed preference), as in the job choice example, or presenting them with realistic but hypothetical choices and eliciting their preferences (stated preference). WTP studies are used around the world (and particularly in the United States) in cost–benefit analyses of government regulations.

Cost-of-illness (COI)

COI estimates typically include direct medical expenditures, forgone wages, and lost household services associated with illness and premature death. Also known as the human-capital approach, COI does not purport to be a measure of individual or social welfare, since it excludes such intangibles as pain and suffering. Its advantage is its simplicity. COI estimates generally supplement WTP estimates in cost–benefit analyses of government regulations where WTP estimates are missing.

$/QALY

Another approach used to analyze the benefits of policies (or more often medical treatments) is to monetize the QALY estimate of effectiveness. QALYs are converted to dollars generally using a single $/QALY factor and then the resulting monetary estimate of benefits can be used in a CBA. These numbers are generally taken from studies that place ceilings on the cost-effectiveness of various medical interventions. Others have attempted to use value of a statistical life (VSL) estimates to derive a value of a QALY. It was suggested at the conference that these approaches as they have been used to date are not theoretically sound, in part because individuals cannot be expected to have a constant rate of substitution between QALYs and wealth.

Policy-Level Choices

In choosing their metrics, practitioners and the policymakers receiving their work need to be aware of the assumptions, both explicit and implicit, that they are accepting. Some assumptions are not amenable to technical solutions; rather, they lie more in the domain of ethicists and philosophers. Chapter Three takes up these policy-level choices.

Efficiency

Efficiency—meaning better allocation of resources—has two dimensions in the context of regulatory analysis. The first is the normative dimension, that is, does the regulation generate positive public benefits and does this particular regulatory design generate the largest public benefits? The second is a relative dimension, that is, does this regulation rank highest in terms of incurring the lowest social cost per chosen measure of effectiveness? Under the first perspective, CBA using WTP measures of value are favored. Under the second perspective, the choice of type of analysis is unclear because both CEA and CBA rank alternative regulatory designs.

Equity

The tools themselves—CBA and CEA—have no equity implications. Equity concerns with respect to the valuation measures, at a minimum, may involve how age, health status, and income factor into the analysis. Health indices, unlike WTP measures, implicitly value extensions to younger people's lives and healthy people's lives more than life extensions for those who are older or infirm, primarily because of lower life expectancies in elderly populations and lower health status in infirm populations. Adjustments could be made to eliminate such biases, as has been suggested by OMB in Circular A-4 with respect to the bias against the disabled and ill. For their part, WTP measures are constrained by income, which may be regarded by some as unethical, although in practice, values averaged over all income classes are usually used in CBAs and therefore do not discriminate against any income groups.

Normative guidance or relative rankings

How much normative guidance does the decisionmaker want? CBA offers a possibility of rejecting a course of action on efficiency grounds, if it generates net social losses. CEA can provide this service only when net costs are negative. With CBA, policies can be ranked from the largest net benefit to the smallest. With CEA, they can be ranked from the smallest cost-effectiveness ratio to the largest.

Interpersonal utility

In adding up the gains (or losses) from a policy, aggregation over individuals is required. CUA assumes that individual utilities can be added by simply summing QALYs. Thus, for any given intervention, some will gain, others will lose, and the choice of intervention will be ordered for cost-effectiveness based on the net effect on QALYs and cost. Similarly, WTP measures assume individual values can be summed.

Individual versus social perspective

Welfare economics places individual preferences at the center of the "story," with government intervention to correct market failures. In this market story, consumers are sovereign, their preferences create demand for goods, and government should not interfere with this demand if it wishes to maximize social welfare—defined as the sum of individuals' welfare.

Whether individual preferences should be at the core of government activity is an open question. Individuals have preferences for what their own health states are; they also have preferences for what happens to others (in their household, in the community, and so on). Either or both of these objects of individual preferences may be important for decisionmakers to take into account. They may also want social preferences to play a role in their decisions, where such preferences are different from the sum of the individual preferences. This issue affects WTP and QALYs alike. WTP and QALY using the person trade-off approach—where respondents choose between helping individuals in one health state over those in another—can be designed to capture preferences for community health improvements.

Underlying the choice between an individual and social perspective may be, at least in part, beliefs about the reliability of individual preferences. Certainly, there is ample evidence of gaps between individual risk perceptions and scientific estimates of risk and other factors relevant to valuing health. Some of these differences may be cognitive difficulties, say in understanding

probabilities, which would argue for reducing the weight given to individual preferences. Then again, arguing in the other direction is the well-known phenomenon that individuals imbue risk preferences with many qualitative attributes, such as degree of voluntariness and dread, which lie outside of the standard probabilistic treatment of risk. Both WTP and QALY approaches are conceptually indistinguishable on this perception issue, although this issue has received far more attention by economists in the WTP literature.

Health versus utility

What measure is appropriate to maximize in distinguishing among regulations—health or social welfare? If an aggregate measure of health changes associated with a regulatory design is the preferred measure of effectiveness, then QALYs are favored. Needless to say, describing changes in social welfare, not just health, would seem more important in a policy context, and this view would favor WTP measures.

Avoiding controversy

Using monetary measures of health effects such as WTP, particularly where premature mortality is at issue, can be more controversial than using either physical effects or QALYs (although conferees agreed that this controversy was unwarranted and is largely the result of a misunderstanding of what is being valued and how it is being valued). However, the use of physical criteria does not permit aggregation with other health or nonhealth effects. Using QALYs does eliminate the need to express benefits in monetary terms, but this merely postpones the problem, since eventually a decision must be made about whether to spend a given amount of money to save a specified number of QALYs.

Completeness

Developing WTP measures is more labor-intensive than developing QALYs because the latter approach provides weights for many different health states or domains in a given survey, while WTP studies yield values for at most a few health endpoints at a time. Attempts to develop benefit-transfer approaches to extend the range of health effects valued by WTP generally have not been successful, but choice experiment (conjoint analysis) techniques (where attributes of a health state are valued) may prove useful for this purpose. "Scores" for health states are generally more available than WTP values, although they tend to be for endpoints that are more detailed than those typically specified in epidemiological studies used in RIAs.

Credibility

A large part of Chapter Four is devoted to analyzing the credibility of the various valuation measures. The bottom line is that, under a welfare economics paradigm, WTP measures are theoretically more credible than QALYs. QALYs provide a valid utility measure only under very restrictive conditions. Within the WTP literature, much attention has been devoted to validating the credibility of WTP measures. Within the QALY literature, however, treatment of the issue of credibility (at least according to critics) has been far less extensive, though this is a flaw in the literature rather than the measure itself. Within the monetary measures, those based on willingness to pay using stated preference techniques are beginning to be viewed as more credible for use in evaluation of policy interventions related to health than those based on revealed

preference techniques, either because of the paucity of the latter or because the market behavior being studied is too far removed from the policy context. Typically, WTP measures derived from revealed preference (RP) techniques are based on wage-risk tradeoffs, which differ in context from most environmental and health risks in terms of the populations and the nature of the risk at hand. Stated preference (SP) methods can be more easily tailored to the policy context, thus potentially providing more credible estimates for policy evaluation. A panel of QALY experts and others to be convened by the Institute of Medicine later in 2004 will be attempting to sort out the advantages and disadvantages of various health indices for use in the policy context.

Consistency

Consistency can be assessed along two dimensions: consistency of the approaches used in studies and consistency in values across studies. WTP measures tend to be reasonably consistent regarding approaches—the two approaches in use (SP and RP) are derived from the same economic theory, and both approaches have been standardized (and improved upon) within the literature. The QALY literature also recognizes a number of approaches to calculating health status scores, but these approaches do not descend from the same theoretical origins. These include standard gamble (SG), which asks individuals to give the probability of death that would make them indifferent to a certain health state; time trade-off (TTO), which asks how much time—that is, how many years of life—people would trade to forgo certain symptoms; and ratings scale (RS), in which respondents are asked to simply rate various conditions on a numeric scale. While SG is based on expected utility theory, TTO and RS are not.

Transparency

Perhaps the least transparent of the WTP measures is the VSL. VSL measures are frequently misinterpreted as representing a market value for human life, rather than their true meaning, which is a value of a *statistical* life, derived by aggregating individuals' willingness to pay for small changes in risk. The metric, then, is representative of choices and trade-offs individuals make (or say they would make) in the face of risk. WTP measures for acute effects are more transparent in that they are simply the average willingness to pay to avoid a case of an illness.

On one level, QALYs are more transparent than willingness-to-pay measures, as they are simply the product of two components: a health state score and its duration. What is less transparent about QALYs, however, is the information underlying the health state scores. Such scores can be taken from a number of different indices, which in turn are developed using a number of different approaches for deriving preference weights (RS, TTO, SG) that have implications for health state scores (just as using RP or SP methods for WTP has implications for VSL values).

Time preferences

Another issue is how to account for the incidence of costs and benefits over time. Discounting the future is appropriate because it is perceived that getting something today is worth more than getting it later. Both costs and benefits should be expressed in terms of present discounted value, which involves applying discount rates to future costs and benefits. QALY gains should also be discounted, depending on their time of realization.

Consensus among economists once was that the same rate should be used to discount both costs and benefits. However, some health economists have now concluded that there is no com-

pelling basis for that approach. In the OMB guidelines, it is said that costs can be discounted using the real market rate of interest and benefits using the rate of time preference, generally from 1% to 5%.

Technical Choices

This report examines from a technical perspective the advantages, disadvantages, and underlying assumptions associated with QALYs, WTP, COI, and $/QALY measures. Comparisons between WTP and QALY measures are summarized in Table ES-1, according to a list of attributes identified as desirable during the Valuing Health Outcomes conference. The criteria for judging these measures include different types of validity (criterion, context, convergent, construct, and content validity); comprehensiveness; ease of application; costs of developing estimates; how well uncertainty is addressed; whether averting behavior is captured; whether qualitative risk attributes are included; and whether these measures bias choices toward certain groups. This detailed treatment is provided because agencies tend to be familiar with and use one measure to the exclusion of the other.

Those looking for a clear winner in the WTP versus QALYs comparison will be disappointed. According to experts at the conference, WTP studies can provide reasonable and credible social welfare-based estimates of value for some health endpoints, but not for others. The labor market studies provide a particularly robust set of studies on the VSL, with a growing body of contingent valuation method (CVM) studies on this health endpoint, which is generally considered to be the most important.

The literature on health effects valued with QALY indices is extensive and covers a wide variety of health endpoints, although these do not necessarily match endpoints appearing in RIAs. These indices enjoy wide acceptance by the medical community for discriminating the efficacy of alternative medical interventions. Judging on the comprehensiveness of estimates, QALYs do better than WTP, generally because any given survey to develop weights covers many types of health states that can be repackaged as a particular health effect is redefined. In general, WTP methods only apply to one health effect at a time, but newer studies taking the choice experiment approach promise to develop "prices" for a variety of health attributes. This approach, also called conjoint analysis, asks individuals to choose among different attributes such as health states. However, another important comparison is between the credibility of the underlying weights on health states in calculating QALYs and the credibility of the WTP estimates for individual health conditions. On this dimension, the literature on refining WTP approaches to improve their credibility is more advanced.

Concerning the QALY measures in the RIA context, some of the health indices do not pass this test and there is ambiguity and controversy about the others. Specifically, using indices based on the standard gamble to develop weights are favored because they incorporate the notion of trade-offs and some notions of risk. On the other hand, indices based on the person trade-off approach to weighting also have some appeal, as this weighting approach is the only one that may address individual preferences for effects to the community, in contrast to the standard gamble, which is concerned with individual preferences for one's own health. Not surprisingly, most studies use the simpler, less defensible indices (as measures of utility) because they are easier to understand and use when scoring aspects of disease.

TABLE E-1

Comparisons of Health Valuation Measures for Technical Attributes

Attributes	Quality-Adjusted Life Year (QALY)	Willingness to Pay (WTP)
Criterion Validity		
Tested against conditions for preferences to represent utility	Key assumptions violated by individuals, but may perform better in the aggregate.	Performs well.
Comparison to actual choices	Standard gamble (SG) scores predict treatment choices	Concern over hypothetical bias for stated preference (SP) studies; difficult to make head-to-head comparisons of SP with actual choices
Context Validity	SG does fairly well in invoking trade-offs, but not in context of reduced health risks; person trade-off (PTO) reflects community-level choices; health domains/states defined on medical interventions may not match health outcomes relevant for policy interventions.	Performs well; however, most health valuation studies are for individual preferences rather than community preferences.
Convergent Validity	Differences in preference weights by approach; SG is the only utility-consistent approach and depends on cardinal utility assumption, but is insensitive to changes in health status.	Differences in revealed preference (RP) and stated preference (SP); SP has potential to better match choice context.
Construct Validity	Focus is more on testing validity of indices than validity of weights. Weights are sensitive to duration of effect, violating independence assumption. Difficult for people to make SG tradeoffs. Duration estimates often unreliable or ad hoc. Yet, QALY indices can predict medical consumption.	Performs well, except proportionality to scope/scale for contingent valuation method (CVM).
Content Validity	Critics charge little attention given to "weights" surveys except in the construction of health state descriptions. Proponents say there is extensive work on this topic.	Major thrust of SP literature
Comprehensiveness	More comprehensive than WTP, but for health only. Combines mortality and morbidity.	Less comprehensive than QALYs, but covers more than health; doesn't combine mortality and morbidity.
Ease of Application	Easy	Easy
Cost	Cheap to apply, but getting weights is expensive (though only a one-time effort).	Cheap to apply, but getting unit values is more expensive per endpoint than QALYs. Presumption is that measures have to be estimated for each health effect–duration combination and by context, but research approaches are changing.
Address Uncertainty in Weights (QALYs)/Prices (WTP)	Relatively little attention here, only in sensitivity analysis. Uncertainty in duration of health states not addressed.	Yes
Recognizes Avoidance Behavior	No	Yes
Inclusion of Qualitative Elements of Risk	Embedded in preferences to unknown degree.	Embedded in preferences to unknown degree; beginning to be an object of research.

WTP measures may be better than QALYs at capturing preferences regarding acute health effects and can, at least in theory, capture qualitative attributes of risk (voluntariness, dread, and so on) that are not quantified in standard surveys to derive weights used in QALY indices. WTP measures can also be applied in consistent fashion to nonhealth effects, such as the effects of emissions on both ecosystems and childhood asthma. For QALYs to incorporate nonhealth effects in a CEA, the effects would have to be monetized using WTP techniques and then subtracted from costs—a confusing hybridization.

Many of the shortcomings of the QALY literature could be remedied by using best practices or even reforming some practices. Two examples of remedies are better for describing uncertainty of the scores and performing studies to better test the credibility of the weights. The WTP literature expends great effort on these features, routinely reporting uncertainty of the WTP estimates and performing a variety of content validity tests.

Improvements in WTP methods as well as research that could integrate the two approaches also are addressed in this report.

Concluding Thoughts

Regulatory decisionmaking by agencies will always be complicated, as decisionmakers evaluate incomplete and uncertain information while seeking to meet their legislative requirements, respond to stakeholders, and take their own readings about what is best for society. The new, more complex RIAs that will be produced under Circular A-4 could lead to better decisionmaking if they produce multifaceted, more complete results.

More information is not always better, however, and unfortunately WTP and QALY measures cannot be unambiguously ranked in their usefulness for policy. Such rankings would depend on the policymaker's philosophical choices and on the relative weight given to the various technical criteria. Similarly, despite CBA's advantages in measuring and ranking regulatory outcomes, the gaps in valuation and other practical problems posed by CBA that are not present in CEA would discourage sole reliance on the former method. Researching to improve all the tools and measures, evaluating each new RIA for how it approaches these issues, and examining how decisions are influenced by RIAs before and after implementation of the new guidelines would make improvements in decisionmaking more likely.

■ ■ ■