ELSEVIER Journal of Health Economics 16 (1997) 1-31

JOURNAL OF HEALTH ECONOMICS

Economic foundations of cost-effectiveness analysis

Alan M. Garber [a,b,c,*], Charles E. Phelps [c,d]

[a] *Veterans Affairs Palo Alto Health Care System, Palo Alto, CA, USA*
[b] *Department of Medicine, Stanford University, 204 Junipero Serra Boulevard, Stanford, CA 94305-8006 USA*
[c] *National Bureau of Economic Research, Inc., 204 Junipero Serra Boulevard, Stanford, CA 94305-8006 USA*
[d] *Office of the Provost, University of Rochester, 200 Administration Building, Rochester, NY 14627-0021 USA*

Received 1 January 1995; accepted 1 July 1996

Abstract

To address controversies in the application of cost-effectiveness analysis, we investigate the principles underlying the technique and discuss the implications for the evaluation of medical interventions. Using a standard von Neumann-Morgenstern utility framework, we show how a cost-effectiveness criterion can be derived to guide resource allocation decisions, and how it varies with age, gender, income level, and risk aversion. Although cost-effectiveness analysis can be a useful and powerful tool for resource allocation decisions, a uniform cost-effectiveness criterion that is applied to a heterogeneous population level is unlikely to yield Pareto-optimal resource allocations.

JEL classification: I18; D61

Keywords: Cost-effectiveness analysis; Utility maximization; Welfare economics; Cost-benefit analysis; Health care costs

[*] Corresponding author. Tel: 415/326-7160; fax: 415/328-4163; e-mail: garber@stanford.edu.

Published by Elsevier Science B.V.
PII S0167-6296(96)00506-1

1. Issues in cost–effectiveness analysis

A substantial body of literature uses cost–effectiveness (CE) analysis to rank (or at least to provide guidance about) the desirability of using alternative medical interventions. Despite its many similarities to the usual cost–benefit (CB) analysis practiced by economists, CE is widely used by practitioners who consider it "different" from CB analysis (Phelps and Mushlin, 1991). Many physicians, and others who perform CE analysis, prefer it to CB analysis because it does not require placing a dollar value on a health outcome. Typically, CE analysis describes an intervention in terms of the ratio of incremental costs per unit of incremental health effect (i.e., marginal cost/marginal health effect), in contrast to the usual CB analysis approach of describing net benefits of a project in dollars. CE studies translate the output of a medical intervention into a common denominator, such as life years saved. As techniques have emerged to "quality-adjust" those life years (see Torrance, 1986 for an excellent summary), the Quality Adjusted Life Year (QALY, pronounced kwa-lee) has become the common currency for sophisticated CE analyses. Decision analyses typically compare the cost per QALY of the intervention of interest to that for other commonly used medical interventions, arguing that the use of a new technique or technology can be justified if it has at least as favorable a cost per QALY as generally accepted interventions.

Despite the widespread use of CE analysis, we are unaware of any published formal justification of the technique on the basis of first principles. The intuitive appeal of the logic of CE (minimizing the cost of producing a given level of health, or correspondingly, maximizing the achievable level of health for a given budget) *sounds like* a familiar economic problem, and for the most part, practitioners have *assumed* that CE analysis could be a tool for utility maximization.

Yet, even within this broad level of agreement (unsupported by any formal proof of the conclusion), a number of thorny problems remain prominent in the CE literature, including: (1) How, if at all, should one include "indirect" time-related costs of treatment or benefits (e.g., work-loss prevented)? (2) Should CE estimates include "unrelated" future medical costs incurred during years of life "extended" by a current medical intervention? (3) Does the use of incremental life-years (or variants thereof) as measures of effectiveness discriminate against older persons? (4) Can one find the proper "cutoff" CE ratio in ways other than looking at the CE ratios of other commonly used medical interventions? This last issue is particularly challenging, since the range of CE ratios for common interventions is wide. Furthermore, health insurance alters the incentives for using medical care, leading to the widely held belief that our society uses too much of it (Pauly, 1986, and references therein). Although one might conclude that equalization of CE ratios at the margin is necessary for Pareto optimality, there is still the question of the proper level. We discuss each of these issues briefly, then turn to our formal model of utility maximization to answer them.

1.1. Indirect costs

Time can be a significant input into the production of health care, and convalescence from an operation or disease can require substantial time away from work and leisure. Estimates of "indirect costs" of an illness or treatment typically consist of lost wages or, for someone who is not in the labor market, imputed value of time lost. How should such costs, whether due to a health care intervention or the condition it prevents or treats, be incorporated into a CE analysis?

There is no uniformly accepted "standard practice" for incorporating such costs. Some argue that the increased life-long earnings from longer life, due to a life-saving intervention, should count as (reduced) indirect costs. Others claim that because the effectiveness measure (e.g., life years) already accounts for the value of living longer, including the increased lifetime earnings as reduced costs amounts to double-counting.

Sometimes treatment incapacitates patients for days, weeks, or even longer. Conversely, by alleviating or preventing disease, treatment may reduce a period of incapacity. Although there is widespread agreement that such effects should be incorporated in the CE analysis, some have argued that a value should be attached to the time gained or lost, and that such time costs should appear as an adjustment to the numerator of the CE ratio. Others have argued that, in a CE analysis whose effectiveness measure is sufficiently broad (e.g., QALYs), the true costs should appear as an adjustment to the effectiveness measure. Which approach should be used?

1.2. Future medical costs

The issue of including future medical costs creates a vague discomfort for the authors of many CE studies. There is no controversy about including in a CE analysis those future health care costs (or savings) that are directly attributable to the intervention. But what about health expenditures that result simply from living longer? "If we extend life," some authors argue, "then we will have to spend more for the medical care of future diseases. Therefore these medical costs are a consequence of the current treatment, and should count as a relevant cost." As Weinstein and Fineberg (1980) say in their classic text on clinical decision analysis:

> Often ignored are the costs of medical care received during extended years of life. Credit given to control of blood pressure for reducing costs associated with treatment of strokes and myocardial infarctions must be balanced against the costs for other diseases incurred during the added years of life. (p. 36)

Similarly, in their book on economic evaluation of health care programs, Drummond et al. (1987) note that:

> if hypertension therapy does extend the lives of people, there is nothing to say that they should have to be given cancer therapy at a later date. This is a decision that should be made on its own merits. ... However, in the calculation of the life extension from instituting the hypertension screening programme, if it has been assumed that those developing cancer will have the benefits of life extension from the therapies available, then consistency would demand that the costs of cancer therapy be also included. (p. 80)

In a book about disease prevention, Russell (1986) reached a wholly different conclusion, arguing that:

> if the purpose of the analysis is instead to determine whether the program is a good investment, only the costs of the preventive program should be counted. Added years of life involve added expenditures for food, clothes, and housing as well as medical care. None ... is relevant to deciding whether the program is a good investment ... (pp. 35–36)

The handling of "unrelated future medical costs" is important because they can be large enough to raise the CE ratio substantially. The impact is greatest when the intervention primarily extends life, such as for vaccines against potentially fatal contagious diseases. Several studies have highlighted the sensitivity of CE estimates to the inclusion of future medical costs. [1]

1.3. The question of age bias

A wholly separate concern arises from the usual (but not universal) practice of describing the "benefit" of the intervention in terms of life years saved, increases in life expectancy, or quality-adjusted versions thereof. Do these methods of analysis intrinsically bias the results against older persons, for whom the potential increase in life expectancy from any intervention is of necessity limited? Avorn (1984) has asserted:

> [Cost–benefit and cost–effectiveness analysis] can have major shortcomings when applied to the care of several high-risk populations, particularly the elderly ... As usually applied, these methods embody a set of hidden value

[1] In a study of influenza vaccine (Office of Technology Assessment, 1981), the cost per healthy life year (QALY) depended on the age of the vaccine recipient, ranging from $258 (for children 1–3 years of age) to $23 (for persons 45–64 years of age), when future medical costs were omitted. By contrast, when future medical costs of extended life years were included, the relevant costs per healthy life year increased by $1745 (for children 1–3) to $2084 (for adults age 45–64). Thus, including future medical costs raised the incremental CE ratio by two orders of magnitude. A later analysis by Michael A. Riddiough et al. (1983) reached similar conclusions.

> assumptions that virtually guarantee an anti-geriatric bias to their purportedly objective data. (p. 1295)

How, if at all, are these formal methods biased, and against what criterion should bias be measured?

1.4. How should one select the ''optimal'' cost–effectiveness ratio?

A final problem emerges when one considers the common uses of CE analysis for decision making. Typically, practitioners of CE analysis calculate the incremental costs and incremental effectiveness (e.g., in QALYs) of an intervention, then they compare that ratio to those found for commonly used interventions. Table 1 provides estimates from studies of various interventions, updated to 1993

Table 1
Estimated cost–effectiveness of commonly used medical interventions. (All interventions compared to ''usual care'' unless otherwise noted)

Intervention	Cost/life-year ($1993)
Low-dose lovastatin for high cholesterol [a]	
Male heart attack survivors, age 55–64, cholesterol level ≥ 250	2,158
Male heart attack survivors, age 55–64, cholesterol level < 250	2,293
Female nonsmokers, age 35–44	2,023,440
Exercise electrocardiogram as screening test [b]	
40-year-old males	124,374
40-year-old females	335,217
Hypertension screening [c]	
40-year-old males	27,519
40-year-old females	42,222
Breast cancer screening [d]	
Annual breast examination and mammography, females age 55–65	41,008
Physician advice about smoking cessation [e]	
1% quit rate, males age 45–50	3,777
Pap smear starting at age 20, continuing to 74 [f]	
Every 3 years, versus not screening	24,011
Coronary artery bypass graft [g]	
Left main coronary artery disease	8,768
Single vessel disease with moderate angina	88,087
Neonatal intensive care units [h]	
Infants 1000–1500 g	10,927
Infants 500–999 g	77,161

[a] Goldman et al. (1991); [b] Sox et al. (1989); [c] Littenberg et al. (1990); [d] Eddy (1989); [e] Cummings et al. (1989); [f] Eddy (1990); [g] Weinstein (1981); [h] Boyle et al. (1983).

dollars through use of the medical Consumer Price Index (CPI) from the calculations in the original articles.

As should be clear from examination of this table, inferring the CE ratios of "common practices" provides little guidance regarding the *optimal* CE ratio – that is, the willingness to pay for a health effect. Most practitioners of CE analysis discard interventions with CE values at the top range of a table such as this one, and conclude that interventions in the realm of $50,000 (or so) per QALY are "OK" but that more expensive technologies become more and more "out of bounds;" the $50,000 criterion is arbitrary and owes more to being a round number than to a well-formulated justification for a specific dollar value. [2]

1.5. Plan for analysis

To address these problems, we first set up a simple model of expected utility maximization (Section 2), from which we seek to answer the first question ("which costs to include"). We then generalize this model to explore a number of other issues associated with CE analysis, including whether or not the approach is internally consistent (Section 3), and the nature of an optimal lifetime medical spending plan and the estimation of an optimal CE ratio (Section 4).

Implicit in our formulation is the idea that CE analysis is applied to maximize (an aggregate of) individual utilities, such as a perfect agent might perform for an individual or a group of individuals with similar health prospects and preferences. The CE criterion could be used to determine what interventions should be covered, and in what quantity, by an insurer attempting to offer the optimal policy for a homogeneous population. Although, as we will show subsequently, the results of the analysis may not be highly sensitive to age heterogeneity, ordinarily it will not be appropriate to apply a uniform CE criterion to groups of people whose preferences or health status vary greatly. Healthy individuals and patients with a serious chronic disease may not have the same "optimal" CE ratio.

2. A simple model of cost–effectiveness

We begin with a simple three-period model in which the individual has medical care expenditures C_i and exogenously given income Y_i in period i. Utility in each period is a function of income net of medical care expenditures. All individuals are alive in period 1 and survive to period 2 with probability P_2. Given survival of period 2, they survive into period 3 with probability P_3. In this model, C_1 affects P_2, but not P_3, and C_2 affects only P_3. Medical care affects utility only by

[2] This leaves unresolved, of course, why interventions with relatively low marginal CE ratios are not expanded in scope at the expense of more-costly interventions, a shift of medical resources that would surely increase overall health absolutely (Phelps and Mushlin, 1991).

altering the survival probabilities. Thus, the individual's von Neumann–Morgenstern expected utility is:

$$E(U) = U_1(Y_1 - C_1) + P_2(C_1)U_2(Y_2 - C_2) + P_2(C_1)P_3(C_2)U_3(Y_3) \tag{1}$$

The only relevant choice variable is C_1, since C_2 is independent of C_1.[3] Following a similar analysis, we can solve for an optimal investment in C_2, which we denote as C_2^*, and corresponding outcome P_3^*. "Effectiveness" in this simple model is the increase in P_2 that results from investment in health care during period 1 (C_1), which in turn increases expected utility. Maximizing expected utility with respect to C_1 leads to an equation involving $\mathrm{d}P_2/\mathrm{d}C_1$, which (when inverted) provides the optimal incremental CE ratio. This ratio includes only current costs (not C_2 in this simple model), and comes directly from maximization of expected utility. Define $U_i' = \mathrm{d}U_i/\mathrm{d}Y_i$. Then

$$\frac{\mathrm{d}C_1}{\mathrm{d}P_2} = \frac{U_2(Y_2 - C_2^*) + P_3^* U_3(Y_3)}{U_1'} \tag{2}$$

Specifically, Eq. (2) says that an optimum is reached when the CE ratio equals the sum of future *expected utility* normalized by the marginal utility of income in period 1. This result generalizes to more periods, and allows discounting of future consumption, but this insight remains throughout such generalizations.

In a sense, this answers our first question. When one approaches the problem of defining an optimal CE ratio for medical resource decisions by using expected utility maximization principles, one can derive a CE "cutoff" for decision making that does not include future costs (C_2).

2.1. How (if at all) should one include future costs?

What are the consequences of including unrelated future costs in the CE analysis? Define expected total lifetime costs as $C^{\text{tot}} = C_1 + P_2(C_1)C_2$. Consider a CE ratio, $\mathrm{d}C^{\text{tot}}/\mathrm{d}P_2$, that includes future unrelated costs C_2 as well as the current costs included in $\mathrm{d}C_1/\mathrm{d}P_2$. From the definition of C^{tot}, we know immediately that $\mathrm{d}C^{\text{tot}}/\mathrm{d}P_2 = \mathrm{d}C_1/\mathrm{d}P_2 + C_2$. The optimization problem tells us that, if we wish to optimize using total costs, the optimal cutoff is the same as that in the previous problem *plus* C_2. One must only be consistent in practice: use the CE cutoff for decision making that corresponds to the cost accounting method one has chosen.

Are there reasons to prefer one approach over the other? The most important

[3] Formally, $\mathrm{d}C_2/\mathrm{d}C_1 = 0$. This reflects the basic principle of dynamic programming that future decisions do not depend on past decisions, given the current state. This would not be true if either P_3, Y_2, or the function U_2 were a function of C_1.

consideration is consistency, so that comparisons of the CE ratios of alternative interventions are meaningful. Insofar as it is difficult to measure unrelated future health expenditures, there is an advantage to omitting them from the analysis. However, if it is not possible to measure these costs, the assumption that they are truly unrelated to the intervention – that is, that C_2 is independent of C_1 – cannot be tested. Thus, when unrelated future costs can be identified, there may be no compelling reason to select one method over the other. However, because it frequently is not possible to determine that all changes in future health care costs are due to the "unrelated" expenditures, it is reasonable to include future costs as the default option.

2.2. Decision making with life expectancy as the "effectiveness" measure

The effectiveness measure of the preceding discussion is the probability of surviving a single period. We now show that the results also hold when we measure CE in more conventional terms, namely in terms of the cost per life year (or cost per year of life expectancy). For notational convenience, we suppress the dependence of the probability terms on prior health expenditures, and observe that in this simple model, life expectancy is given by:

$$\mathrm{LE} = 1 + P_2 + P_2 P_3 \tag{3}$$

so

$$\frac{\mathrm{dLE}}{\mathrm{d}P_2} = 1 + P_3 \tag{4}$$

and

$$\frac{\mathrm{d}C_1}{\mathrm{dLE}} = \frac{U_2(Y_2 - C_2^*) + P_3^* U_3(Y_3)}{(1 + P_3^*) U_1'} \tag{5}$$

The independence of future spending decisions (conditional on survival) from past spending decisions implies that here, too, the results will be equivalent. Since P_3^* is selected optimally by altering C_2 but is independent of C_1, Eq. (5) only differs from Eq. (2) by a multiplicative constant. We generalize this result further below, but this very simple model of medical "effectiveness" provides the basic insight for much of what follows.

The above discussion counts gains in $E(U)$ only from improvements in life expectancy, that is, through the effects of medical interventions on survival probabilities P_i. Our more general framework allows changes in utility from quality of life improvements as well. A broad literature describes methods developed to measure quality of life and to assess the value of quality improvements in terms of the increases in life expectancy that would provide equivalent increases in utility (Torrance, 1986 and references therein) in a strict von Neumann–Morgenstern framework. Indeed, these approaches provide the basic framework for computing the quality adjustments in QALY measures.

3. Does cost effectiveness provide an internally consistent way to maximize E(U)?

Common practice in CE analysis says that in order to maximize expected utility, one should adjust the intensity of all medical interventions so that they have a common CE ratio. [4] The intuition of the dictum derives from the idea that one should seek to equate the marginal benefit and marginal cost of all inputs in a productive process, as in other contexts. In this section we elaborate on our previous simple model of expected utility, incorporating more than one medical intervention, allowing those interventions to have differing effects on all future period survival probabilities, and introducing discounting and quality of life considerations. Using this model, we show that CE analysis provides a consistent criterion for selecting health interventions: the optimal CE cutoff is the same for all interventions, regardless of when they exert their effects. We thereby provide rigorous support for the common practice. We model this problem in discrete time using two interventions (*a* and *b*) available at constant cost w_a and w_b respectively, each with the ability to alter future quality of life and survival.

We first give precise definition to the three measures of effectiveness most commonly used in CE analysis. The most general measure is QALYs. [5] If P_j is the probability that a person alive the preceding period will be alive during period j, then the cumulative probability that a person is alive (the survivor function) at period i is

$$F_i = \prod_{j=1}^{i} P_j$$

.

The expected number of QALYs can be written as

$$\mathrm{QALY} = \sum_{i=1}^{N} F_i \delta^i k_i$$

where N is the maximum life span, $\delta = 1/(1+r)$ is a time discount factor, and the k_i terms represent quality adjustments. The value of k_i can range from 0 (for the worst state of health, usually assumed to be death or its equivalent) to 1 (corresponding to "perfect" health). In many preference assessment surveys, the state of perfect health is not defined, so there may be some ambiguity about its interpretation. In particular, surveys usually do not specify whether a score of 1 represents best health imaginable, or only best health for age; nor do they specify

[4] We address later the question of how one might select that ratio. We also note that corner solutions are likely to be frequent; some forms of treatment should not be used at all.

[5] Our use of the term cost–effectiveness analysis to encompass QALYs differs from the practice of some authors (Drummond et al., 1987), who attach the label "cost–utility" analysis to describe any CE study that uses QALYs as the outcome measure.

what is to be held constant – such as income and other arguments of utility. Here we interpret the k terms more broadly, so that they can incorporate general aspects of quality of life, not just those that are narrowly health-related. Each k_i term is the expected value of quality adjustments for all possible states of health in period i; if $\delta = 1$, 2 years of life in which $k_i = 0.5$ contribute the same number of QALYs as 1 year in which $k_i = 1$. The other two commonly used measures of effectiveness are special cases of QALYs. The simplest measure, life expectancy, sets $k_i = \delta = 1$ for all i; discounted life expectancy sets $k_i = 1$ for all i, but $\delta < 1$.

We now turn to the framework of utility maximization and relate it to the definition of QALYs. We posit von Neumann–Morgenstern utility maximization, and assume that lifetime expected utility as viewed from time 0, which we denote by E_0U, takes the form:

$$E_0U = U_0(Y_0 - w_a\boldsymbol{a} - w_b\boldsymbol{b}) + \sum_{i=1}^{N} U_i(Y_i)F_i \tag{6}$$

Period-specific utility, U_i, takes the form $U_i = v\delta^i k_i$, where $v = U_0(Y)$ and Y is a constant (in real terms) across time periods, and k_i is a period-specific multiplier interpreted as a quality adjustment above. In this form, the utility function and its argument, income, are constant over time, but period-specific utility can change by the multiplicative terms k_i and can be discounted. This assumption implies that expected utility can be rewritten as:

$$E_0U = U_0(Y - w_a\boldsymbol{a} - w_b\boldsymbol{b}) + v\sum_{i=1}^{N}\left[\delta^i k_i \prod_{j=1}^{i} P_j\right] \tag{7}$$

Thus the summation above is the number of QALYs remaining as of period 1. Define $\mathrm{d}U_0(Y - w_a\boldsymbol{a} - w_b\boldsymbol{b})/\mathrm{d}Y = U_0'$. The dependence of U_0' on $\boldsymbol{a}$ and $\boldsymbol{b}$ will be suppressed notationally from here forward, but it is important to remember this relationship.

The two available medical interventions, $\boldsymbol{a}$ and $\boldsymbol{b}$, can affect both the survival probabilities (the P_i terms) and the expected quality adjustments (the k_i terms) in future periods; we further assume that the functional relationships satisfy the usual continuity and differentiability conditions. Define $\partial P_i/\partial \boldsymbol{a} = \epsilon_i^a$, $\partial P_i/\partial \boldsymbol{b} = \epsilon_i^b$, $\partial k_i/\partial \boldsymbol{a} = \psi_i^a$, $\partial k_i/\partial \boldsymbol{b} = \psi_i^b$, and $V_i = k_i\ F_i$. Now optimize with respect to $\boldsymbol{a}$ and $\boldsymbol{b}$ in the usual fashion. Differentiation with respect to $\boldsymbol{a}$ yields the following condition:

$$\frac{\partial E_0U}{\partial \boldsymbol{a}} = -w_aU_0' + v\sum_{i=1}^{N}\delta^i\frac{\partial V_i}{\partial \boldsymbol{a}} \tag{8}$$

The change in expected utility consists of an expenditure-induced loss of period 0 utility and a gain in future expected utility, which can result from changes in the survival distribution as well as changes in the quality adjustments k_i. The derivative of each V_i term has the form

$$\frac{\partial V_i}{\partial \boldsymbol{a}} = \psi_i^a F_i + k_i \frac{\partial F_i}{\partial \boldsymbol{a}} \tag{9}$$

which decomposes the change in period i's expected utility into a change in the quality factor expected during period i, weighted by the probability of being alive then, and the change in the survival probability, weighted by the expected quality. Rewriting Eq. (9), and substituting the definition of the survival probability, we have

$$\frac{\partial E_0 U}{\partial \boldsymbol{a}} = -w_a U_0' + v\left\{\sum_{i=1}^{N} \delta^i \prod_{j=1}^{i} P_j \left(\psi_i^a + k_i \sum_{k=1}^{i} \frac{\epsilon_k^a}{P_k} \right)\right\} \tag{10}$$

which we set to zero for optimality.[6] An equivalent expression arises for intervention $\boldsymbol{b}$, replacing ϵ_i^a with ϵ_i^b and ψ_i^a with ψ_i^b.

In this form, utility is a function of (discounted) QALYs, and the term in braces represents the incremental effect of $\boldsymbol{a}$ on QALYs, which we denote as $\partial Q/\partial \boldsymbol{a}$. This term plays a central role in the analysis that follows.

If utility is the same in every period (except for discounting and the period-specific quality adjustment k_i), then (and only then) the problem in expected utility maximization becomes equivalent to a problem in discounted life expectancy, since each period's utility is assumed to be proportional to the first period's. In standard models of lifetime consumption planning, optimization implies equating the *marginal* utility of income in each period, rather than *total* (i.e., the absolute level of) utility (Hirshleifer, 1966; Ehrlich and Becker, 1972). Only if $k_i = 1 \ \forall \ i$ would optimal income transfers equalize both marginal utility and income. Allowing for quality adjustment greatly relieves this restriction, since the quality adjustment is designed to account for differences in the level of utility across states of health and across ages. Differences in quality of life could arise from shifts in health, changes in the utility function with age, or changes in the values of other arguments of utility functions, such as exogenously determined consumption of complements or substitutes for consumer goods and services.

The above equations yield the simple result that optimal investment in $\boldsymbol{a}$ is defined by:

$$w_a = \frac{v}{U_0'} \frac{\partial Q}{\partial \boldsymbol{a}} \tag{11}$$

[6] We assumed that there was no immediate effect of the intervention on quality of life (hence there is no k_0 term). If there was such an immediate effect, the marginal utility of expenditures on $\boldsymbol{a}$ in period 0 would consist of two terms, the negative one from the loss of income, and a positive term from the increase in k_0. This generalization does not affect any of the substantive conclusions that we draw from the analysis.

With this utility structure, the marginal benefit of medical care is simply the scaled utility (v/U_0') of the incremental QALYs derived from incremental $\boldsymbol{a}$, and at the optimum, incremental benefit equals incremental cost (w_a). A comparable result holds for intervention $\boldsymbol{b}$:

$$w_b = \frac{v}{U_0'} \frac{\partial Q}{\partial \boldsymbol{b}} \tag{12}$$

3.1. Consistency of cost–effectiveness ranking, ignoring future unrelated costs

With these tools, we can return to the problem we visited above, namely to consider whether one can define a utility-maximizing program using CE ratios. This time, we have two interventions, rather than one, and current medical cost is $C = w_a\boldsymbol{a} + w_b\boldsymbol{b}$. If the CE method is internally consistent, the optimal CE cutoff for interventions $\boldsymbol{a}$ and $\boldsymbol{b}$ must be the same, even if they exert their health effects at different times. If the CE method is not consistent, it cannot be used to allocate resources efficiently. We also need to allow for substitution in production of health between $\boldsymbol{a}$ and $\boldsymbol{b}$; define the marginal rate of substitution as $z = (\mathrm{d}\boldsymbol{b}/\mathrm{d}\boldsymbol{a})$. By definition, $\mathrm{d}C/\mathrm{d}\boldsymbol{a} = w_a + zw_b$. Now, define the CE ratio for intervention $\boldsymbol{a}$ as:

$$\left(\frac{\mathrm{d}C}{\mathrm{d}Q}\right)_a = \frac{\dfrac{\mathrm{d}C}{\mathrm{d}\boldsymbol{a}}}{\dfrac{\mathrm{d}Q}{\mathrm{d}\boldsymbol{a}}} = \frac{\dfrac{\partial C}{\partial \boldsymbol{a}} + zw_b}{\dfrac{\partial Q}{\partial \boldsymbol{a}} + z\dfrac{\partial Q}{\partial \boldsymbol{b}}} \tag{13}$$

Substituting the optimal values for $\partial Q/\partial \boldsymbol{a}$ and $\partial Q/\partial \boldsymbol{b}$ from Eqs. (11) and (12) leads to an extremely simple but important result. At the optimum investment in intervention $\boldsymbol{a}$,

$$\left(\frac{\mathrm{d}C}{\mathrm{d}Q}\right)_a = \frac{w_a + zw_b}{(w_a + zw_b)\left(\dfrac{U_0'}{v}\right)} = \frac{v}{U_0'} \tag{14}$$

At the optimum, the ratio of incremental costs to incremental QALYs from further investment in intervention $\boldsymbol{a}$ equals v scaled by U_0'. Thus the optimal CE cutoff is the ratio of future period-specific utility v to marginal utility in the base period.

Note that the optimal CE cutoff depends on total medical spending in the initial period. Recall that U_0' depends on income net of medical spending, that is, $Y_0 - w_a\boldsymbol{a} - w_b\boldsymbol{b}$. As current health expenditures increase, the U_0' term in the denominator rises, making the optimal CE cutoff smaller, and hence a more stringent test for a medical intervention. We explore this phenomenon in Section 4.

An exactly parallel development shows that for intervention $\boldsymbol{b}$, the same condition holds. Tracing through similar steps, we find that optimal investment implies

$$\left(\frac{\mathrm{d}C}{\mathrm{d}Q}\right)_b = \frac{\dfrac{w_a}{z} + w_b}{\left(\dfrac{w_a}{z} + w_b\right)\left(\dfrac{U_0'}{v}\right)} = \frac{v}{U_0'} \tag{15}$$

This proves the internal consistency of CE analysis, since the optimal CE cutoff, $\mathrm{d}C/\mathrm{d}Q$, is the same for both interventions. This result obviously generalizes to multiple interventions that exert their effects at different times. In the two-intervention model, intervention $\boldsymbol{a}$ might be a treatment for heart attacks, which has an immediate effect only, while $\boldsymbol{b}$ is a preventive intervention that has no immediate effects but diminishes mortality rates in the future.

3.2. Consistency when future unrelated costs are included

Do the same results hold if unrelated future costs are included in the definition of the costs for the CE ratio? To answer, we need to define the present value of expected *total* costs of care, which are

$$C^{\mathrm{tot}} = w_a\boldsymbol{a} + w_b\boldsymbol{b} + P_1\delta c_1 + P_1P_2\delta^2c_2 + \ldots \tag{16}$$

where c_i = total health expenditures in period i. The change in costs due to an intervention includes direct expenditures for the intervention, the change in expenditures for the other intervention, and the expenditures that result from living longer:

$$\begin{aligned}\frac{\mathrm{d}C^{\mathrm{tot}}}{\mathrm{d}\boldsymbol{a}} &= w_a + w_b\frac{\mathrm{d}\boldsymbol{b}}{\mathrm{d}\boldsymbol{a}} + \frac{1}{P_1}\left[\frac{\partial P_1}{\partial \boldsymbol{a}} + \frac{\partial P_1}{\partial \boldsymbol{b}}\frac{d\boldsymbol{b}}{d\boldsymbol{a}}\right]\left[\delta P_1c_1 + \delta^2P_1P_2c_2 + \ldots\right]\\ &\quad + \frac{1}{P_2}\left[\frac{\partial P_2}{\partial \boldsymbol{a}} + \frac{\partial P_2}{\partial \boldsymbol{b}}\cdot\frac{\mathrm{d}\boldsymbol{b}}{\mathrm{d}\boldsymbol{a}}\right]\left[P_1P_2\delta^2c_2 + \ldots\right] + \ldots\end{aligned} \tag{17}$$

The above expression can also be written as

$$\frac{\mathrm{d}C^{\mathrm{tot}}}{\mathrm{d}\boldsymbol{a}} = w_a + w_b\frac{\mathrm{d}\boldsymbol{b}}{\mathrm{d}\boldsymbol{a}} + \frac{\partial E}{\partial \boldsymbol{a}} + z\frac{\partial E}{\partial \boldsymbol{b}} = \frac{\mathrm{d}C}{\mathrm{d}\boldsymbol{a}} + \frac{\partial E}{\partial \boldsymbol{a}} + z\frac{\partial E}{\partial \boldsymbol{b}} \tag{18}$$

where E = the present value of expected health expenditures.

When combined with the logic we used to demonstrate consistency when future costs are excluded, these results imply that

$$\left(\frac{\mathrm{d}C^{\mathrm{tot}}}{\mathrm{d}Q}\right)_b=\left(\frac{\mathrm{d}C}{\mathrm{d}Q}\right)_b+\frac{\frac{1}{z}\left(\frac{\partial E}{\partial a}\right)+\frac{\partial E}{\partial b}}{\frac{1}{z}\left(\frac{\partial Q}{\partial a}\right)+\frac{\partial Q}{\partial b}} \tag{19}$$

and, by similar reasoning,

$$\left(\frac{\mathrm{d}C^{\mathrm{tot}}}{\mathrm{d}Q}\right)_a=\left(\frac{\mathrm{d}C}{\mathrm{d}Q}\right)_a+\frac{\frac{\partial E}{\partial a}+z\frac{\partial E}{\partial b}}{\frac{\partial Q}{\partial a}+z\frac{\partial Q}{\partial b}} \tag{20}$$

The preceding analysis showed that the first terms on the right-hand sides of these two equations are equal at the optimum. By multiplying the numerator and denominator of the second terms in Eqs. (21) and (22) by z, it is seen that the second terms are also equal. Thus, at the optimum,

$$\left(\frac{\mathrm{d}C^{\mathrm{tot}}}{\mathrm{d}Q}\right)^*_a=\left(\frac{\mathrm{d}C^{\mathrm{tot}}}{\mathrm{d}Q}\right)^*_b \tag{21}$$

implying that the CE ratio is also consistent when unrelated future costs are included.

Thus, if utility can be expressed in terms of QALYs, our model can be used to show that the optimal CE cutoff is the same for all medical interventions, so that CE methods are internally consistent. The result does not depend on our choice of including or excluding unrelated future costs. In addition, we have shown that this optimal cutoff depends strictly on the ratio of a utility level to marginal utility, v/U_0'. This result allows, at least in concept, inferences about the optimal cutoff for CE analysis that flow directly from the preference structure of consumers, rather than relying on the often distorted and confusing inferences that one can draw from calculating CE ratios for observed medical practices (see, for example, Table 1). We examine this issue in greater detail below.

This formulation is based on a straightforward mathematical representation of what we believe is usually meant by "costs of health care that result solely from living longer." Conditional on reaching a given age, a person's expenditures on health care do not change with an increase in the quantities of intervention $\boldsymbol{a}$ or $\boldsymbol{b}$ consumed. Thus the goods under study cannot be close substitutes or complements for other forms of health care (nor can there be changes in the rates of substitution between quality-enhancing and life-prolonging health care). Large shifts in future consumption of health care may well result from use of an intervention -- for example, because drugs for hypertension prevent future strokes, their consumption reduces expenditures for future stroke care – but we believe that most people would recognize that in such a situation, the future costs are truly "related."

3.3. Time costs

This framework provides a natural mechanism for examining the incorporation of time costs into a CE analysis. Assume that the treatment has non-negligible effects on time available for other activities (e.g., it requires prolonged hospitalization or convalescence from surgery, or alternatively by preventing illness in the future it produces a net time savings). There are two obvious methods for incorporating time costs into CE calculations: (1) using a suitable measure of opportunity cost, treat the time cost as a dollar cost that goes in the numerator of the CE ratio; (2) directly subtract the time the intervention takes from the QALYs attributed to it, so that the QALYs in the denominator are net of the time expenditure on the intervention. Do the two approaches give the same rankings of interventions, and if not, which one is correct? In the following discussion, we abstract from considerations of variation of quality of life under various conditions, treating time spent ill as equivalent to time dead and equivalent in disutility to time spent working. If the utility of time spent on work is not equivalent to death, time spent at work must be adjusted for its utility, or wages must be adjusted to reflect non-pecuniary compensation for work time if they are used as a measure of opportunity cost. Suitable adjustments in these numbers do not change the basic conclusions of the analysis.

3.3.1. Method 1: counting time costs as part of the numerator

For this method, follow the approach above, starting with Eq. (6), but set the costs of the intervention to $w_a + wt_a$, where w is the suitably defined opportunity cost (i.e., the shadow price of time, as would be measured by the wage rate in a perfectly competitive market, in which there is no on-the-job investment or non-pecuniary compensation). Now Eq. (11) becomes Eq. (11*):

$$w_a + t_a w = g\frac{\partial Q}{\partial a} \tag{11*}$$

where $g = v/U_0'$.

Furthermore,

$$\tilde{C} = a(w_a + t_a w) + b(w_b + t_b w)$$

.

Using the marginal rate of substitution $z = \mathrm{d}b/\mathrm{d}a$ implies that

$$\frac{\mathrm{d}\tilde{C}}{\mathrm{d}a} = w_a + t_a w + z[w_b + t_b w]$$

.

Because $\mathrm{d}Q/\mathrm{d}a = \partial Q/\partial a + z\,\partial Q/\partial b$, Eq. (13) becomes Eq. (13*):

$$\left(\frac{\mathrm{d}\tilde{C}}{\mathrm{d}Q}\right)_a = \frac{w_a + t_a w + z[w_b + t_b w]}{\dfrac{\partial Q}{\partial a} + z\dfrac{\partial Q}{\partial b}} \tag{13*}$$

and at the optimum, going through the same exercise for intervention $\boldsymbol{b}$, we find that

$$\left(\frac{\mathrm{d}\tilde{C}}{\mathrm{d}Q}\right)_a = \left(\frac{\mathrm{d}\tilde{C}}{\mathrm{d}Q}\right)_b = g$$

This proves that the same cutoff CE ratio is utility-maximizing for different interventions, because the same analysis holds for $\boldsymbol{a}$ and $\boldsymbol{b}$. Hence including the time cost as a cost in the numerator is acceptable. This imposes a requirement on w; wt_a must equal the income loss that produces the same reduction in utility as the loss of QALYs.

3.3.2. Method 2: treating time costs as a reduction in the number of QALYs gained

Now, instead of adding a dollar valuation of time costs to the numerator, every unit of $\boldsymbol{a}$ is treated as producing a loss of QALYs equal to t_a. Neither the costs nor the QALY measure that go into the CE ratio in this approach are the same as in Method 1.

Begin by subtracting t_a from the term in the curly braces in Eq. (10). Let $\partial Q/\partial a$ equal the term in curly braces, and let $\partial Q^*/\partial \boldsymbol{a} = \partial Q/\partial \boldsymbol{a} - t_a$. Q^* is the effectiveness measure for Method 2.

Now Eq. (11) becomes Eq. (11**).

$$w_a = g\left(\frac{\partial Q}{\partial a} - t_a\right) \tag{11**}$$

Carrying through the remaining equations as before, the CE ratio is defined as

$$g = \frac{\mathrm{d}C}{\mathrm{d}Q^*}$$

for each intervention. Thus each method produces the same cutoff CE ratio, or g, for each intervention. The equality holds as long as $w = g$; in other words, if at the margin the wage rate is equal to the willingness-to-pay for an additional unit of time, and as long as the appropriate definitions of both costs and QALYs are used.

Of course, if the market wage is unequal to "true" w, applying Method 1 by using the observed market wage will not result in the correct valuation. Similarly, neither method will produce consistent ranking in the presence of time costs if it fails to make appropriate quality adjustments for differences in the utility of time spent in states of illness or recuperation, time spent at work, and the utility of zero assigned to death.

4. Optimal lifetime medical spending program

This model also provides a framework for defining an optimal lifetime spending program for medical care and for exploring its relationship to the CE criterion. One health intervention can differ from another in many respects, including its marginal productivity in producing health (i.e., the size of the effects on survival probabilities and quality of life), the time course of its impact (''treatment'' expenditures are ordinarily those for existing, symptomatic illnesses, and tend to have an immediate effect, while ''preventive'' care usually has the aim of preventing future disease), and its costs. Optimal expenditures on health care can also vary because of person-specific characteristics – factors that cause them either to have different optimal CE cutoffs or to have the same cutoffs, but different utility-maximizing expenditures. In this section, we explore the causes of variation in optimal CE cutoffs and in expenditures, emphasizing the personal factors responsible for variation in optimal expenditures.

We first note the implications of Eq. (14) and Eq. (15) for variation in the optimal CE cutoff. These equations say that the CE cutoff is just the ratio of the fixed component of future period-specific utility (v) to marginal utility in the ''initial'' period. A number of factors might cause this ratio to vary among individuals, such as variation in risk aversion or other characteristics of the utility function. Even if the utility function does not differ among persons, the values of its arguments, such as income, may. The ratio of the level of utility to the marginal utility rises with income (or wealth). Furthermore, changes in health status that increase the utility of expenditures for goods and services designed to mitigate the effects of illness, such as arthritis-induced expenditures for mineral baths and pain relievers, tend to diminish the level of utility and to raise the marginal utility of income.

Even if different individuals have the same CE cutoff, there will be several reasons for their optimal expenditures to differ. For example, since advancing age is associated with a decrease in the number of potential years of life left (i.e., a decrease in annual survival probabilities), a life-saving intervention might not be capable of increasing life expectancy or QALYs by as great an amount at advanced ages as in youth. It also seems obvious that individuals with a high rate of time preference (low value of δ) would spend less on preventive care than those with a low rate of time preference.[7] Although these findings are true in general, we now explore them in detail by analyzing specific examples. We begin by specifying an intertemporal production function for health and a specific utility

[7] Insofar as we use the correct value of δ for each person, people with different values of δ can have the same CE cutoff but it will correspond to different amounts of care. If we use a single value of δ for a heterogeneous population, the CE criterion may not lead to optimal expenditures, since individual utilities will not be functions of the population-level value of δ. The same is true of other parameters of the utility function.

function. For expositional simplicity, we assume that medical spending only alters future probabilities of death, but these ideas readily generalize to the improvement of quality of life (Pliskin et al., 1980; Miyamako and Eraker, 1985).

Define $\mu_i(a) = 1 - P_i(a)$ as the age-specific probability of death in period i as a function of the level of a used in period 0. Production of health can be characterized as the relative mortality reduction that results from the expenditure on a:

$$\frac{\mu_i(a)}{\mu_i(0)} = \left[1 - \alpha\rho^i\left(1 - e^{-\phi a}\right)\right] \tag{22}$$

In this equation, $\mu_i(0)$ is the mortality rate when $a = 0$ and α reflects the largest percentage reduction in the risk of dying that an expenditure can provide. Insofar as an intervention is targeted toward a single cause of death, α will be much less than unity. The parameter ρ (ordinarily $0 \le \rho \le 1$) represents the persistence of the treatment's effect over time, and ϕ scales the impact of a on the relative mortality rate. These relations imply that the marginal productivity of a in increasing the age-specific probability of survival is

$$\frac{\mathrm{d}P_i(a)}{\mathrm{d}a} = \left(1 - P_i(0)\right)\alpha\rho^i\phi e^{-\phi a} \tag{23}$$

If we think of disease-specific expenditures for preventive care applied to the general population, then the above derivative will be very small, because $P_i \approx 1$; further, if no single cause of death predominates, then α will be small (even if the intervention completely eliminates mortality from the specific cause).[8] Thus, for example, the widely publicized effort to get Americans to reduce fat consumption will have little effect on mortality. Under fairly optimistic assumptions, which include a reduction in mortality from certain forms of cancer as well as from heart

[8] Overall mortality rates represent an unattainable upper bound, of course, on the potential change in the probability of death, while the value of α is bounded by the fraction of mortality due to the cause the intervention targets. It is difficult to have a large impact on mortality rates in young adulthood because mortality is so low at such ages, as the following 5-year mortality rates reveal:

Age interval	All-cause 5-year mortality rates
30–35	0.008
35–40	0.010
40–45	0.013
45–50	0.018
50–55	0.028
55–60	0.044
60–65	0.068

These figures are for both sexes and all races, US Life Tables for 1992 (National Center for Health Statistics, 1996, p. 7).

disease, if Americans reduced fat consumption to 30% of calories, life expectancy for a 50-year-old man would increase by about 4 days. A 50-year-old woman would only live about 2 days longer (Browner et al., 1991). Only for patients with life-threatening diseases is the potential improvement in life expectancy very large. Hence, with the exception of effective treatments applied to people at high risk of death because of illness or extraordinary predisposition to illness, we expect the product of α and $(1 - P_0)$ to be small.

The parameterization in Eq. (22) and Eq. (23) implies that the intervention's effectiveness will decay exponentially over time, at rate $1 - \rho$. Unless they are used on an ongoing basis, the effectiveness of many preventive interventions for the control of such risk factors as hypertension and hypercholesterolemia declines with time. The protective effects of vaccines also diminish with time. They prevent infectious diseases by stimulating the production of specific antibodies, whose levels gradually decline after the initial response to the vaccine. Effectiveness falls as the antibody levels drop; in this context, ρ might represent the proportion of the antibodies, or the rate of effectiveness, persisting from one year to the next.

In order to assess the implications of this model, we specify a separable utility function convex in Y_i. For the period-specific utility, a convenient and commonly used functional form specifies $U = \beta(1 - e^{-\gamma Y})$, with corresponding $U' = \gamma\beta e^{-\gamma Y}$, $U'' = -\beta\gamma^2 e^{-\gamma Y}$, absolute risk aversion $r = -U''/U' = \gamma$, and relative risk aversion $r^* = \gamma Y$ (see Pratt, 1964, and Arrow, 1974). The expressions for $\boldsymbol{a}$ and $\boldsymbol{b}$ contain the ratio U/U' as a central component. With this utility function, we can specify the ratio U/U' if we know the relative risk aversion measure $r^* = \gamma Y$. This function serves as the period-specific component expected utility, Eq. (7). This functional form allows us to assess the impact of variation in risk aversion and in other parameters of the utility function on the optimal CE ratio.

Note that this utility function with constant absolute risk aversion approximates more general functions. Consider the Taylor series expansion of an arbitrary utility function around base income Y^*. Then $U(Y) = U(Y^*) + U'(Y^*)(Y - Y^*) + U''(Y^*)(Y - Y^*)^2/2 + \dots$. If, as is common in economic analyses, we truncate the Taylor series at the second-order term, utility (scaled by $U'(Y^*)$) is completely specified by Y and the risk aversion ratio $r = -U''(Y^*)/U'(Y^*)$. Thus our simple functional form offers a second-order Taylor series approximation around Y^* to any well-behaved utility function.

To analyze the dependence of the CE ratio on the value of γ, the risk aversion parameter, we first recall that U varies with total medical spending, since we evaluate it at $Y - w_a \boldsymbol{a} - w_b \boldsymbol{b}$. Thus, the optimal CE cutoff will depend both on the utility function and the degree of medical spending. (Of course, medical spending also depends on these same preferences.) Combining the utility function, Eq. (7), and the production function, Eq. (22), and maximizing with respect to $\boldsymbol{a}$, gives the optimal spending on medical care and (from that) the optimal CE ratio.

Because optimal $\boldsymbol{a}$ cannot be readily determined analytically (Eq. (10), which

Table 2
Base case optimal spending and CE cutoff by age: women, income = $18,000 ($1989)

Age	Optimal spending ($)	Optimal CE ratio ($)
30	0	36,870
40	0	36,870
50	300	35,950
60	1010	33,890
70	1480	32,600

gives the derivative of expected utility with respect to a, does not have a simple closed-form solution (the equation contains an arbitrary number of survival probability terms that are each functions of a), we used iterative techniques to solve for the optimal values of medical spending and the CE cutoff. Solutions were computed for a wide range of parameters for income, risk aversion levels (r^*), discount rates (δ), maximal reduction in mortality rates (α), persistence of the medical intervention's effects (ρ), and gender. These simulations use actual mortality tables for US citizens, specific to gender (but not race). For the base case, we selected the median annual per capita income in the USA ($18,000 in 1989), a maximal reduction in mortality of $\alpha = 0.3$ (as appropriate for an effective treatment of a quite dangerous disease), a persistence parameter of $\rho = 0.6$, and a discount rate of 5% ($\delta = 0.95$). The resulting optimal spending rates and CE cutoffs are shown in Table 2 for females; the patterns are quite similar for males, but the optimal spending is slightly higher at each age interval because of the higher age-specific risk of death for males.

These results convey two major features of the model's behavior. First, optimal spending rises rapidly with age (as does the actual pattern of spending in the USA and elsewhere), a consequence of the increase in mortality that accompanies aging; as illness risks increase, the demand for medical intervention rises. Second, and more subtle, the optimal CE ratio falls with increasing expenditure (and hence with age), since the foregone utility from not spending income on other goods increases as medical spending increases.

The corner solution at younger ages – zero medical spending – does not result from a low CE ratio (the young have higher CE cutoffs than the old), but rather reflects the infrequency of death at younger ages. The risk of dying is so small in the youngest age groups that it cannot be reduced much more by expenditures on health care, so spending the entire budget on other goods and services provides the greatest utility. The other feature driving these results is that, as a person ages, annual mortality rates rise, so that any treatment affecting future mortality risks is "amortized" over a shorter and shorter period. These results enter our model through use of actual life tables.

A variety of sensitivity analyses (see Table 3) show that the optimal spending pattern behaves much as one might expect, and the optimal CE cutoff is remark-

Table 3
Optimal spending and sensitivity of CE ratio to production function, for 60-year-old men and women

	Women		Men	
	Spending ($)	CE ratio ($)	Spending ($)	CE ratio ($)
Maximum risk reduction α				
0.05	0	36,870	0	36,870
0.10	0	36,870	350	35,800
0.15	310	35,940	770	34,590
0.20	600	35,080	1,060	33,750
0.25	830	34,420	1,290	33,111
0.30 *	1,010	33,890	1,480	32,600
Persistence of effect (ρ)				
0.20	$310	$35,920	$790	$34,530
0.40	600	35,060	1,090	33,710
0.60 *	1,010	33,890	1,480	32,600
0.80	1,690	32,040	2,130	30,880
0.90	2,250	30,570	2,660	29,530

* = base case.

ably stable over a wide range of production function parameters (α and ρ). Increasing either α or ρ increases the marginal productivity of medical spending and optimal spending on a. This spending increase reduces the income available for other goods and services, thus making the optimal CE cutoff slightly more stringent (CE falls). This same pattern occurs at all age intervals, although optimal spending remains zero for younger persons over a broader range of the production parameters. [9] Higher values of α are possible for people who have potentially fatal illnesses (such as cancer) for which the treatment is reasonably effective. Our inclusion of an upper limit of 0.3 for α corresponds to such a case.

Varying the discount rate has effects on optimal spending and the CE cutoff as one might anticipate, although the effect interacts with age much more than in the case of the production function parameters. We vary the discount rate from 0 to 0.1 ($\delta = 1$ to 0.91), reflecting values found in the literature (see, for example, Viscusi and Moore, 1989, and Cropper et al., 1992; but see Fuchs, 1982 for a larger estimate). Table 4 shows the results for males (the pattern is very similar for females, but optimal spending is zero for a greater range of age and values of δ and lower in general, given the lower mortality risk for females at any age). While the optimal spending depends importantly on δ, the optimal CE cutoff remains stable over values we have tested.

Variability in rates of time preference may pose a special problem for most CE analyses, which are usually based on the assumption that the appropriate rate of

[9] More detailed results are available from the authors upon request.

Table 4
Optimal spending and CE ratio: men, income = $18,000

Age	$\delta = 1$		$\delta = 0.98$		$\delta = 0.95$		$\delta = 0.91$	
	Spending ($)	CE ($)	Spending ($)	CE ($)	Spending ($)	CE ($)	Spending ($)	CE ($)
30	780	34,560	320	35,890	0	36,870	0	36,870
40	1000	33,930	620	35,020	150	36,410	0	36,870
50	1140	33,520	850	34,340	480	35,430	0	36,870
60	2000	31,210	1780	31,790	1480	32,600	1060	33,760
70	2260	30,540	2110	30,930	1900	31,470	1590	32,290

time discount for the health benefits of an intervention is the same as the market rate of interest (Keeler and Cretin, 1983). If rates of time preference are the same for all people, and if capital/savings markets are perfect, the market rate of interest equals the rate of time preference (approximately 0.02 to 0.03; see Barro, 1987). But some estimates of rates of time preference are much higher than the usual values assumed for the real (or even nominal) rate of interest. Of course, there are multiple rates of interest to which one could refer, but the variability among them is evidence of capital market imperfections, perhaps explaining why market interest rates could fall short of average rates of time preference. Usual arguments about why the same discount rate should be used for health effects as for costs have less force, under the circumstances. Fortunately for social planning purposes, the optimal CE ratio does not vary importantly with the discount rate. [10]

These results confirm our earlier assertion: ordinarily a diminished planning horizon implies that preventive spending should decline with age. The same finding holds true if the effect of aging is captured instead in a greater mortality rate; for a given change in the survival probabilities, Eq. (10) implies that the marginal utility of expenditures on $\boldsymbol{a}$ is negative if the levels of the survival probabilities (F_i terms) are small enough. In other words, if a person is very unlikely to survive the current period, he or she would prefer to increase current utility by spending money on current consumption, rather than modifying a small probability of survival.

Why, then, do expenditures typically *rise* with age? Primarily because the benefit of treatment must be small when there is little disease to treat; neither survival nor quality of life can be improved significantly when there is little mortality or morbidity. Thus the CE criterion tends to promote large treatment expenditures (i.e., in which ρ may be small but the derivative of survival with

[10] Significant variation in the value of δ poses problems for CE analysis, even if it affects the cutoff CE ratio little, because the CE ratio of an intervention is a function of δ. Otherwise identical people with different rates of time preference would thus gain different levels of effectiveness from the same intervention: the cost-effectiveness of the intervention is a function of preferences as well as health status.

respect to the treatment may be large because P_i is small) at older ages and in persons who have diseases.

An important consideration is the role of the quality adjustments k_i. Thus far we have assumed that the CE analysis properly incorporates quality of life measures. Many CE analyses do not, implicitly assuming that $k_i = 1$ for all i. By omitting quality of life, they miss the effects of the intervention on future quality of life (i.e., implicitly assume $\psi_i^a = 0$), and they fail to discount properly years of life in which the expected level of utility is relatively low. Omitting the quality impact of treatment means that particular treatments will be undervalued, such as many forms of rehabilitative and long-term care that are used most commonly in old age. On the other hand, failure to recognize that years of life extended at older ages are often characterized by worsened health status tends to bias expenditures in favor of the elderly. If one accepts the notion that quality of life falls as physical and mental disability increase (see, for example, Torrance, 1987), then the usual pattern of declining physical function that accompanies aging implies that the pattern of multipliers k_i becomes smaller as a person grows older. If so, simplifying CE analyses by assuming $k_i = 1$ for all i, that is, assuming that utility is a function of discounted life expectancy alone, will result in *overestimating* the optimal spending for persons in their later years of life.

In all of the sensitivity analyses discussed above, the optimal CE ratio remains fairly constant over a wide range of values of the parameters of the utility function and over a range of personal characteristics, although optimal spending varies considerably with age, the marginal productivity of medical care, and the discount rate. However, the optimal CE ratio is sensitive to two characteristics that vary among individuals – income and risk aversion. These findings have important consequences for private and public allocation of medical care resources and for social planning of medical investments.

Fig. 1 shows how the optimal CE cutoff varies by income and the degree of risk aversion. The two income levels – \$18,000 and \$29,000 – correspond to median per capita and per family income in 1989. The utility function captures risk aversion in terms of $r^* = \gamma Y$. When researchers have estimated the degree of risk aversion using various methods, the estimates center on a relative risk aversion of about 2.0 (see Weber, 1970, 1975; Friend and Blume, 1975; Blume and Friend, 1975; Farber, 1978; Siegel and Hoban, 1982; Hansen and Singleton, 1983; Litzenberger and Ronn, 1986; Szpiro, 1986; Hall, 1988; Caballero, 1991; Gruber and Madrian, 1995), with a range of about 1 to 4 (hence our choice of these parameters). [11] These figures also show the interaction of the effects of age and risk aversion: at higher degrees of risk aversion, the optimal CE ratio shifts more with age, and as people become less risk averse, age has a diminishing (and finally nearly zero) effect on the optimal CE ratio. Fig. 1 also shows the results

[11] Our thanks to Mark Machina for guiding us to many of these references.

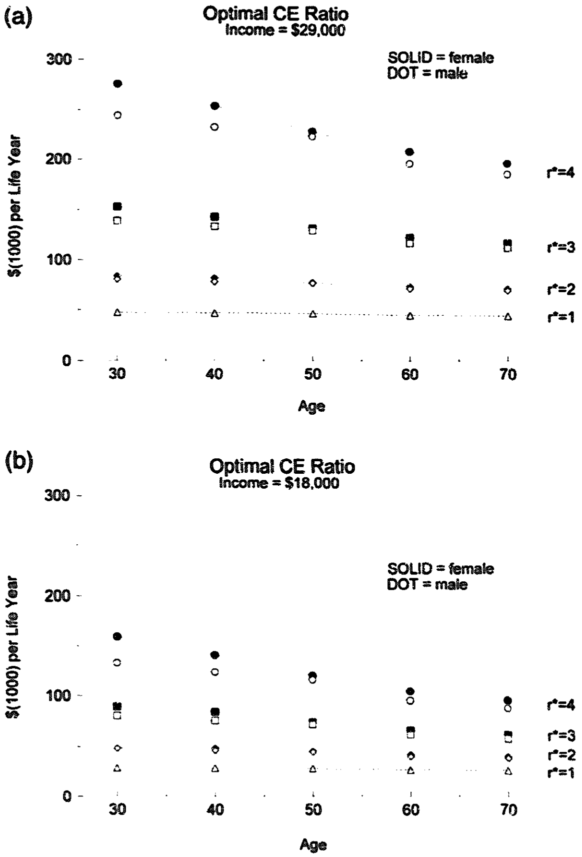

Fig. 1. Optimal cost–effectiveness ratio for women and men, by age and relative risk aversion (R^*).

previously mentioned, namely that the effects of gender matter only a very little in determining the optimal CE cutoff, entering this model solely through the effects of differential risks of mortality on the optimal spending program, and hence on the optimal CE cutoff.

5. Discussion

CE analysis has long been recognized as a convenient approach to guiding health care decisions. Its validity, however, has not been rigorously established. We have shown that, within the framework of standard von Neumann–Morgenstern utility maximization, CE analysis can offer a valid criterion for choosing among health interventions. Surprisingly, the inclusion of unrelated future costs is without consequence so long as the practice is consistent. Although our analysis is based on a specific family of utility functions and a strict definition of unrelated future health expenditures (i.e., they are conditionally independent of prior expenditures) the use of quality adjustments allows it to approximate a wide range of functional forms. The frequent use of life expectancy as the chief outcome variable in CE analysis is considerably more restrictive. With the quality adjustments, CE analysis can be a powerful guide for decision making.

Insofar as the observed CE ratio of medical interventions in common use varies by at least an order of magnitude, the usual practice of comparing the CE of a particular intervention with that for others offers little guidance for planning or resource allocation. Although it is clear that CE ratios should be equalized across interventions at the margin, seemingly arbitrary criteria are often used to select a specific cutoff CE ratio (one rule of thumb widely applied in the USA, for example, is to deem as acceptable any CE ratio less than the annual cost of hemodialysis for a patient with end-stage renal disease). We have presented an alternative method for picking the optimal cutoff, showing how it can be derived from the parameters of a flexible utility function. Our estimates imply that, over the range we estimated, CE cutoffs should be about double the annual income.

When effectiveness is measured in terms of life expectancy, the optimal CE ratio represents the same concept as the (marginal) ''willingness to pay'' – the amount an individual would pay to reduce a risk of death. Empirical work has shown that the ''willingness to accept'' – the amount of money that individuals would require to voluntarily accept a risk of death from job causes – is on the order of $300,000 per year of life expectancy in jeopardy (Viscusi and Moore, 1989). The willingness to pay for a reduction in the risk of death may be quite different, and would ordinarily be substantially lower (Hanemann, 1991). [12] Labor market data that are used to infer willingness-to-accept, therefore, provide an

[12] Hanemann explored public goods, of which environmental risk is an example. The obvious reason willingness to pay and willingness to accept can differ is the income effect, but it is usually negligible for environmental risks. The income effect is also likely to be small for preventive health care and other care delivered to low-risk individuals, but it can be much larger for some health conditions (falling ill with a serious disease is equivalent to a large loss in endowment). Furthermore, Hanemann showed that even when the income effect is small, willingness to pay and willingness to accept can differ greatly, as long as private goods are poor substitutes for the public good.

upper bound on the optimal CE ratio, although the appropriate cutoff for a given individual's utility function may be substantially lower.

Under the assumptions employed in our models, CE analysis leads to the same decisions as CB analysis. When these assumptions are violated, however, CE analysis will not necessarily lead to potential Pareto improvement. The grounding of CE analysis in von Neumann–Morgenstern utility theory and welfare economics more generally depends heavily on whether QALYs adequately represent preferences. Several authors have commented on the restrictive assumptions required to represent utilities in this form, chief among them additive separability, risk neutrality over the length of life, and the constant rate of time preference (see, for example, Pliskin et al., 1980, and Kamlet, 1992). Furthermore, in much of the literature on quality of life effects in CE analysis, the definition of the quality adjustments (k_i terms) is somewhat vague – usually (but not always) they are designed to measure only "health-related" quality of life. It is not certain, however, that survey respondents hold everything else (including income and other non-health arguments of utility) constant when they rate health states. Insofar as the k_i terms are defined narrowly, QALYs are unlikely to serve as comprehensive measures of utility.

When QALYs do not represent utility adequately, the usual CE approaches cannot offer reliable guidance to welfare improvement. It is not hard to think of such circumstances – for example, individuals may have a non-constant rate of time preference (placing great weight on surviving to a particular event), or significant intertemporal dependencies in utility may violate separability (experience in the consumption of some goods and services influences future utilities from their consumption; see Becker and Murphy, 1988). Under these circumstances, CB calculations, and economic evaluation generally, will be particularly difficult, and a far more complex calculation may need to be performed for each individual. However, QALYs are likely to offer a reasonable approximation to utilities in many other circumstances, particularly for local changes in outcomes.

Our analysis relies on the assumption that the marginal cost effectiveness of alternative health interventions can be equated by varying their quantity continuously. Thus, each intervention that is used at all can be used until its CE ratio just equals the cutoff or threshold CE ratio (equivalent to equating marginal benefit with marginal cost). Many interventions, however, appear to be discrete, hence not subject to the divisibility necessary for marginal conditions to hold. For some of them, despite the appearance of "lumpiness," the quantity can be varied continuously, or nearly so. For example, a decision to undergo mammographic screening is, at first glance, an either/or decision for a woman. But the frequency of mammography can be varied continuously. Similarly, within a population, the margin at which the quantity of the intervention is varied might be based upon a continuously distributed underlying ability to benefit from the procedure. Such variation in benefit might derive from variation in demographic and physiological characteristics. The ability to benefit from cholesterol-lowering drugs (hence the cost-effectiveness ratio of expenditures on the drug), for example, varies with an individual's underlying risk of developing coronary heart disease, which in turn is a function of cholesterol level, age, gender, blood pressure, and other characteristics (Goldman et al., 1991; Garber et al., 1996a). We suspect that the quantity or intensity of many health interventions, particularly preventive interventions, can be continuously varied in this manner.

What if, however, an intervention and its required expenditures come in truly discrete, indivisible quantities? Then the validity of the main conclusions of our analysis depends upon the budget constraint. Suppose first that expenditures on such "lumpy" interventions are small relative to the budget. Use of the intervention will be welfare-enhancing as long as the value of the increase in QALYs it produces is at least as great as its cost. If the value of a QALY varies little with the quantity of the intervention, at least in the relevant range, the CE criterion will remain valid: any intervention whose CE ratio is less than (greater than) the CE threshold is (not) welfare-enhancing. Under these circumstances, for any set of interventions whose CE ratios are on the same side of the CE threshold, their ranks (by individual CE ratios) are irrelevant, since all will be chosen (rejected) if their CE ratio is less than (more than) the CE threshold. Thus in this case, the lumpiness of the intervention has no real consequence for the analysis.

Lumpiness matters if there is an explicit, potentially binding budget constraint, or if the threshold CE ratio varies across the relevant range of QALYs. Under a binding budget constraint, some interventions or combinations of interventions whose CE ratio is less than the threshold CE ratio may no longer be feasible. Then projects can no longer be ranked solely by their CE ratios. Thus, just as the CB ratio cannot be the sole criterion to select projects that are indivisible, CE ratios cannot by themselves guide resource allocation under these circumstances. Rather, each project's benefits and costs must be tested against the budget constraint, and the combination of projects that meet the budget constraint and provide the greatest increase in QALYs will be the one selected. Hence, we speculate that under the circumstances in which CE ratios offer sufficient information to rank alternative health interventions, the results of our analysis – such as the invariance of time costs and of unrelated future costs of care – remain valid.

Implicit in our discussion is the assumption that CE analysis is used to improve decision making at an individual level. Ordinarily an apparatus like CE analysis is unnecessary for individual consumption decisions, in the absence of externalities or public good considerations. In health care, however, the familiar informational failures are sufficient reason for CE analysis to be performed as an aid to individual decisions. A more common application, however, is to decisions about the scope of health insurance: the technique can be used to help determine which forms of health care should be reimbursed by a private or governmental insurer, or provided by a health-maintenance organization. The optimal CE criterion is equivalent to determining optimal coverage for an actuarially fair insurance policy, under perfect information.

Often, however, the assistance of CE analysis is sought for making broad public policy decisions (see Kamlet, 1992; Tolley et al., 1994; Pauly, 1995; Russell et al., 1996; Garber et al., 1996b, for discussions of the ways in which CE analysis and other health valuation techniques are or could be used for such purposes). In such settings, the purpose of the analysis is to improve social welfare. Then the principal issues are whether application of the technique generates a potential Pareto improvement (Kaldor–Hicks criterion), and whether it can have favorable distributional properties. Do social decisions based on CE (applying a fixed CE cutoff to all forms of health expenditure) have the desired properties?

Our analysis showed that a CE criterion applied at the individual level, like a CB criterion, can lead to optimal consumption. But CB analysis is usually applied in a very different set of circumstances. Most CB analysis is designed to assist in the evaluation of public goods, when the chief task is to measure the total benefit by *summing* individual surpluses. CE analysis in health care, by contrast, addresses the consumption of goods and services that are mostly private (i.e., both excludable and rival) and, when applied to a population, estimates an *average* measure of valuation rather than a sum. In a population in which demand for QALYs (the optimal CE ratio) varies, application of a uniform ratio means that some individuals will receive health care whose marginal benefit exceeds marginal cost, while for others the opposite will hold true. Thus the distribution of care is no longer likely to satisfy the Kaldor–Hicks criterion. Preference variability poses measurement challenges for the evaluation of public goods but is unlikely to generate the inefficiencies that result from the uniform consumption of private goods.

Inter-individual variability in the optimal CE ratio leads to a fundamental tension in using CE analysis to guide the allocation of health care resources: insurers and policy makers may wish to equate CE ratios across interventions *and* across individuals, yet the CE of an intervention varies within heterogeneous populations and members of the population can have very different optimal CE ratios. Individual variability in demand may be an important reason for the persistence of a pluralistic health care system in the USA, despite its perceived inefficiencies.

The use of CE analysis to improve both equity and efficiency is particularly congenial to social insurance approaches to health care, and can be justified either by appeal to welfare economic principals or to the claim that the maximization of health (as measured by QALYs) itself is the goal or a goal for social policy (see Williams, 1993, and Culyer, 1991). Ordinarily health care provision with a uniform CE cutoff will be more "equal" than the market-based distribution of health care, since the wealthy possess relatively high cutoff CE ratios and will purchase more care than others. From a social welfare perspective, the improvements in equity may offset the loss of (Pareto) efficiency. Nevertheless, the uniform CE cutoff does not imply that QALYs will themselves be equalized; a more egalitarian distribution of well-being would require using a higher CE cutoff for those whose endowments of utility (QALYs) are lower. The framework of this paper is not designed to address distributional iinplications of the application of CE analysis, but we do not believe that global evaluations of social welfare can be made on the basis of health alone. The principal goal of CE analysis, we believe, is to promote economic efficiency in the allocation of health services. If applied with appropriate recognition of its limitations, it can succeed.

Acknowledgements

This research was supported in part by grants AG07651 from the National Institute on Aging, HL4629 from the National Heart, Lung, and Blood Institute, HS5477 from the Agency for Health Care Policy and Research, and 84R-2459-HPE from the Henry J. Kaiser Family Foundation. Some of the work was performed while Garber was a Fellow at the Center for Advanced Study in the Behavioral Sciences. Garber is an HSR&D Senior Research Associate of the Department of Veterans Affairs. We are grateful to Elisabeth Coutts for research assistance, and to Andrew Dick, Dana Goldman, David Meltzer, Alvin Mushlin, Milton Weinstein, Joe Lipscomb, and Will Manning for helpful comments.

References

Arrow, K.J., 1974, Essays in the theory of risk bearing (North-Holland, Amsterdam).

Avorn, J., 1984, Benefit and cost analysis in geriatric care: Turning age discrimination into health policy, New England Journal of Medicine 310, 1294–1301.

Barro, R.J., 1987, Macroeconomics, 2nd edn. (Wiley, New York).

Becker, G.S. and K.M. Murphy, 1988, A theory of rational addiction, Journal of Political Economy 96, 675–700.

Blume, M.E. and I.E. Friend, 1975, The asset struuture of individual portfolios and some implications for utility functions, Journal of Finance 30, 585–603.

Boyle, M.H., G.W. Torrance, J.C. Sinclair and S.P. Horwood, 1983, Economic evaluation of neonatal intensive care of very-low-birth-weight infants, New England Journal of Medicine 308, 1330–1337.

Browner, W.S., J. Westenhouse and J.A. Tice, 1991, What if Americans ate less fat? A quantitative estimate of the effect on mortality, Journal of the American Medical Association 265, 3285–3291.

Caballero, R.J., 1991, Earnings uncertainty and aggregate wealth accumulation, American Economic Review 81, 859–871.

Cropper, M.L., S.K. Aydede and P.R. Portney, 1992, Rates of time preference for saving lives, American Economic Review 82, 469–472.

Culyer, A.J., 1991, The normative economics of health care finance and provision, in: A. McGuire, P. Fenn and K. Mayhew, eds., Providing health care (Oxford University Press, Oxford).

Cummings, S.R., S.M. Rubin and G. Oster, 1989, The cost-effectiveness of counseling smokers to quit, Journal of the American Medical Association 261, 75–79.

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

Eddy, D.M., 1989, Screening for breast cancer, Annals of Internal Medicine 111, 389–399.

Eddy, D.M., 1990, Screening for cervical cancer, Annals of Internal Medicine 113, 214–226.

Ehrlich, I. and G.S. Becker, 1972, Market insurance, self-insurance, and self-protection, Journal of Political Economy 80, 623–648.

Farber, H.S., 1978, Individual preferences and union wage determination, Journal of Political Economy 86, 923–942.

Friend, I.E. and M.E. Blume, 1975, The demand for risky assets, American Economic Review 65, 900–923.

Fuchs, V.R., 1982, Time preference and health: An exploratory study, in: V.R. Fuchs, ed., Economic aspects of health (University of Chicago Press, Chicago), 93–120.

Garber, A.M., W.S. Browner and S.B. Hulley, 1996a, Cholesterol screening in asymptomatic adults, revisited, Annals of Internal Medicine 124, 518–531.

Garber A.M., M.C. Weinstein, G.W. Torrance and M.S. Kamlet, 1996, Theoretical foundations of cost–effectiveness analysis, in: M.R. Gold, J.E. Siegel, L.B. Russell and M.E. Weinstein, eds., Cost–effectiveness in health and medicine (Oxford University Press, New York).

Goldman, L., M.C. Weinstein, P.A. Goldman and L.W. Williams, 1991, Cost–effectiveness of HMG-CoA reductase inhibition for primary and secondary prevention of coronary heart disease, Journal of the American Medical Association 265, 1145–1151.

Gruber, J. and B.C. Madrian, 1995, Health insurance availability and the retirement decision, American Economic Review 85, 938–948.

Hall, R.E., 1988, Intertemporal substitution, Journal of Political Economy 96, 339–357.

Hanemann, W.M., 1991, Willingness to pay and willingness to accept: How much can they differ? American Economic Review 81, 635–647.

Hansen, L.P. and K.J. Singleton, 1983, Stochastic consumption, risk aversion, and the temporal behavior of asset returns, Journal of Political Economy 91, 249–265.

Hirshleifer, J., 1966, Investment under uncertainty: Applications of the state-preference approach, Quarterly Journal of Economics 80, 252–277.

Kamlet, M.S., 1992, The comparative benefits modeling project: A framework for cost–utility analysis of government health care programs (Office of Disease Prevention and Health Promotion, Public Health Service, US Department of Health and Human Services, Washington, DC).

Keeler, E.B. and S. Cretin, 1983, Discounting of life-saving and other nonmonetary effects, Management Science 29, 300–306.

Littenberg, B., A.M. Garber and H.C. Sox, Jr., 1990, Screening for hypertension, Annals of Internal Medicine 112, 192–202.

Litzenberger, R.H. and E.I. Ronn, 1986, A utility-based model of common stock price movements, Journal of Finance 15, 67–92.

Miyamako, J.M. and S.A. Eraker, 1985, Parameter estimates for a QALY utility model, Medical Decision Making 5, 191–213.

National Center for Health Statistics, 1996, Vital statistics of the United States, 1992; vol. II, sec. 6 Life tables (US Public Health Service, Washington, DC).

Office of Technology Assessment, US Congress, 1981, Cost effectiveness of influenza vaccination (US Government Printing Office, Washington, DC).

Pauly, M.V., 1986, Taxation, health insurance, and market failure in the medical economy, Journal of Economic Literature 24, 629–675.

Pauly, M.V., 1995, Valuing health benefits in money terms, in: F.A. Sloan, ed., Valuing health care: Costs, benefits, and effectiveness of pharmaceuticals and other medical technologies (Cambridge University Press, Cambridge, UK) 99–124.

Phelps, C.E. and A.I. Mushlin, 1991, On the (near) equivalence of cost effectiveness and cost benefit analysis, International Journal of Technology Assessment in Health Care 7, 12–21.

Pliskin, J.S., D.S. Shepherd and M.C. Weinstein, 1980, Utility functions for life years and health status, Operations Research 28, 206–224.

Pratt, J.W., 1964, Risk aversion in the small and in the large, Econometrica 32, 122–136.

Riddiough, M.A., J.E. Sisk and J.C. Bell, 1983, Influenza vaccination: Cost-effectiveness and public policy, Journal of the American Medical Association 249, 189–195.

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

Russell, L.B., J.E. Siegel, N. Daniels, M.R. Gold, B.R. Luce and J.S. Mandelblatt, 1996, Cost-effectiveness analysis as a guide to resource allocation in health: Roles and limitations, in M.R. Gold, J.E. Siegel, L.B. Russell and M.E. Weinstein, eds., Cost-effectiveness in health and medicine (Oxford University Press, New York).

Siegel, F.W. and J.P. Hoban, 1982, Relative risk aversion revisited, Review of Economics and Statistics 64, 481–487.

Sox, H.C. Jr., B. Littenberg and A.M. Garber, 1989, The role of exercise testing in screening for coronary artery disease, Annals of Internal Medicine 110, 456–469.

Szpiro, G.A., 1986, Measuring risk aversion: An alternative approach, Review of Economics and Statistics 68, 156–159.

Tolley, G., D. Kenkel, R. Fabian and D. Webster, 1994, The use of health values in policy, in: G. Tolley, D. Kenkel and R. Fabian, eds., Valuing health for policy (University of Chicago Press, Chicago), 345–391.

Torrance, G.W., 1986, Measurement of health state utilities for economic appraisal, Journal of Health Economics 5, 1–30.

Torrance, G.W., 1987, Utility approach to measuring health-related quality of life, Journal of Chronic Diseases 40, 593–600.

Viscusi, W.K. and M.J. Moore, 1989, Rates of time preference and valuations of the duration of life, Journal of Political Economy 38, 297–317.

Weber, W.E., 1970, The effects of interest rates on aggregate consumption, American Economic Review 60, 591–600.

Weber, W.E., 1975, Interest rates, inflation and consumer expenditures, American Economic Review 65, 843–858.

Weinstein, M.C., 1981, Economic assessment of medical practices and technologies, Medical Decision Making 1, 309–330.

Weinstein, M.C. and H.V. Fineberg, 1980, Clinical decision analysis (W.B. Saunders, Philadelphia).

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