2
Theoretical Foundations of Cost-Effectiveness Analysis

A.M. GARBER, M.C. WEINSTEIN, G.W. TORRANCE, and M.S. KAMLET

Cost-effectiveness analysis (CEA) informs resource allocation decisions in health and medicine: How well it does so depends on the comparability and consistency of analyses of diverse interventions. But even a cursory examination of the literature reveals that investigators have made different assumptions about such issues as which costs and effects should be included in the analysis, which rate (if any) should be used to discount health effects that occur in the future, and the ways in which the cost of people's time should be incorporated. In the absence of uniform methods and perspectives—or of time-consuming efforts to reconstruct analyses that have used disparate methods—the results of different analyses cannot be compared.

If cost-effectiveness studies adhered to a fixed set of methodological standards, such problems might disappear. But why should one set of standards be adopted in preference to others? One way to answer this question is to seek consistency with a theoretical foundation that is broadly acceptable and informative. Such a theoretical construct, if followed through to its logical consequences, would have specific implications for the structure of CEA. An examination of the theoretical foundations of CEA will potentially resolve controversies and assist in the development of standards. This chapter describes possible theoretical foundations of CEA and their implications for the performance and interpretation of analyses.

Historically, there is no single theoretical foundation for CEA. Its roots can be traced to a variety of sources, prominent among them such fields as decision analysis and operations research. Many tools of CEA, such as the optimization techniques required for its application and the instruments developed to measure health-related quality of life, reflect the contributions of researchers of diverse backgrounds. Indeed, it might be said that CEA developed as an applied engineering technique for allocating resources.

Only recently (see, for example, Garber and Phelps, 1995) have economists sought to graft CEA to theoretical roots in welfare economics.

What kind of theory can serve as a foundation for CEA? Consider firstwhat one may mean by a theory. A theory can be defined as a coherent group of general propositions or principles (*Random House Dictionary of the English Language*, College Edition, 1968). A theory of decision making can be (1) *descriptive* if its objective is to explain phenomena or (2) *normative* if its objective is to define a standard of correctness or a norm. To the extent that CEA is designed to be a practical tool for achieving societal goals, we believe that such a theory must be normative. We do not claim that CEA adequately *describes* the behavior of health care decision makers; if it did, it would not be needed. Hence, the following discussion focuses on normative theory underlying CEA.

Perhaps no theoretical foundation can answer all of the questions that arise in setting policies for the allocation of health care. In this chapter we emphasize welfare economics as a theoretical foundation for CEA. We do so because welfare economics represents a comprehensive framework that provides answers to more methodologic questions that arise in decisions from the "societal perspective" than do any alternatives. We acknowledge, however, that CEA can be based on first principles outside of welfare economics and, therefore, that not *all* of the principles of welfare economics are essential to the practice of CEA.

The particular advantage of the welfare-theoretic framework, however—and the basis for our reliance on this theoretical foundation—is that it can inform specific issues in the application of CEA from the "societal perspective." Welfare economics provides guidance on such elements of CEA as how *society* should value resource costs and choose a discount rate for evaluation. This is not true, for example, for optimization techniques—themselves based on theoretical principles from applied mathematics. Optimization techniques are essential to any application of CEA, but they address the question of which approach is best *if* one adopts a particular decision-making perspective in which the constrained resources and the decision maker's objectives are explicit. They cannot directly answer questions that require reference to a fundamental set of values.

Despite its appeal as a comprehensive framework, the values implicit in welfare economics are not shared by all decision makers, even those working from the societal perspective. Hence, on some matters it may be preferable to depart from the recommendations of welfare economics to accommodate alternative formulations of social goals regarding health and health care. We return to alternative, "extra-welfarist" perspectives later in the chapter.

What Is Cost-Effectiveness Analysis?

Cost-effectiveness analysis is a method designed to assess the comparative impacts of expenditures on different health interventions. As Weinstein and Stason (1977) state, it is based on the premise that "for any given level of resources available, society . . . wishes to maximize the total aggregate health benefits conferred." For example, we might wish to know whether spending a certain amount of money on a public campaign to stop smoking will have greater or lesser effect on health than spending the same amount on colorectal screening. Cost-effectiveness analysis can also be used in decision making by groups or individuals, but we focus here on resource allocations at the societal level.

The Cost-Effectiveness Ratio

The central measure used in CEA is the cost-effectiveness ratio. Implicit in the cost-effectiveness ratio is a comparison between alternatives. One alternative is the intervention under study, while the other is a suitably chosen alternative—"usual care," another intervention, or no intervention. The cost-effectiveness ratio for comparing the two alternatives is the difference in their costs divided by the difference in their effectiveness, or C/E.

The C/E ratio is essentially the incremental price of obtaining a unit health effect (such as dollars per year, or per quality-adjusted year, of life expectancy) from a given health intervention when compared with an alternative. When the intervention under study is both more effective and less costly than the alternative, it is said to *dominate* the alternative; in this situation there is no need to calculate a cost-effectiveness ratio. In the circumstances under which C/E analysis is typically performed, though, the intervention is both more costly and more effective than the alternative. Interventions that have a relatively low C/E ratio are "good buys" and would have high priority for resources. In the contemporary climate of cost-consciousness, C/E analysis can also inform decisions in which a new intervention is less costly but somewhat less effective than existing alternatives. In either case, the value of a unit of the health effect is the greatest "price," or incremental C/E ratio, that we would pay for an intervention relative to its less costly alternative.

A decision rule based on adopting all interventions with C/E ratios less than or equal to a particular value will be optimal in the following two respects: (1) the resulting set of interventions will maximize the aggregate health effect achievable by the resources used, and (2) the resulting aggregate health effect will have been achieved at the lowest possible cost. There are other ways to use cost-effectiveness ratios as well. For example, they can be used to provide information to consumers about the relative values of alternative health interventions. As discussed in Chapter 1, there may be contexts in which optimization strictly according to cost-effectiveness ratios may not be ethically acceptable owing to concerns about distributive fairness, but in which knowledge of the ratios may be informative nonetheless.

Cost-Effectiveness Analysis and Cost-Benefit Analysis

Cost-benefit analysis (CBA) is similar to cost-effectiveness analysis in many respects but has a closer and better-established connection with welfare economics. The usual cost-benefit criterion from program evaluation in CBA, that the benefits of a program exceed its costs, leads to decisions that meet the requirements for an "optimal" solution under the welfare-economic framework.

Because of CBA's explicit grounding in welfare-economic principles, it is natural to ask why one would use cost-effectiveness rather than cost-benefit analysis if one wants to build from a welfare-economic foundation. Our interest in cost-effectiveness analysis derives largely from its broad acceptance within the health care field, in contrast to the skepticism that often greets cost-benefit analyses in that arena.

It is the distinguishing feature of CBA that offends some sensibilities: In CBA, the benefit of the health intervention is expressed in dollar terms rather than in terms of a nonmonetary effectiveness measure (Kamlet, 1992). The monetary measurement is obtained by estimating individuals' willingness to pay for life-saving or health-improving interventions, a measure that inherently favors the wealthy over the poor.[1] It is thus the dependence of CBA on the monetary valuation of health benefit and the method for obtaining this estimate that have motivated the reliance on CEA in the field of health and medicine.

The valuation requirement for CBA is both its greatest disadvantage and its greatest strength. It presents a difficult measurement challenge, requiring the dollar valuation of all health outcomes of importance, including changes in pain, suffering, functional status, and mortality. The valuation exercise is so daunting that few analyses attempt it. But because CBA values health in dollars rather than in units of health outcomes, it entails no distinctions between cost and effect, input or outcome. Perhaps more importantly, its scope of application is broader than that of CEA. CEA can only compare interventions whose benefits are measured in the same units of effectiveness. Thus CEA cannot be used to inform decisions about how much we should spend on housing, food, or education in relation to health care. At least in principle, CBA can handle such disparate comparisons.

For those who are uncomfortable about attaching dollar valuations to health outcomes such as life expectancy, CEA offers much of the same information. Often the two techniques will lead to similar or identical decisions concerning the allocation of health resources, so the distinction may be more important for the sake of appearance than for its practical consequences (Phelps and Mushlin, 1991).

A Metric of Health Effect: Quality-Adjusted Life Years

It may appear that CEA cannot even be used to compare interventions whose effects on health are qualitatively different, such as prevention of coronary artery disease and

treatment of arthritis. However, such a comparison is possible if the measure of effectiveness is general enough to capture all of the important health dimensions of the effects of the interventions. Using the quality-adjusted life year (QALY) as the unit of effectiveness approaches this ideal within the framework of CEA, thus expanding considerably the range of application of CEA. The QALY is a measure of health outcome which assigns to each period of time a weight, ranging from 0 to 1, corresponding to the quality of life during that period, where a weight of 1 corresponds to perfect health and a weight of 0 corresponds to a health state judged equivalent to death. The number of quality-adjusted life years, then, represents the number of healthy years of life that are valued equivalently to the actual health outcome. Chapter 4 gives a more detailed description of the theory and methods of quality-adjusted life years in CEA. The following discussion assumes that health effects are measured in QALY units.

Theoretical Foundations for Valuing Individual and Social Well-Being

If the fundamental purpose of CEA is to serve as an instrument to improve well-being by improving health, the overriding question is: Under what circumstances do decisions made on the basis of CEA lead to better distributions of resources? If such circumstances are artificial or uncommon, the technique is unlikely to be broadly useful. But if the circumstances pertain approximately in the settings in which it might be applied, CEA can have great value. Even when reality and the conditions of the theoretical model fundamentally differ, an exploration of the theoretical framework can reveal how and why CEA might need to be modified to remain a valid guide to decisions under such conditions.

By describing CEA as a tool for improving general welfare, we place it squarely within the context of welfare economics. Welfare economics is concerned with the means by which we can assess the desirability—from the societal point of view—of alternative allocations of resources. The central problem of welfare economics has been described by Arrow (1963) as "achieving a social maximum derived from individual desires." Welfare economics is based on the assumptions (1) that individuals maximize a well-defined *preference function* (in other words, their "utility" or sense of well-being depends on, among other things, material consumption, and the utility or preference function follows certain conditions of rationality and logical consistency), and (2) that the overall welfare of society is a function of these individual preferences. Much of the literature of welfare economics is concerned with developing criteria to determine whether a program improves the welfare of the affected population.

To make that determination, then, it is necessary to first measure well-being at the individual level and then aggregate individual well-being to measure welfare at the societal level.

Individual Utility Maximization

The starting point of economic theory, including welfare economics, is the behavior of individuals and the implications of individual economic behavior for interactions of groups of people in markets. Individuals are assumed to have well-defined preferences. These preferences are represented by individuals' utility functions, which relate their well-being to their levels of consumption of a number of goods and services.

The simplest economic models pertain to a world of complete certainty. Prices are known, there are no random events, and all information is freely available to everyone. These conditions bear little relation to the usual circumstances of health and medical care. Disease and its treatment have at their core substantial uncertainty. Kenneth Arrow's classic essay (1963) on the welfare economics of medical care claims that many of the distinguishing characteristics of health services delivery are direct consequences of uncertainty: the uncertainty inherent in the risk of disease and the uncertainty attending treatment—because our knowledge of its impact is imperfect.

Because both health status and the effects of health care involve pervasive uncertainty, the principal approach used in modeling preferences in cost-effectiveness analysis, as well as in other applications of health economics, has been expected utility theory. It has proven to be an extremely useful, if imperfect, descriptive framework with which to analyze individual behavior under uncertainty. When risk and uncertainty are significant factors, it has been used even more successfully to prescriptively guide decisions. According to expected utility theory, alternative actions are characterized by a set of possible outcomes and a set of probabilities corresponding to each outcome. Quantitative representations of preference, or utilities, are assigned to each possible outcome (e.g., health state) that may occur. To choose the best action, the probability of each outcome is multiplied by the utility of that outcome; these products are then summed across all possible outcomes in order to derive the expected value of utility. The numerical quantities used as utilities, then, reflect both ordinal rankings of outcomes and strength of preference for these outcomes when they are embedded in uncertain gambles. Expected utility theory is presented in many textbooks of economics (Hirshleifer and Riley, 1992), and it is at the heart of the prescriptive methodology of decision analysis (Raiffa, 1968; Holloway, 1979; Weinstein et al., 1980; Sox et al., 1988).

Valuing Individual Health Effects

Expected utility theory supplies a theoretical foundation for the quantification of effectiveness in cost-effectiveness analysis conducted at the individual level. Many analysts agree that the measure of effectiveness should reflect individual preferences under uncertainty: Specifically, the measure of health benefit to an individual should reflect the gain in expected utility for the individual. Quality-adjusted life years (QALYs) are one such measure.

The theoretical foundations of expected utility theory may be applied to answer the question: Under what circumstances can health-related utility be represented in terms of quality-adjusted life years? Pliskin et al. (1980), as modified by Johannesson et al. (1994), have shown that QALYs can be used to represent utility only if (1) individuals are willing to trade off years of life in a given health state for fewer years at an ideal health state at a constant rate, irrespective of the number of years spent in the state (the constant proportional tradeoff assumption), and if (2) individuals are indifferent among various survival curves that have the same life expectancy (they are risk neutral).[2] These assumptions may not hold in practice, but QALYs may still offer a close enough approximation to health-related utility to justify their use in cost-effectiveness analysis, especially when one views CEA as an input to, rather than a procedure for, decision making.[3]

Having defined individual health-related utility in terms of quality-adjusted life expectancy, the question of how to aggregate changes in health-related utility across individuals remains. We turn next to the issues at the level of a group or population.

The Role of Health in Determining Social Welfare

Health is an important component of individual utility, but not the sole consideration. Consumption of other goods and services, such as food, shelter, clothing, and recreation, also contributes to overall well-being. Different people may be willing to exchange other sources of utility for health at different rates. For example, a wealthy individual might be willing to reduce other consumption of nonhealth goods more sharply in order to improve health than would a poor person, who cannot afford to give up as much. One issue from the point of view of social welfare is whether to accept individual preferences for health vis-à-vis other commodities or whether to base social policy on the assumption that the goal of health policy is to maximize health.

To illustrate the distinction between these two approaches to social policy, consider a society consisting of rich people and poor people in which opportunities to provide health services to both groups are available. Suppose that society has allocated health care resources in order to maximize the aggregate number of quality-adjusted life years across the population. Now it may be that the poor people would gladly give up some of their health care (say, 100 QALYs' worth) in exchange for cash (which they can use to buy other valued items), and the rich people would gladly give up an equivalent amount of cash in order to get more health care (say, 90 QALYs' worth). Welfare economics would recognize this situation as an opportunity for a trade which could make both rich and poor people better off, according to their own preferences. Such a trade would, however, result in less aggregate health for the society as a whole. More-

over, it would leave the poor people in worse health than the rich, although they would consider themselves better off than under the initial state of affairs. Is this trade socially desirable? Neoclassical welfare economics says yes, because everyone perceives themselves as better off. However, an "extra-welfarist" perspective (Williams, 1993) might regard this trade as unacceptable because society values health as a "merit good," that is, a good which people should have regardless of their willingness to pay for it. According to the latter view, since the posttrade society has 10 fewer QALYs' worth of health than the pretrade society, society is worse off. This illustrates a fundamental difference in values between the implications of defining the output of health care in terms of its contribution to overall well-being and instead, defining it in terms of its contribution to health itself. In either case, individual preferences determine the magnitude of health improvements, but society's approach to aggregating these would be very different.

Welfare Economics as a Theoretical Foundation for CEA

In welfare economics, a *social utility function* is defined as some aggregate of individual utilities; economists view the maximization of the social utility function as the ultimate goal of any resource allocation scheme. One approach, which is frequently and incorrectly equated with the welfare-economic approach generally, is *strict utilitarianism*. The specific form of the social welfare function under strict utilitarianism is the sum of the utilities of the individuals who comprise society. But the usual reason to address social welfare in this framework is to propose or at least explore other forms of aggregation; typical measures allow for the possibility that different people should receive different weights in the social accounting. For example, greater weight might be given to the welfare of persons who are either in poor health or impoverished.

A substantial literature, spanning economics, philosophy, and political science, addresses the possible specifications of the social welfare function and the ways that such a distributive scheme might be elicited from the views of members of society. The literature suggests that there is no consensus on the specific form the social utility function should take; it appears to be impossible to select a specific weighting scheme from any universally accepted set of first principles (Sen, 1995). Consequently, much of the economic literature concerned with improvements in well-being avoids choosing weights to be attached to the utilities of different individuals. Instead it seeks less-demanding assumptions under which it is possible to make firm statements about the relative desirability of alternative resource allocations.

If there is no consensus about how individual utilities should be combined to form a social utility function, can anything useful be said about the effect of any reallocation of resources on social welfare? The concept of *Pareto optimality*, which is the benchmark used in nearly all mainstream microeconomics, has proven to be a simple but powerful guide to testing for whether a resource reallocation might improve social

welfare. A resource distribution is considered to be Pareto-optimal when any change in the distribution must make someone worse off, even if others are better off. This implies, of course, that if an allocation is not Pareto-optimal, it is possible to reallocate so as to improve at least one person's welfare without making anyone worse off. A strict criterion for deciding whether a reallocation of resources represents an improvement in welfare is closely related to this concept. If the reallocation makes at least one person better off, and no one worse off, it is said to represent a *Pareto improvement*. Thus, when the effects of a change in policy or prices on individual utilities are known, but the specific social welfare function is not, the Pareto criterion can be used to test whether social welfare is improved.

A reallocation that makes some people better off and none worse off seems unexceptionable, but unfortunately it is rarely attainable. Few public programs produce only winners; typically, funds must be raised by taxes or another mechanism that imposes costs on some people that exceed the benefits they can expect to receive. In fact, packages of programs are often constructed to enable every voter to gain in at least some dimension, while perhaps sacrificing in others—or at least to appear to offer gains to everyone—but such efforts rarely achieve unqualified success. Thus, although this criterion is extremely useful in economic theory for determining the optimality of alternative schemes for pricing, taxation, and so on, it has limited applicability in testing the consequences of real-world policy options.

A less-restrictive standard, variously called *potential Pareto improvement*, the *Kaldor-Hicks criterion*, or the *compensation test*, has been proposed to evaluate situations in which there are both gainers and losers from a reallocation. Under the compensation test a program is considered to be welfare-enhancing if the gainers are willing to pay enough for their gains in order to compensate the losers. The rationale behind this standard is that *if* there were a mechanism for such payment to occur, the program would result in an actual Pareto improvement. Cost-benefit analysis is directly based on the potential Pareto-improvement criterion. It can be shown that if a program is undertaken whose (properly measured) benefits exceed their costs, a potential Pareto improvement will occur.[4]

Central to the compensation test for potential Pareto improvement is the proposition that the appropriate measures of value are the amounts of money that individuals are willing to pay for goods and services. The compensation test is tantamount to the following thought experiment. When a program is being considered, imagine passing a hat to each member of society. Individuals who would gain from the program must put into the hat the maximum amount of money that they are willing to pay for the program. Individuals who would lose, including taxpayers who would pay a share of the cost without receiving any benefit, take from the hat the amount of money that would be just enough to compensate them for their losses or tax payments. (For this reason, willingness to pay is also called "compensating variation" in welfare economics.) If there is more money in the hat at the end than there was at the beginning, then the program represents a potential Pareto improvement.

The drawback of this approach, of course, is that the reallocation from gainers to losers may not occur. Then the desirability of a program from the societal perspective cannot be determined without reference to the distribution of welfare, and a well-defined way of combining the welfare of different people into a social welfare function must again be invoked.

The welfare-economic framework facilitates derivation of the cost-effectiveness approach from fundamental principles, and in particular clarifies the conditions under which decisions based on C/E ratios are equivalent to tests of the Kaldor-Hicks criterion. Garber and Phelps (1995) describe a set of assumptions under which rankings derived from cost-effectiveness ratios provide the optimal expenditure of health resources. One such assumption is that individual utility in any period of life is the product of two factors: the utility attached to health-related quality of life in that period (i.e., a quality weight) and the utility attached to the individual's material consumption in that period. Under this assumption, Garber and Phelps show that individuals will optimally set priorities for health care expenditures by selecting those with cost/QALY ratios less than some threshold.

In essence, this approach rests on the assumption that QALYs are a valid representation of individual utilities for health outcomes. Because of the flexibility afforded by the adjustments for health-related quality of life, in many instances this will be reasonable; although the QALY formulation appears restrictive, it represents a close approximation for a much broader set of plausible utility functions than those that can be described in precise terms as QALYs (Garber and Phelps, 1995). Sometimes, however, QALYs will not be adequate; for example, an individual with a terminal illness may place very high value on living until a particular milestone (a child's wedding, a holiday, a reunion with a relative or friend), and care less about length of life after the event. The approximation to health-related utility that QALYs offer will be inexact. However, such phenomena may be unimportant when CEA is applied at the *population* level.

Maximizing QALYs as Social Policy

As an alternative to defining social utility as an aggregate of individual utilities, special status may be given to health in the social accounting. According to this view, health per se is viewed as the output of the health care sector, and the social objective is to maximize health subject to resource constraints (Culyer, 1991; Williams, 1993).[5]

The connection with individual expected utility theory is not that individual utilities provide any normative basis for aggregation, since clearly they do not, but that individual utilities allow for the possibility of creating a social utility function based on explicitly stated societal preferences, as determined, for example, by a decision-making body or official. For example, it might be asserted, as an ethical principle, that the marginal social utility of 1 year of quality-adjusted life expectancy is equal for all

individuals. This assumption would lead to the use of aggregate QALYs as the quantity to be maximized in health resource allocation.

One approach to justifying a procedure for aggregating utilities (i.e., QALYs) appeals to a hypothetical choice situation, or "contract," among citizens who we assume are impartial because they operate behind a "veil of ignorance" (see Harsanyi, 1953, 1955). Imagine individual citizens in a state prior to their birth, uncertain of which of many prospects, including possible health scenarios, await them. Then rational individuals, seeking to make themselves as well-off as possible but blinded to the specifics of their futures, would opt for societal decision rules based on maximizing aggregate (or average) utility across the "population" of possible lives; they would choose a pure utilitarian distribution.[6] This conceptual basis for maximizing aggregate health-related utility has been described also as a "Constitutional Convention" by Kamlet (1992). If (1) deliberators behind such a veil of ignorance would choose to maximize expected utility across possible life scenarios, and (2) we assume that individual preferences for health outcomes are expressed by quality-adjusted life years, then we are led to a societal effectiveness measure equal to the sum of quality-adjusted life years gained.

Others have challenged the claim that rational citizens behind such a veil of ignorance would choose to maximize expected utility in this way. Rawls (1971)—who in any case rejects social utility as an appropriate measure of well-being for purposes of justice, appealing instead to "primary social goods"—argues that agents deliberating upon their life prospects behind a veil of ignorance should adopt as their principle of rational choice a "maximin" rule, that is, a rule which seeks to maximize the well-being of the worst-off member of society. Rawls refuses to assume that life prospects are equally probable in the absence of any information when the stakes are so high. Other contractarian theorists (Scanlon, 1982) have also argued that our moral concerns about the "separateness of persons," including the fact that the losses of some people are not compensated for by the gains of others, preclude accepting the "gamble" involved in choosing a utilitarian distribution or its specific implication here, namely, a measure of social effectiveness equal to the sum of QALYs.

It is possible to accommodate some of these worries about distributive effects, since we might aggregate individual utilities in ways other than the simple sum of QALYs. In extreme form, this could lead to a distributional principle based on maximizing the utility of the worst-off individual (the maximin rule). However, the maximin aggregation rule attaches no weight to improvements in the utility of the better-off or even average members of society. In the health context, the question is whether, behind the veil of ignorance, people would rather have increases in quality-adjusted life expectancy if their initial endowment of quality-adjusted life years turns out to be low, or if they would choose to receive the same gains in quality-adjusted life expectancy under all life scenarios. As an alternative, *changes* in QALYs could be weighted more heavily for members of society whose initial level of health is poorer (Nord, 1992). The problem is to justify any weighting scheme in a principled or morally acceptable way.

Implications of Alternative Foundations for Distributional Equity

It should be noted that, in the formulation of cost-effectiveness analysis founded upon the compensation test, the optimal cost-effectiveness threshold differs across individuals (Garber and Phelps, 1995). Wealthier individuals would spend a larger amount per QALY than poorer individuals, reflecting their greater willingness to sacrifice material consumption for increased quality of life and probability of survival. The resulting potential Pareto improvement can be converted to an actual Pareto improvement by requiring the wealthy, who would receive health interventions according to a more generous criterion, to compensate their poorer counterparts with a portion of their wealth. But because there is no guarantee that this redistribution will occur, the resulting distribution of health benefits may be unacceptable.

Would an allocation rule based on assigning equal value to all QALYs result in a more equitable distribution of welfare than an allocation rule based on the Kaldor-Hicks criterion? One cannot say. It depends on whether the transfers of wealth from rich to poor, as compensation for a greater willingness to invest in their health care, outweigh the inequities in the provision of health care based on allowing the C/E cutoff to vary by income.

Even the assumption that all QALYs are valued equally may lead to some ethically unsettling distributional implications. Applying this principle rigorously in CEAs would lead to calculations of societal benefit that give less weight to saving the lives of persons with life expectancies that are reduced because of age, race, or socioeconomic status. Similarly, the extension of lives of persons with chronic disabilities would count for fewer QALYs gained than the extension of healthy lives. From behind the veil of ignorance, perhaps this practice can be justified ethically, but some observers may find unacceptable the ethical implications of counting all QALYs equally. We return to this issue in Chapter 4.

Theoretical Foundations for Valuing Costs in CEA

The welfare-theoretic foundation of CEA facilitates resolution of numerous methodologic issues relating to the valuation of costs. Of particular importance for CEA, it provides guidance about how to assign monetary costs to the resources that are used or freed up by health care services.

The real cost to society of a resource consumed or freed up as part of a health intervention (or as a result of it) is the value of that resouce in its next best use to society. Because resources are more scarce than the needs for which they can be used, doing more of a given health service—employing more doctors or nurses, utilizing more space and equipment for hospital beds, using more chemical or biological products—means forgoing something else of value. In an ideal analysis from the societal

perspective, therefore, resources should be valued at an amount equal to their best alternative use—their opportunity cost.

Economic theory shows that if the economy exhibits certain characteristics, then the prices prevailing in the marketplace fully reflect the values for resources in alternative uses. That is, the price of a good or service equals the resource cost of producing the last unit produced, and the resource cost of the marginal unit produced equals the value of its inputs used elsewhere. A common and tractable method useful in calculating the societal opportunity cost of a health intervention in a cost-effectiveness analysis thus locates and assigns a price to each of the resources consumed or saved by the intervention. Market prices are multiplied by incremental quantities of consumer goods or inputs to health care to calculate incremental costs.

The practice of substituting market prices for value in cost-effectiveness analyses may be less than ideal for two reasons. First, the theoretical equivalence between market prices and the value of the resources consumed does not hold in many circumstances. It assumes (1) the existence of perfectly competitive[7] markets for all goods and services, (2) the absence of externalities and public goods, and (3) the absence of distorting incentives (e.g., due to insurance, subsidies, or taxes). It is generally agreed that these conditions do not hold generally in the health sector. Second, the use of market prices does not account for changes in price that may occur as a result of the implementation of an intervention. A cost-effectiveness analysis performed before widespread use of a treatment, based on existing prices, might not reflect the true marginal cost of the treatment if substantially more (or less) of that treatment were consumed.[8] For example, if a national program began to cover the costs of bone marrow transplantation, demand would likely increase, causing a price increase due to limited short-term supply. A long-run price decrease might also occur as a result of improvements in the technology over time through a learning curve.

If observed market or transaction prices are inadequate as measures of value, the analyst may need to adjust current market prices or investigate alternatives, as discussed in Chapter 6. However, the principle of using opportunity costs provides a guide for determining the value of resources consumed and for use of market prices.

Applications of Theory to Methodologic Controversies

If the goal is to define and adopt a uniform set of practices to be followed by all cost-effectiveness analyses (see Drummond et al., 1993), then an appeal to theoretical foundations may seem unnecessary. For example, investigators might reach an agreement to ignore time costs in computing cost-effectiveness ratios. There might be strong reasons to favor such an approach, not the least among them the practical difficulty of measuring and valuing time costs. But the consensus is more likely to endure and earn wide acceptance when the logic supporting it is clear and persuasive. We now discuss how economic theory can provide a logical foundation for use in analysis. We will look

at the consequences of adopting different approaches to three controversial issues in cost-effectiveness analysis—handling time costs, incorporating health care costs that occur during years of added life, and discounting future costs and health effects.

Time Costs

A complete analysis of the costs and benefits of an intervention should include all costs, including those that are due to time lost during illness or while in treatment. The need to incorporate time costs is widely accepted, yet many details about which time costs should be included in cost-effectiveness analyses and how they should be included are unresolved. Published analyses include three categories of time costs: (1) costs related to the treatment in question that involve the time of patients, their families, or others not considered to be formal health care providers; (2) costs associated with lost or impaired ability to work or to enjoy leisure activities due to morbidity; and (3) lost economic productivity due to death. Although some authors regard each of these categories as "indirect costs" of health care, we will refer to them as "time costs." An exception is the time spent by uncompensated caregivers, which will be considered to be included among the health care services costs.

Useful guidelines for handling of all three categories of costs emerge directly from the principle that time costs should be counted but not double-counted (either included as a [health] consequence or a change in monetary cost, but not both). The need to incorporate time costs follows from the motivation for performing cost-effectiveness analysis—to use limited resources as effectively as possible. Because time, like money, is a limited resource that can be put to other (valuable) uses, time should be incorporated in the analysis. Clearly two alternative interventions that are similar in every way, except that one requires more time to travel to obtain health care, are not equally desirable.

Once it is recognized that time costs must be included, the question for cost-effectiveness methodology is whether they should be included as monetary costs (i.e., in the numerator of the C/E ratio) or as decrements to utility (i.e., in the denominator). Placing the costs in both locations, of course, would amount to double-counting; if the financial implications of lost time are reflected in the utility weights assigned to health states in the calculation of QALYs, then it would be incorrect to count the lost productivity again as costs in the numerator. In that case, only the costs borne externally to the individual whose health is affected, such as frictional costs to the employer, would be counted additionally in the numerator. If, however, respondents to the utility questions are specifically instructed *not* to consider loss of income when assessing their preferences for health states, then the full time costs must be counted in the numerator. We return to the question of which time costs to place in the numerator or denominator after discussing two pertinent theoretical issues: whether and under what conditions it matters, in principle, if time costs go in the numerator or the denominator, and the conceptual basis for assigning monetary value to time costs.

Does it matter whether time costs are valued in dollars or QALYs?

Garber and Phelps (1995) show that under conditions of perfect markets, the cost-effectiveness method leads to the same decision rules for allocating health resources whether one places time costs in the numerator or the denominator of the C/E ratio. The optimal resource allocation can be achieved by comparing C/E ratios with a threshold value representing the willingness to pay for additional QALYs; interventions with C/E ratios lower than this threshold will be accepted and interventions with C/E ratios higher than this threshold will be rejected. Garber and Phelps show that, under specific conditions, the position of the C/E ratio above or below the threshold is the same whether time costs are valued in the denominator as a decrease in the number of QALYs produced or in the numerator by a dollar value. For an activity whose utility is considered equivalent to death (i.e., whose quality-of-life weight is 0), their result requires that the opportunity cost of time equals the willingness to pay for additional QALYs. For other activities—i.e., whose quality-of-life weight is positive—their result requires that the opportunity cost equal the willingness to pay to improve the quality of life from that experienced in the activity to the level corresponding to a value of unity on the QALY scale. Thus an activity that imposes no disutility (i.e., no decrement in the QALY) has zero opportunity cost.

Are these results heavily dependent on the assumptions underlying the model? As Garber and Phelps (1995) acknowledge, the two methods will not produce equivalent results if the wrong valuation of time is used in either the denominator or the numerator. For example, for reasons discussed below, wages can be used as a proxy for the opportunity cost of time under certain conditions. However, for most people work is not the equivalent of death; on a scale from death to unrestricted leisure in full health, working while otherwise healthy might be assigned a relatively high weight. Therefore, it would be incorrect simply to subtract time spent in a doctor's waiting room from the number of QALYs gained in the denominator of the C/E ratio, unless that time was considered to be equivalent to death. (See Chapter 4 for a discussion of the meaning and sources of health-related quality-of-life weights.) In the numerator, the wage rate would understate the true opportunity cost of time if some of the compensation for the work does not take the form of wages. For example, a manager might accept a lower salary if it meant that she would get a corner office, extensive secretarial support, and flexibility in work hours; a machinist might decline more lucrative job opportunities to take a position that included substantial on-the-job training and offered better opportunities for future advancement. The disparity between wages and opportunity costs poses a challenge that must be surmounted if time costs are to be included in the numerator of the C/E ratio.

Other deviations from the underlying assumptions of perfect markets can mean that health interventions will be ranked differently if time costs are placed in the numerator rather than in the denominator of the C/E ratio. However, under the same circumstances C/E ratios may no longer be valid guides to the alternative ranking of interventions. For example, income taxes cause wages to deviate from opportunity costs. Leisure time

is not taxed, and the worker deciding how many hours to work considers *after-tax* wages, but the employer bears the full cost of the pretax wages. Tax rates and subsidies that differ across people and across inputs into health care greatly complicate the determination of the socially optimal types and levels of medical interventions to use.

The Garber-Phelps model refers to an individual allocating his or her own lifetime resources, and not to resource allocation at the population level. If society applies different cost-effectiveness criteria (dollars per QALY) to each individual, based on their own willingness to pay for QALYs, then the conditions leading to the equivalence between including time costs in the numerator and denominator may be satisfied. However, if C/E ratios are applied to populations, the two approaches will yield equivalent rankings only if the monetary value of time (and QALYs) is the same for everyone.

Thus, the theoretical framework suggests that, under certain circumstances, the choice between numerator and denominator for time costs does not matter. However, those ideal circumstances seldom apply, so choices have to be made.

Valuing time costs in monetary terms

Placing the time costs in the numerator presupposes that there is a method for converting time costs into dollar values. The dollar valuation of time is a central theme of labor economics: It is key to understanding such phenomena as unemployment, job turnover, hours of work, and retirement. The central concept, as described above in the context of valuing health resource costs, is that of *opportunity cost*, or the value of time in its best alternative use. The fundamental assumption of this literature is that people will take their opportunity cost into account when allocating their time, choosing to devote it to the activities that produce the greatest utility. They will work an extra hour, for example, if the compensation they receive exceeds the value they place on their time in other activities.

The well-established basic theory, along with variants that take into account forms of "market imperfection," have been subjected to empirical analysis and can shed some insight into the valuation of time costs for cost-effectiveness analysis. The *labor–leisure tradeoff*, which is at the heart of the theory of labor supply, illustrates the method used to value time that is not spent at work: if there is perfect competition; if workers and employers are perfectly well informed; if the worker has declining marginal utility of leisure time (i.e., the more time spent away from work, the lower the value of each incremental increase in leisure time) and diminishing marginal utility of income; and if the quantity of labor supplied in the market is continuously variable, then the worker "consumes" leisure time up to the point at which the value of an additional hour of leisure equals the (hourly) wage that he or she can receive by working.

Although only chimerical markets may satisfy all the conditions of perfect competition that underlie the simplest, idealized model of value of time, in mainstream economics all efforts to value time build upon the concept of opportunity cost. Even in settings in which market imperfections are prominent and empirical tests of the theory are infeasible, the concept has direct, concrete implications. For example, it leads to

the conclusion that the value of time is not near zero for people who are retired or otherwise out of the labor force. Economists would infer, for example, that people who choose to retire place a higher value on time spent in leisure activities or "household production" (which encompasses diverse activities such as child raising, food preparation, and cleaning) than they place on wages they could receive if they continued in their current job. Although it is not easy to infer the exact value of their time (i.e., the wage rate that would induce them to continue to work), there is no reason to believe that the number is negligible.[9]

If we accept the principle that time costs should be valued by their opportunity costs, then it follows from the theory that the time of people with differing opportunity costs should be valued differently. To the degree that wages reflect opportunity cost, the time of persons in demographic groups that tend to have lower-paying occupations would be valued less. It remains controversial whether it is ethically acceptable, for example, to value the time of women less than that of men in CEAs, although this is the implication of the theory.[10] Like the issue of whether to count the QALYs of disabled persons the same as those of nondisabled persons, ethical concerns may sometimes override the strict interpretation of the theory. We return to this question in Chapters 4 and 6.

Should time costs go in the numerator or denominator?

Despite the practical difficulties, then, there is at least a conceptual basis for valuing time costs in either dollar terms or in utility terms so that it will often be possible to choose either to place such costs in the numerator or the denominator of the C/E ratio. In some circumstances, however, it is clear that the numerator and denominator are not equally appropriate for this purpose. For example, the common practice in dealing with lost earnings due to death is unambiguous. The value of lost life is included in natural units (adjusted or unadjusted) in the denominator of the C/E ratio precisely to distinguish it from CBA, in which the value of life is monetized. Subtracting from the numerator to reflect a monetary valuation of savings due to deaths averted clearly amounts to double-counting.[11]

In contrast to the handling of lost productivity due to death, there is no convention guiding the placement of lost productivity due to morbidity in the numerator or denominator of a CEA. However, the principle of not double-counting is also relevant in considering morbidity costs. In principle, the answer depends, at least in part, on the framing of the question used to elicit utility weights for health states. If we choose the convention of eliciting utility weights for health states in such a way that the opportunity cost of morbidity time is in the denominator, this principle dictates that the monetary value of this time should not also be placed in the numerator. If we choose the opposite convention, and explicitly exclude monetary costs from consideration in the utility assessment procedure by stating that the respondent would be compensated financially for lost earnings, then these costs must be in the numerator. We consider these two situations in turn.

First, consider the situation in which the preference weights for health states are assessed under the assumption that the respondent receives full monetary compensation for the loss of work time directly resulting from impaired health status. In that case, the full societal cost of that time must be included in the numerator. If the disutility of work exactly equals the disutility of the illness, then the lost earnings can serve as a measure of the dollar value of the morbidity. Moreover, from a social perspective, the time costs are real even if the worker who is in a hospital or at home with an illness receives sick pay or disability pay; the payments the worker receives are transfer payments, a concept discussed in Chapter 6. Even though the worker may be compensated fully by these transfer payments, society is not, since the disability pay must come out of somebody's pocket. There may be additional frictional or transactions costs that result from the illness—for example, the worker who replaces another who is unable to work may receive as much compensation but be less productive in the position. The productivity loss imposes genuine social costs which, if they are large enough, should be included in the analysis (Johannesson, 1994; Koopmanschap et al., 1995). A similar approach applies to men and women who are not in the labor force—the opportunity cost must be assessed for them just as it is assessed for a worker. If the individual loses leisure time, the appropriate cost is based on the opportunity cost of their leisure time rather than the wage rate. Furthermore, the same principles apply to time costs that result from using health care services.

Second, alternatively, suppose that preference weights are assessed *without* an explicit proviso that there would be financial compensation. In this case, *part* of the cost of the lost time would already be reflected in the (dis)utility weight assigned to the health states that impair ability to work or perform valued leisure activity. The part that would *not* be reflected in the (dis)utility weights, however, pertains to the loss of time per se, independently of any effect on health status. For example, time spent traveling to health care, spent in a physician's office, or recuperating in a hospital or at home, *while otherwise unimpaired in terms of health status*, would not reduce the number of QALYs but would nonetheless represent a time cost. Such time costs would still have to be captured in the numerator, even though the effect of the impaired health status would have been reflected in the denominator.

The quality of time may vary in different activities. Variation in quality of time does not raise major conceptual difficulties, since one can define an opportunity cost for alternative states of health or activities; thus the time spent in a doctor's office may be considered more pleasant than death but less pleasant than work, in which case the dollar value of time in the doctor's office exceeds the wages lost. Appropriate adjustments can be made to the opportunity cost, if time costs are included in the numerator of the C/E ratio, or in the quality adjustments, if they are mediated by health status changes and included in the denominator.

To return to the question of which time costs should be counted as costs (in the numerator) and which should be counted as losses of health-related quality of life (in the denominator), consider two examples. The first example is a major operation which

requires a painful period of convalescence during which work is impossible. Should this period be subtracted from the number of life years or QALYs that the intervention produces? Should one place a dollar value on the time spent in recuperation and add it to the costs in the numerator? Or should some costs appear in the numerator and others in the denominator? If the utility weights are elicited under the assumption of full compensation for lost earnings, then the loss of QALYs will reflect only the pain itself and not the opportunity cost of the time. In that case, to fulfill the requirement that all resource costs be included in the analysis, the full societal cost of that time must be included in the numerator. Hence, the lost productivity (as a proxy for the opportunity cost of the time) would be included in the numerator as a component of the costs. If, alternatively, the utility weights are elicited without any implication of financial compensation for lost time, then it may be inferred that the loss of utility due to the inability to work has been captured as a loss of QALYs; to count the lost earnings in the numerator would be double-counting in this case. In the latter case, only the frictional, or transitional, costs of lost productivity should be included in the numerator (Johannesson, 1994; Koopmanschap et al., 1995).

As a second example, consider the valuation of time spent in an exercise program. If the individual values the time spent exercising as equivalent to time spent in other leisure activities, then the time cost is zero. If the time spent exercising is valued as equivalent to time spent at work, then the time cost is equal to the opportunity cost of leisure, as measured by lost earnings. If exercise is considered so onerous that it impairs health-related quality of life, then its cost would exceed the opportunity cost of leisure. The issue of numerator versus denominator rests on whether the time spent in exercise is incorporated into the calculation of QALYs. If so, and if exercise results in an impairment (or improvement) of health-related quality of life, then the opportunity cost of the time per se must still be counted in the numerator.

Thus, while the handling of time costs associated with mortality is relatively clear in CEA, the costs of other patient time consumed could be incorporated into either the numerator (as a monetary cost) or the denominator (as a decrease in QALYs). Either approach is theoretically justified, and either is feasible. The social welfare approach indicates only that these time costs, like other resource use, should be included. Further guidance is provided by the principle of not double-counting, which requires that if such costs are incorporated in the denominator they should not appear in the numerator (or vice versa) and by the motivation to achieve consistency across C/E ratios, which requires that a decision be made as to which costs are included in the numerator and the denominator of Reference Case C/E ratios.

What about time spent by family members or paid helpers, either as part of treatment or consequential to the illness? The social welfare framework clearly implies that such costs must not be ignored. The above logic implies that when unpaid work is performed by people who are not in the labor force, the value of the time should again be based on opportunity cost. Insofar as QALYs usually refer to health outcomes for the patient receiving treatment, the time costs borne by others do not appear in the QALY weights;

hence, to ensure that it is not overlooked, caregiver time that is not incorporated in the QALY measure should be valued in dollar terms and included in the numerator of the C/E ratio.

It must be noted that when CEAs are conducted from perspectives other than societal, the answer to the question of what belongs in the numerator and what belongs in the denominator could be different. For example, the "costs" from the point of view of a government agency that administers a health program might be limited to the payments it makes; if it pays for a visiting nurse, the cost will be included in the numerator of the C/E ratio, but if services are provided by a family member or the patient, the time costs might be ignored (or treated as a reduction in the number of QALYs produced). But our focus is on the social perspective, in which all costs count. Thus we cannot avoid making a decision about whether to put time costs in the numerator or the denominator of the C/E ratio.

Summary: theoretical considerations in handling time costs

1. *Mortality costs.* By definition in CEA, mortality is incorporated into either life years or QALYs as the effectiveness measure. Therefore, it would be double-counting to include a monetary value for lost life years in the numerator of the C/E ratio. To do so would be tantamount to performing a complete cost-benefit analysis in the numerator, which would render the C/E ratio meaningless.

2. *Morbidity costs and time spent receiving care.* Under specific circumstances, it can be shown that it does not matter whether time costs are incorporated in the numerator (in dollar terms) or in the denominator (in QALYs) of the C/E ratio, as long as the practice is consistent. A choice about the best practice must be made, however, both for those occasions when these circumstances are not valid and to ensure consistency across cost-effectiveness estimates.

As discussed in Chapter 4, standardization of QALYs can be achieved only if the denominator is used solely to represent health-related quality of life and not the value of time per se. If this argument is accepted, and, therefore, the value of time spent receiving health services is excluded from the denominator, then it must be placed in the numerator. This would imply that the monetary value of time spent receiving health services must be placed in the numerator of the C/E ratio. To the extent that these activities also result in an impairment of health-related quality of life which is reflected in, *and measured as*, a loss of QALYs, these reductions in QALYs can be included (in the denominator); their consequences must not, however, be doubly counted in the numerator as opportunity costs in excess of the cost of time per se. These time costs should appear regardless of whether they arise from the illness, are associated with receiving health care, or are part of recuperation.

If the full consequences of morbidity to patients, including lost productivity and leisure, are included in the QALY measure in the denominator, then they must not be

double-counted in the numerator. Under these circumstances, only the costs borne by persons other than the patient, such as frictional costs to employers and co-workers due to disability, should be included in the numerator. If the full consequences of morbidity to patients are not included in the denominator—for example, if preference weights for QALYs are assessed under the explicit assumption that the individual will be financially compensated for lost ability to work—the monetary value of that financial compensation must be included among the time costs in the numerator.

The panel's recommendations on these issues are contained in Chapters 4 and 6.

3. *Placing a dollar value on time.* When it is necessary to value time to include in the numerator, each hour should be valued at its opportunity cost. The wage rate can be used as a proxy for the opportunity cost of time for employed persons, but it does not adequately reflect the value of time for persons engaged primarily in leisure or in activities for which they are not compensated.

Unrelated Future Costs of Health Care

One of the most persistent of the unresolved issues in the application of cost-effectiveness analysis is the handling of so-called "unrelated" future costs of health care. Should health care costs that result solely from the fact that a successfully treated patient lives longer be attributed to the health intervention? Suppose, for example, that we contemplate instituting a suicide prevention program in a high school. It is highly effective and reduces teenage suicides by 50%. Students who would otherwise have died now lead lives of average length and have medical care utilization comparable to those of average persons their age. Should the future costs of health care that they consume be counted as costs of the intervention? The literature contains diametrically opposed opinions on this issue. Weinstein and Stason (1977) and Drummond et al. (1987) have argued that they should be counted while Russell (1986) has argued that they should not. Adherents to the former view argue that insofar as health care expenditures rise when people live longer, the true cost of the intervention exceeds the simple expenditures for the treatment. According to the alternative view, however, health care is but one of many costs of living longer: If we count future health care costs in added years of life, why not also count future expenditures on food, clothing, and shelter as part of the cost of the intervention?

Garber and Phelps (1995) claim that the method of accounting for truly "unrelated" future costs of health care does not matter, under the circumstances described above, in the section on time costs. In defining "unrelated" costs, they consider those future costs of care that are conditionally independent of expenditures on the intervention under consideration, as in the suicide prevention program.[12] They further assume that the future stream of health expenditures meets certain optimality conditions. These assumptions imply that the decision to include or exclude the unrelated costs merely

changes every cost-effectiveness ratio, as well as the cutoff cost-effectiveness ratio, by a constant amount. Then it does not matter whether the cost-effectiveness analysis incorporates changes in future unrelated costs of health care, as long as the practice is entirely consistent. The calculated cost-effectiveness ratio for any intervention that prolongs life, of course, will be greater if these costs are included, but the ranking of interventions will not be affected.

An important limitation of this theoretical result is that it applies only when comparing programs targeted at persons with the same remaining survival, that is, persons of the same age who are not known to differ in ways that would cause their age-specific risks of death to diverge. Otherwise, the amount by which the cost per life year will increase when these costs are included will not be constant but will depend on the age-specific pattern of health care costs.

Many interventions can be expected to alter future patterns of health care significantly, so future costs of health care cannot be considered conditionally independent of current expenditures. Failure to measure or anticipate such effects will alter not only the estimates of the effectiveness of the therapy but also the estimates of the long-term costs. It is fair to ask whether the pattern of future expenditures is ever truly unaffected by an intervention that has a large impact on longevity. Often we don't know and can't easily find out; for example, an unanticipated long-term side effect of a drug usually takes years to be discovered, and a cost-effectiveness analysis cannot be expected to reveal such consequences of treatment if clinical studies do not.

Even if there are no long-term side effects of therapy, it is possible that no costs will be truly unrelated because any treatment that has a sizable impact on mortality acts (by reducing "competing risks") to change the rates of other diseases. For example, if we were to cut heart disease death rates by a large amount, such as 50%, we would increase the prevalence of cancer solely because people who would have died of heart disease, the most common cause of death among adults, now live to die of other common diseases. If cancers are associated with more expensive treatments, and if we were to treat such costs as unrelated, we would fail to anticipate a potential increase in total health expenditures that reductions in heart disease mortality would provoke. Such arguments are quantitatively important only when an intervention is highly effective and in a population with high mortality rates, because competing risk effects essentially represent the product of two (small) mortality terms, and for most preventive interventions in the general population such effects are negligible.

To illustrate the importance of including costs of "unrelated" diseases whose incidence is affected by competing risk, consider the following hypothetical scenario. Suppose, for purposes of illustration, that all causes of death are associated with "terminal care" costs of $10,000. This cost is incurred, for example, in attempting to save a patient with a fatal heart attack or metastatic cancer. In performing a cost-effectiveness analysis of an intervention to prevent heart attacks (such as cholesterol lowering), suppose the costs of "unrelated" health care in the added years of life were excluded.

Then, the $10,000 saved by preventing a fatal heart attack would be credited to the intervention, but the $10,000 cost of dying from cancer would not be counted. Such an analysis would be predicated on an illusory saving of $10,000—the unavoidable cost of terminal care in this illustration—when in fact this cost is merely shifted by the intervention from one disease to another.

Toward a resolution of the dispute over future costs

To clarify the issues, we define three categories of induced costs that may or may not be germane in a cost-effectiveness analysis. These are: (1) costs related to the intervention, which are incurred during years of life that would have been lived without the intervention; (2) costs unrelated to the intervention, which are incurred during years of life that would have been lived without the intervention; and (3) costs that occur in years of life added (or subtracted) by the intervention. The third category may be subdivided further into three subcategories: (a) health care costs for the disease or diseases affected by the intervention, (b) health care costs for other diseases, and (c) nonhealth costs such as food, shelter, and clothing.

Costs in category (1), related diseases in the original life span, are not controversial; they must be included in the analysis. Analyses of cardiovascular prevention programs must include the costs or savings of treating heart attacks and strokes if these events are affected by the program. Likewise, costs of treating complications of treatment must be included.

Costs in category (2), unrelated health and nonhealth costs occurring during the original life span, are also not controversial. By definition, these costs are the same with and without the intervention. They cancel from the calculation of incremental cost in the numerator of the C/E ratio and, therefore, may be excluded. Furthermore, because their measurement may induce error in the estimation of costs with and without the intervention, it is usually preferable to exclude them.[13]

Category (3) is more complicated. First consider category (3)(a), costs for diseases related to the intervention but occurring in added years of life. These are typically included in cost-effectiveness analyses. For example, if a coronary bypass operation or a cholesterol-lowering intervention delays a fatal heart attack by 5 years, the costs of treating coronary events that occur during those 5 years are included. Likewise, the costs of an ongoing treatment during added years of life, such as lifelong antihypertensive therapy and its side effects, are always included.

Next consider category (3)(b), costs for diseases unrelated to the intervention and occurring in added years of life. This has been the source of much controversy. As a first step, we argue that in practice—that is, under usual circumstances—it matters whether these costs are included or excluded from all analyses if cost-effectiveness ratios are to be comparable. One important reason is that health care costs are not independent of age. Adding an 80th year of life truly costs more to maintain in good

health than adding a 20th year of life. Thus, if different interventions add years of life for different age groups, a set of C/E ratios calculated including these costs could be ranked differently from a set calculated for the same interventions if these costs were excluded.

Setting aside the fact that these costs vary with age, the Garber-Phelps model might seem to suggest that these costs could be either consistently excluded or consistently included without changing the ranking of C/E ratios. However, in order to apply this principle correctly, one would have to note that some of the costs in category (3)(a) are actually "unrelated" by the Garber-Phelps definition. For example, persons who are not candidates for a cholesterol-lowering intervention may nonetheless experience cardiovascular costs in future years of life. These age-specific "background" costs of coronary heart disease are no different conceptually from the costs of clearly unrelated diseases such as arthritis and Alzheimer's disease; they may be consistently included or consistently excluded without changing C/E rankings, but the key is consistency.

This means that if we choose to exclude the costs of "unrelated" diseases, we would also have to exclude the "unrelated" component of the costs of "related" diseases. To fail to do so would create an uneven playing field for comparing interventions into different diseases: Life-prolonging heart disease interventions would be burdened with *all* of the future costs of heart disease, while suicide prevention programs would not. There are practical and conceptual problems in disentangling the "related" and "unrelated" components of costs for "related" diseases, both of which are included in category (3)(a). The comprehensive exclusion of future "unrelated" costs would therefore be difficult, if not impossible, in practice.

We turn finally to category (3)(c), nonhealth costs in added years of life. Theoretically, these costs should be included, if health care costs in added years of life are included. However, if these nonhealth costs meet the Garber-Phelps definition of "unrelated," then their consistent inclusion or exclusion would only add or subtract a constant from the C/E ratio. Whether nonhealth costs are truly "unrelated," or at least approximately so, is an unresolved empirical question. If it were true, for example, that non–health care consumption is more closely constant with age than health care, then the constant added for consumption in each year of life at different ages would be, approximately, truly constant across ages. The question then becomes whether the Garber-Phelps result allows us to exclude these nonhealth costs without affecting the ranking of C/E ratios.[14] The Garber-Phelps argument does not, however, apply to health care costs, because they are not nearly constant with age and because a portion of the apparently related costs is, in fact, unrelated in complex and often unknown ways and would have to be excluded along with the costs of unrelated diseases in order to achieve consistency.

Like other costs and consequences, the rule of reason applies to these health care costs in added years of life. If they are small compared to the magnitude of the C/E ratio, they can be omitted without affecting the conclusions of the analysis.

Discounting

The practice of discounting health care expenditures—adjusting the dollar amounts to reflect the time value of money by assigning lower values to dollars paid in the future than to dollars paid in the present—has never been controversial. In modern economies people pay interest when they borrow money and receive interest payments when they lend or save. Thus, a dollar paid in the future is worth less than a dollar today, and for health interventions whose costs are spread over many years or whose savings are spread over many years, the practice of discounting is essential.

Discounting is more controversial, however, when it is applied to health effects. At first glance, it is not obvious why health effects that are obtained in the future should count less than immediate health effects. Is it less valuable to avert a heart attack 10 years from now, for example, than a heart attack next year, if they have the same impact on health-related quality of life and on life expectancy? Economists who work on cost-effectiveness analysis have long accepted that health effects should be discounted in the same way that the dollar expenditures are and that the same discount or interest rate should be used. Others have argued that a year of life is a year of life, whether it occurs today or in the future, and therefore health effects should not be discounted.

The social welfare foundation of CEA depends heavily on the fidelity with which an outcome measure, such as QALYs, approximates utility. QALYs are construed to have a particular functional form, usually with constant-rate discounting; a zero rate of time discount is a special case. Whether QALYs serve to approximate utilities when the personal rate of time preference is set to zero is an empirical question. If individuals place the same weight on future events as on those that will occur soon, or if they are as happy to receive a reward in the future as now, then a zero rate of time discount may be consistent with utility maximization. If they apply positive rates of time preference, the social welfare foundation only applies if the QALYs include nonzero time discounting.

An empirical question—What are appropriate rates of time preference?—thus drives the theory regarding discounting. The empirical literature on rates of time preference involves determining the rate at which individuals trade off future gains (or losses) against current gains (losses) from either their response to surveys (which ask them to consider a set of hypothetical alternatives) or from observations of their actual behavior (particularly with regard to life-saving investments or financial behavior). This issue is explored more fully in Chapter 7. In brief, estimates of personal rates of time preference vary widely, but it appears that few people have a rate of time preference near zero. In fact, much of the literature implies that the rates of time preference are implausibly large, suggesting that individuals place far greater weight on costs and benefits that occur soon as compared to delayed costs and benefits, regardless of the domain of the question (i.e., financial tradeoffs or health tradeoffs).

Even if one accepts the need for discounting, there is substantial disagreement about

whether the same rate of discount should be applied to nonmarket outcomes as to market outcomes. For example, if a person is willing to save money at a 5% annual interest rate, does it imply that the same individual will trade off the benefits of preventive therapy for current risks at the same 5% annual rate of discounting? Much of the conventional wisdom suggests that the same discount rate should be applied to all outcomes, but cogent arguments have been made that when a market good that can serve as a close substitute for a nonmarket good (such as health) is not readily available, rates of time preference need not be uniform across goods and services. Thus the welfare-economic foundations suggest that discounting is ordinarily appropriate, but it does not always provide unambiguous guidance to the particular discount rate to use.

When viewed in terms of welfare-economic foundations, the argument for discounting health effects rests on the implicit assumption that a rich and virtually continuous set of opportunities exists for exchanging money for current and future health effects. This assumption is needed so that individual marginal rates of substitution between current and future health equal societal rates of time preference. Because such opportunities to buy and sell health are not infinitely rich in an individual's lifetime, we observe wide variations between individual discount rates for health (Redelmeier and Heller, 1993), some of which are different from societal discount rates. The implications of interindividual variation in rates of time preference, and the interpretation of empirical time preferences estimates, are discussed in detail in Chapter 7.

We defer our recommendations regarding the practice of discounting in CEA until Chapter 7, where we elaborate further on the theoretical and empirical basis for discounting future costs and health consequences and for choosing a discount rate.

Conclusion

Cost-effectiveness analysis is, in the end, a pragmatic approach to measuring relative value for money in health care. It evolved as a practical response to the need to allocate limited resources for health care, not as a practical implementation of social welfare theory. Nevertheless, decision-making rules based on cost-effectiveness criteria can, under some circumstances, be directly justified on the basis of social welfare theory. Exploration of these foundations offers more than an intellectual justification for the techniques of C/E analysis because, insofar as the technique is viewed in isolation from any theoretical foundation, the answers to thorny questions in its application—such as whether to discount future health outcomes and how to account for time costs—are often arbitrary. The theoretical foundations can expose the implications of alternative responses to these questions and reveal that some practices are more useful and readily justified than others. In the subsequent chapters of this document, we describe issues that arise in different aspects of cost-effectiveness analysis. Some of the areas of uncertainty that we describe can be resolved by exploring the theoretical foundations. In other cases, the theoretical foundations help us understand what the results of cost-effectiveness analysis mean, what uses they have, and what their limitations are.

Notes

1. The "human capital" method, which values health according to the economic productivity of individuals, is still used, but it has been shown not to be consistent with welfare-economic theory (Mishan, 1988). In any case, the human capital method raises at least as many objections as the more theoretically sound willingness-to-pay method.
2. The modification by Johannesson et al. (1994) is that the constant proportional tradeoff assumption and risk neutrality should apply to discounted life years rather than to undiscounted life years.
3. It is possible to modify the analysis so that risk neutrality is not required. However, much of the power and simplicity of CEA are lost when risk neutrality is violated.
4. This standard result of public finance has been explained in a number of articles and textbooks; see, for example, Harberger (1971) or Mishan (1988).
5. In this sense, health would satisfy Rawls's (1971) definition of a "primary good."
6. Note that each individual's utility may depend on the well-being of others; thus, individual utilities could, for example, reflect altruistic values. Thus, individuals maximizing utility from behind the veil of ignorance might choose a more egalitarian distribution of well-being than if their concept of well-being were purely individualistic.
7. No individual economic agent, either seller or buyer, has sufficient market presence to affect the market price. This rules out monopoly (single seller) and oligopoly (small number of sellers), and also monopsony (single purchaser).
8. For a discussion of what to do when the price varies with the amount consumed, and of the role of taxes and other distortions, see Thompson (1980) and Gramlich (1990). Although most of their discussions are in the context of cost-benefit analysis, many of the solutions also apply to cost-effectiveness analysis.
9. The literature that addresses these issues includes work by Becker (1964), Ghez and Becker (1975), Mincer (1974), Heckman (1974), and MaCurdy (1981). Research on labor supply and the value of time are reviewed in the book by Killingsworth (1983).
10. It is possible that women receive greater nonpecuniary compensation for their time than men, for example, in the form of flexible hours, or less stressful jobs which facilitate child care responsibilities. If those factors fully explained the wage differential between men and women, then after adjusting for these factors the valuation of the opportunity cost of time might be the same for both. This remains an unresolved empirical question.
11. Although most analyses do not include lost earnings due to earlier death as part of the numerator of the cost-effectiveness analysis, some studies and government agencies list these figures as either the indirect costs of treatment or (reduced) indirect costs of disease. This practice, however, often amounts to conducting a cost-benefit analysis, since the dollar valuation of early death averted is a measure of the dollar benefit of treatment. If the analyst has such data, and if the dollar losses averted are valid measures of the benefits of prolonging life, it would seem that there is little reason to perform a cost-effectiveness analysis instead of a cost-benefit analysis.
12. Formally, costs in period 2 (C_2) are defined as "unrelated" to costs in period 1 (C_1) if $dC_2/dC_1 = 0$ (Garber and Phelps, 1995, p. 6). "Related" costs are not explicitly included in the Garber-Phelps model.

13. This issue of measurement error is particularly germane in the context of clinical trials. To include clearly unrelated diseases or unrelated costs that may be larger in magnitude than the related costs would greatly reduce the precision of estimation of the incremental cost between interventions. However, if there is uncertainty as to what costs are "related," it may be prudent to measure them nonetheless.

14. The same rationale permits the exclusion of external benefits from continued productivity during added years of life. Specifically, individuals who live longer would transfer a portion of their productivity to the rest of society, through taxes and other mechanisms, in part to finance health care. However, these benefits, like the nonhealth costs of added life expectancy, are "unrelated" in the sense of Garber and Phelps, and may therefore be excluded as long as this practice is consistently followed.

References

Arrow, K.J. 1963. Uncertainty and the welfare economics of medical care. *American Economic Review* 53:941–73.

Becker, G.S. 1964. *Human capital*. New York: National Bureau of Economic Research.

Culyer, A.J. 1991. The normative economics of health care finance and provision. In *Providing health care*, ed. A. McGuire, P. Fenn, and K. Mayhew. Oxford: Oxford University Press.

Drummond, M.F., G.L. Stoddart, and G.W. Torrance. 1987. *Methods for the economic evaluation of health care programmes*. Oxford: Oxford University Press.

Drummond, M., G. Torrance, and J. Mason. 1993. Cost-effectiveness league tables: More harm than good? *Soc Sci Med* 37:33–40.

Garber, A.M., and C.E. Phelps. 1995. Economic foundations of cost-effectiveness analysis. National Bureau of Economic Research.

Ghez, G.R., and G.S. Becker. 1975. *The allocation of time and goods over the life cycle*. New York: National Bureau of Economic Research.

Gramlich, E.M. 1990. *A guide to benefit-cost analysis*. Englewood Cliffs, NJ: Prentice-Hall.

Harberger, A.C. 1971. Three basic postulates for applied welfare economics:An interpretive essay. *J Economic Literature* 9:785–97.

Harsanyi, J.C. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *J Political Economy* 63:309–21.

Harsanyi, J.C. 1953. Cardinal utility in welfare economics and in the theory of risk taking. *J Political Economy* 61:434–35.

Heckman, J.J. 1974. Shadow prices, market wages, and labor supply. *Econometrica* 42:679–94.

Hirshleifer, J., and J.G. Riley. 1992. *The analytics of uncertainty and information*. Cambridge, England: Cambridge University Press.

Holloway, C.A. 1979. *Decision making under uncertainty: Models and choices*. Englewood Cliffs, NJ: Prentice-Hall.

Johannesson, M. 1994. The concept of cost in the economic evaluation of health care: A theoretical inquiry. *Int J Technol Assess Health Care* 10:675–82.

Johannesson, M., J.S. Pliskin, and M.C. Weinstein. 1994. A note on QALYs, time tradeoff, and discounting. *Med Decis Making* 14:188–93.

Kamlet, M.S. 1992. *The comparative benefits modeling project: A framework for cost-utility analysis of government health care programs*. Washington, DC: U.S. Department of Health and Human Services, Public Health Service.

Killingsworth, M. 1983. *Labor supply*. Cambridge: Cambridge University Press.

Koopmanschap, M.A., F.F.H. Rutten, B.M. van Ineveld, and L. van Roijen. 1995. The friction cost method for measuring indirect costs of disease. *J Health Econ* 14:171–89.

MaCurdy, T.E. 1981. An empirical model of labor supply in a life-cycle setting. *J Political Economy* 89:1059–85.

Mincer, J. 1974. *Schooling, experience, and earnings*. New York: National Bureau of Economic Research.

Mishan, E.J. 1988. *Cost-benefit analysis* 4th ed. London: Unwin Hyman.

Nord, E. 1992. An alternative to QALYs: The saved young life equivalent. *BMJ* 305:875–77.

Phelps, C.E., and A.I. Mushlin. 1991. On the (near) equivalence of cost-effectiveness and cost-benefit analyses. *Int J Technol Assess Health Care* 7:12–21.

Pliskin, J.S., D.S. Shepard, and M.C. Weinstein. 1980. Utility functions for life years and health status. *Management Science* 28:206–24.

Raiffa, H. 1968. *Decision analysis*. Reading, MA: Addison-Wesley.

Rawls, J. 1971. *A theory of justice*. Boston: Harvard University Press.

Redelmeier, D.A., and D.N. Heller. 1993. Time preferences in medical decisionmaking and cost-effectiveness analysis. *Med Decis Making* 13:212–17.

Russell, L.B. 1986. *Is prevention better than cure?* Washington, DC: Brookings Institution.

Scanlon, T.M. 1982. Contractualism and utilitarianism. In *Utilitarianism and beyond*, ed. A. Sen and B. Williams. Cambridge: Cambridge University Press.

Sen, A. 1995. Rationality and social choice. *American Economic Review* 85:1–24.

Sox, H.C., Jr., M.A. Blatt, M.C. Higgins, and K.I. Marton. 1988. *Medical decision making*. Boston: Butterworths.

Thompson, M.S. 1980. *Benefit-cost analysis for program evaluation*. Beverly Hills, CA: Sage Publications.

Weinstein, M.C., and W.B. Stason. 1977. Foundations of cost-effectiveness analysis for health and medical practices. *N Engl J Med* 296:716–21.

Weinstein, M.C., H.V. Fineberg, A.S. Elstein, H.S. Frazier, D. Neuhauser, R.R. Neutra, and B.J. McNeil. 1980. *Clinical decision analysis*. Philadelphia: W. B. Saunders Company.

Williams, A. 1993. Cost-benefit analysis: Applied welfare economics or general decision aid. In *Efficiency in the public sector*, ed. A. Williams and E. Giardina. London: Edward Elgar.