Blomqvist, Åke

Working Paper

Economic efficiency and QALY-based cost-utility analysis in health care

Research Report, No. 2000-7

Provided in Cooperation with:
Department of Economics, University of Western Ontario

Suggested Citation: Blomqvist, Åke (2000) : Economic efficiency and QALY-based cost-utility analysis in health care, Research Report, No. 2000-7, The University of Western Ontario, Department of Economics, London (Ontario)

This Version is available at:
https://hdl.handle.net/10419/70425

Economic efficiency and QALY-based Cost-Utility analysis in health care

Åke Blomqvist*

Abstract.

Economic evaluation in health care often involves cost-utility analysis (CUA), a method based on the cost-effectiveness criterion of dollars per Quality-Adjusted Life Year (QALY), where the quality adjustement factors for years lived in ill health are established through answers to standard-gamble questions. This paper shows that, contrary to a widely held notion, allocation of a fixed health care budget through CUA does not generally result in a second-best efficient allocation. However, CUA can be used to attain second-best efficiency in the sense of meeting a given QALY target at minimum social cost. It also qualifies Meltzer's (1997) result that the cost per QALY for life-saving medical interventions should include the future consumption of those who would otherwise not have survived, by showing that its validity depends on how the QALY index has been defined. Another finding is that failure to specify carefully what respondents to standard gamble question are supposed to assume about the financial consequences of ill health may result in a bias against providing care to older individuals.

1. Introduction

In applied economics, cost-benefit analysis (CBA) is generally the preferred technique for valuing projects or policies, principally because it is based on an intuitively appealing comparison of the money value of both costs and benefits. However, in some areas, such as health care, reliable information on the value of the benefits of projects or policies is not readily available. As a result, evaluation in such areas is often based on cost-effectiveness analysis (CEA) in which benefits are measured in some kind of natural unit (rather than in money terms), and the standard according to which projects are compared is the cost per natural unit.

CEA is most useful when one is interested in comparing projects that yield identical (or at

* Department of Economics, University of Western Ontario, London, Canada, N6A 5C2. Tel: 519-661-2111, Xt. 85207; fax 519-661-3666; e-mail: akeb@julian.uwo.ca

least very similar) kinds of benefits. Resources spent on health services, however, generate many different kinds of benefits, such as reduced mortality rates among individuals who may be young or old, reductions in the productivity losses that occur when individuals are in ill health, and improved quality of life through reduced pain and suffering. If one wants to use CEA in order to compare programs that yield different kinds of benefits, the comparison must be based on some kind of natural unit that aggregates them. Cost-utility analysis (CUA), in which the natural unit is quality-adjusted life years (QALYs) has been developed in response to this objective, and in recent years has become the most commonly used method in economic evaluation in the health care field.

A question of obvious interest to economists concerns the relationship between CUA and the conventional microeconomic concept of efficiency. Although CUA has been in practical use for some time, most of the literature exploring this relationship is of relatively recent origin. One branch of this literature has addressed the question: Under what conditions is maximization of a QALY index equivalent to the maximization of expected utility?[1] Another, more limited, question is: Will allocation of a fixed health care budget in accordance with a QALY criterion lead to a second-best optimum in the sense that resources are allocated efficiently subject to this fixed budget?[2] Explicitly or implicitly, many practitioners believe that, if the QALY criterion is properly used, the answer to this question is yes, and much of the discussion in the literature has concerned the issue how the QALY criterion should be defined in order to accomplish this. The analysis in this paper focusses on the question under what conditions this belief is warranted.

Evidently, the answer depends both on how the QALY measure is operationalized, and on the definition of cost that is used in comparing the cost per QALY of different interventions.

With respect to the cost definition, an issue that has been hotly debated concerns the case of life-saving interventions: Should the cost concept for such interventions include the future resource use of individuals who, without them, would not have survived? Conversely, should the value of the production attributable to the labour of survivors be treated as a benefit (a negative cost)? In practice, these issues are important not only from the viewpoint of assessing the relative cost-effectiveness of life-saving interventions and those that only improve individuals' quality of life, but also for the allocation of resources among the young and the old.[3]

With respect to the definition of QALYs, I consider the case of a QALY index based on the convention of assigning the value 1 to a year lived in good health, while the quality adjustment factors used to assign a value to a year lived in ill health are constructed on the basis of truthful answers to standard-gamble questions in which individuals are asked what decreases in survival probability they would be prepared to accept in return for a specified increase in their expected quality of life. Although this is not the only way of constructing a QALY index, the standard gamble approach is the one most closely related to conventional economic theory since it is supposed to reflect the same individual preferences that are used in the normal definition of economic efficiency.

Briefly, the conclusions of the paper are as follows. First, contrary to what is often believed, cost-utility analysis on the basis of a QALY index as conventionally defined is generally *not* an appropriate tool for solving the problem of allocating a fixed health care budget in a second-best optimal fashion. That is, it will not in general yield the same allocation of a fixed health care budget as CBA would, if CBA were possible. However, CUA based on such a QALY index can be used to attain another kind of second-best optimum: The cost-minimizing

pattern of spending to meet a given population health target (defined as a fixed number of QALYs for a representative individual). Second, when CUA is used in conjunction with a conventional QALY index, I confirm the conclusion in Meltzer (1997) that the incremental consumption of survivors of mortality-reducing projects should be counted as part of the projects' costs. However, I also show that under certain assumptions, answers to standard gamble questions can be used to construct an alternative index of consumers' relative valuation of longer life expectancy and better quality of life; when this index is used, the incremental consumption costs should not be included in the CUA, because they have already been taken into account in the calculation of the valuation index. Third, with respect to the allocation of health resources between the young and the old, CUA does not, in principle, imply an "age bias", provided it is based on an index of discounted QALYs where the discount rate also has been established through answers to a type of standard gamble question. In practice, however, unless care is taken to specify carefully what individuals are supposed to assume about their hypothetical economic circumstances when answering standard gamble questions, I argue that there is a possibility that use of CUA may result in fewer resources than would be socially efficient being allocated to the care of old individuals.

The rest of the paper is organized as follows. In section 2, I sketch a formal model which depicts a life-cycle maximization problem in a situation where health services may be used both to increase survival probabilities among the young and the old, and to increase the expected quality of life for individuals in either group, and show how quality adjustment factors can be constructed on the basis of the standard gamble technique. In section 3, I use the model to consider the answers to the questions referred to above. Section 4 concludes.

2. The model

Consider a model in which *ex ante* identical individuals live for two periods indexed by superscripts *y* (for young) and *o* (for old). At the beginning of the young and old period, respectively, the individual may die, but survives with probability s^y and s^o respectively. (At the end of the old period, everyone dies.) I assume that each individual's utility is an additively separable function of the utility he or she enjoys at different times in their life. Specifically, when individuals are young, their utility depends on the amount *c* of goods they consume, and the leisure *l* they enjoy, as well as on their state of health which is denoted by *h*. For simplicity, the health variable can take on only two values, 1 or 0, with 1 denoting good health and 0 denoting ill health. The state-specific utility of a young individual is written as $U^y(c_h^y, l_h, h)$, where h=0,1. The probability that a young individual will be healthy is denoted by q^y. As old individuals do not work, their utility depends only on their state-specific consumption and their state of health: $U^o = U^o(c_h^o, h)$. The probability that an old individual will enjoy good health (h=1) is denoted by q^o. Without loss of generality, the scale used to measure utility may be defined such that the utility of being dead is zero. A newly born individuals's life-time expected utility is then written

$$E = s^y[q^y U^y(c_1^y, l_1, 1) + (1-q^y)U^y(c_0^y, l_0, 0)] + \beta \cdot s^y s^o[q^o U^o(c_1^o, 1) + (1-q^o)U^o(c_0^o, 0)] \quad (1)$$

where β is a subjective discount factor.

The survival probabilities s^j and the probabilities of good health q^j are assumed to depend

on the amount of medical resources spent on enhancing them, and I denote by m_{sj} and m_{qj} the amount spent on increasing survival probability and the probability of good health, respectively, in the *j*th age group, $j=y,o$. While individuals themselves choose consumption and leisure (when young), I assume that the choices of how much to spend on the different kinds of health services is made by a government decision-maker.

Let N^y and N^o denote the expected consumption net of the individual's own contribution to output for young and old individuals, respectively. Taking into account that, by assumption, old people don't work, one has

$$\begin{aligned} N^y &= q^y(c_1^y - w_1(1-l_1)) + (1-q^y)(c_0^y - w_0(1-l_0)) \\ N^o &= q^o c_1^o + (1-q^o)c_0^o \end{aligned} \qquad (2)$$

where w_i is a labour productivity parameter (wage rate) which depends on the person's health state, and where I have taken into account the possibility that the optimal state-contingent consumption of old individuals may differ depending on their ex post health state. I assume that there is no utility interdependence among young and old individuals, so that no utility is derived from leaving assets behind when you die. Moreover, I assume that there exists a complete set of contingent markets so that individuals' consumption when young and old does not depend on the income they *actually* earn while working (which depends on their state of health when young), but instead on their *expected* earnings. With these assumptions, the expected life-time resource constraint of a representative individual can be written:

$$\Omega \equiv A - s^y(N^y + r \cdot s^o N^o) - M = 0 \qquad (3)$$

where A is non-labour income, r is the market discount rate, and M is the discounted average life-

time cost of health care for a representative person, given by

$M = m_{sy} + s^y(m_{qy} + r(m_{so} + s^o m_{qo}))$. Individuals are assumed to choose age and state specific consumption and leisure so as to maximize expected life-time utility subject to (3), while the values of the m_{ij} are chosen by a government decision-maker.[4]

As usual, the solution to the problem of maximizing (1) subject to (3) can be found by forming the Lagrangean expression Ψ given by

$$\Psi = E + \lambda\Omega \tag{4}$$

where $\lambda>0$, and setting the partial derivatives with respect to the choice variables, equal to zero.

For given values of the survival probabilities and the probabilities of good health, a "project" will simply consist of an increment $d(m_{ij})$ in one of the four kinds of health services, leading to an increase in one of the probabilities s^y, q^y, s^o, and q^o.

In conventional cost-benefit analysis, a project is considered profitable if it leads to a potential Pareto improvement (PPI). In the present case, a project $d(m_{ij})$ passes the PPI test if it increases the representative individual's expected utility. Assuming that individuals choose consumption and leisure optimally, the impact on a representative individual's expected utility of a specified increment $d(m_{ij})$ is simply given by the partial derivative of of (4) with respect to the given kind of expenditure.[5] Performing the differentiations yields the following expressions:

$$\frac{\partial\Psi}{\partial m_{sy}} = \frac{ds^y}{dm_{sy}}\left(\overline{U}^y + \beta s^o \overline{U}^o - \lambda(N^y + r s^o N^o + R_{sy})\right) - \lambda \tag{5}$$

where $R_{sy} = m_{qy} + r(m_{so} + s^o m_{qo})$ is the expected (discounted) spending on medical care for

"young survivors", and $\overline{U}^{j}$, $j = y, o$ are expected utility of the young and old, respectively;

$$\frac{\partial \Psi}{\partial m_{qy}} = s^{y} \frac{dq^{y}}{dm_{qy}} \left(U^{y}(1) - U^{y}(0) - \lambda \Delta N^{y} \right) - s^{y} \lambda \qquad (6)$$

where $\Delta N^{y} = (c_1^{y} - w_1(1 - l_1) - (c_0^{y} - w_0(1 - l_0))$; that is, the difference between the net consumption (consumption less contribution to output) of a person in good and ill health, respectively;

$$\frac{\partial \Psi}{\partial m_{so}} = s^{y} r \frac{ds^{o}}{dm_{so}} \left((\beta / r) \overline{U}^{o} - \lambda (N^{o} + m_{qo}) \right) - s^{y} r \lambda \qquad (7)$$

and finally,

$$\frac{\partial \Psi}{\partial m_{qo}} = s^{y} s^{o} r \frac{dq^{o}}{dm_{qo}} \left((\beta / r)(U^{o}(1) - U^{o}(0)) - \lambda (c_1^{o} - c_0^{o}) \right) - s^{y} s^{o} r \lambda \qquad (8)$$

Denoting by π_{ij} the net profitability of spending an additional unit of resources in any of the four different ways, (5) to (8) can also be rewritten in the following intuitively appealing form:

$$\pi_{sy} = (1 / \lambda)[\overline{U}^{y} + \beta s^{o} \overline{U}^{o}] - [N^{y} + r s^{o} N^{o} + R_{sy}] - [ds^{y} / dm_{sy}]^{-1} \qquad (9)$$

$$\pi_{qy} = (1 / \lambda)[U^{y}(1) - U^{y}(0)] - \Delta N^{y} - [dq^{y} / dm_{qy}]^{-1} \qquad (10)$$

$$\pi_{so} = (\beta / \lambda)\overline{U}^{o} - r(N^{o} + m_{qo}) - r[ds^{o} / dm_{so}]^{-1} \qquad (11)$$

$$\pi_{qo} = (\beta / \lambda)[U^{o}(1) - U^{o}(0)] - r[c_1^{o} - c_0^{o}] - r[dq^{o} / dm_{qo}]^{-1} \qquad (12)$$

In each of these expressions, the last (bracketed) term gives the marginal cost to the health care budget of increasing either survival probability or the probability of good health, among the young and the old respectively. In the discussion below, I will refer to this as the "public cost" (of generating the incremental utility flow associated with the relevant health improvement).The second (middle) term is the value of the net extra resource use (in the form of consumption net of any productivity change or of spending on health care resources) caused by these changes. I will refer to this component as the "consumption cost". The first term, finally, is the direct change in utility associated with the change in life expectancy or expected quality of life, when consumption and labour supply are at their optimal levels.

In expressions (9) to (12), the analyst can be assumed to have estimates of the marginal costs (that is, the third term in each expression). Moreover, the second term in each (the consumption cost) can, in principle, be estimated based on observable behaviour.[6] The problem, of course, is that the terms representing the direct utility gains from survival and better health cannot be directly observed. In conventional Cost-Utility Analysis (CUA), the approach is to try to establish the *relative* magnitude of these unobservable components. This is done by arbitrarily assigning a value of 1 to a year lived in good health, and then finding coefficients smaller than 1 that represent the utility value of a year lived with some degree of ill health. In the present context, this is equivalent to setting $U^{o}(1) = U^{y}(1) = 1$ for both the old and the young, and

then assigning coefficients $U^o(0)/U^o(1) \equiv f^o(0) < 1$ and $U^y(0)/U^y(1) \equiv f^y(0) < 1$ to years lived in ill health by old and young individuals respectively.[7] The value of the discount factor β may be interpreted as a measure of the relative utility of a year lived in good health by an old person. In most applications, β is regarded as a subjective discount rate and is often referred to as a rate of time preference. In the present context, it may simply be regarded as a parameter of the utility function which incorporates both time preference, any difference in the utility function that is purely a function of age, and any adjustment due to the fact that the period of retirement may be shorter than the working age.

The best-known method for estimating the quality of life factors is the standard-gamble technique. In its simplest form, it is based on asking individuals what the maximum increase in the risk of death is that they would be prepared to accept in exchange for a specified improvement in their quality of life over some period of time.

In the present framework, a change in the risk of death can be represented by changes in either s^y (for a young person) or s^o (for an old person), while a change in the expected quality of life corresponds to a change in either q^y or q^o. A truthful answer to a standard gamble question for a given change in either probability q^j presumably yields that change in survival probability that leaves expected utility constant.

Specifically, suppose individuals are asked standard gamble questions involving a trade-off between survival probability and an improved (expected) quality of life both when they are young and when they are old. That is, individuals are given a specific value of Δq^j and are asked what decrease in survival probability Δs^j would leave them equally well off, for $j=y,o$. A critical

issue now is how the answers should be interpreted. Suppose first that, when answering the questions, *individuals take their consumption and labour supply in each state as independent of the probabilities.* In that case, from (5) to (8) it can be shown that their answers will satisfy the following equations:

$$\Delta s^y\left(\frac{1}{\lambda}[\overline{U}^y + \beta s^o \overline{U}^o]\right) = \Delta q^y\left(\frac{1}{\lambda}[U^y(1) - U^y(0)]\right) \qquad (13)$$

and

$$\Delta s^o \overline{U}^o / \lambda = \Delta q^o [U^o(1) - U^o(0)] / \lambda \qquad (14)$$

If β is known, equations (13) and (14) can be used to solve for the quality of life factors $f^j(0)$, $j=y,o$. Note that the value of λ need not be known to do this, since it cancels out in each equation. Moreover, the value of β can be estimated based on a standard gamble question that involves a tradeoff between changes in the probability of illness and survival of the young and old, respectively, such as what improvement in life expectancy in old age would be equivalent to a given improvement in the expected quality of life for a young person. Again assuming that the question was answered on the assumption that a person's consumption and labour supply in each state were given, a truthful answer would satisfy:

$$\Delta q^y (1/\lambda)[U^y(1) - U^y(0)] = \Delta s^o\left(\beta \overline{U}^o / \lambda\right) \qquad (15)$$

which, together with (13) and (14) could be used to solve for β. Again, this can be done without knowing the value of λ. In the following discussion, I will refer to standard gamble questions that satisfy (13) to (15), and the associated QALY index, as case A, or the "conventional case".

Inspection of (9)-(12) now makes it clear that the quality of life factors fj(0) and beta can be used to calculate the values of the unobserved components (the first terms) of the net marginal benefits of increases in each of the four different kinds of health care spending, up to a factor of proportionality given by $(1/\lambda)$. That is, marginal changes in a QALY index constructed in this way measures the individual's willingness to pay for the different kinds of health benefits, up to a given factor of proportionality. Note also that the quantity $(1/\lambda)$ can be interpreted as the willingness to pay for one life year in good health. Thus, a QALY index calculated in this way together with, e.g., an estimate of the willingness to pay for a statistical life of a young person, can be used to estimate the marginal willingness to pay for the health benefits associated with all four kinds of medical services.

However, taking state-specific consumption levels and labour supply as given in answering the standard gamble question neglects the fact that the probabilities q^i and s^i enter the representative individual's life-time budget constraint. If standard-gamble questions are answered taking this into account, the answers will not satisfy (13) to (15) but instead the following equations:

$$\Delta s^y\left(\frac{1}{\lambda}[\overline{U}^y + \beta s^o \overline{U}^o] - (N^y + r s^o N^o + R_{sy})\right) = \\ = \Delta q^y\left(\frac{1}{\lambda}[U^y(1) - U^y(0)] - \Delta N^y\right) \tag{16}$$

$$\Delta s^o\left((\beta/\lambda)\overline{U}^o - r(N^o + m_{qo})\right) = \Delta q^o\left((\beta/\lambda)[U^o(1) - U^o(0)] - r(c_1^o - c_0^o)\right) \tag{17}$$

$$\Delta q^y\left((1/\lambda)[U^y(1) - U^y(0)] - \Delta N^y\right) = \Delta s^o\left(\beta\overline{U}^o/\lambda - r(N^o + m_{qo})\right) \tag{18}$$

Equations (16) to (18) cannot be used to estimate the quality adjustment factors $f^i(0)$ and β without knowledge of λ. That is, if the answers to standard gamble questions are interpreted in this way, they cannot be used to construct an index of QALYs, as conventionally defined. I will henceforth refer to this situation as case B.

While answers to standard gamble questions cannot be used to calculate the conventional quality adjustment factors in case B, they *can* be used to estimate the relative value of the four different types of health care spending. For example, if the term multiplying Δs^y on the left-hand side of (16) (that is, the value of increasing the survival probability of a young person) is taken as the numeraire, (17) and (18) can be used to establish the relative values of increased expected quality of life of the young and the old, and increased survival probability among the old.[8]

3. Implications

In the following paragraphs, the implications of the previous analysis for the questions raised in the introduction are considered. Specifically, I consider the following three questions: 1) To what extent is cost-utility analysis consistent with conventional cost-benefit analysis in a second-best sense? 2) When evaluating projects that involve mortality reduction, should the expected consumption of those who, without it, would not survive, be treated as a cost of the project? 3)To what extent does CUA based on QALYs imply a bias against allocating resources to improve the health of elderly individuals?

1. Is CUA consistent with CBA?

If information were available on individuals' ex ante willingness to pay for marginal changes in life expectancy and probability of illness, it would be possible to perform a cost-benefit analysis (CBA) on a project consisting in increasing the amount of spending on any of the

four different kinds of health services considered in this model. By extension, it would then be possible to determine the optimal amounts of spending on each kind of service; denote this allocation as the first-best optimum. However, since the information contained in the answers to standard gamble questions only establish individuals' *relative* valuation of these services, such answers are not sufficient to find the first-best optimal amounts of spending. Nevertheless, CUA based on answers to standard gamble questions can be used to attain certain kinds of second-best optimality.

Specifically, an often-heard suggestion is that CUA is consistent with CBA in the sense that it can be used to allocate a given total health care budget in the same way that such a budget would be allocated on the basis of CBA. In the present model, this is equivalent to the statement that allocation based on CUA would maximize the expected utility of a representative individual for a given total amount of health services spending.

In terms of the earlier analysis, this suggestion is correct if the information that is used in performing CUA is in the form of answers to ex ante standard-gamble questions as in case B. Since (16) to (18) can be used to determine the relative marginal benefit (taking into account both the direct health effects and the private opportunity cost of a life-year in each state) of a dollar spent by the government on any of the four kinds of health services. A reallocation of health care spending from interventions with a high ratio of (public) cost per unit of benefit, toward with a lower cost, will therefore necessarily be welfare-increasing. At the second-best optimum, spending should be allocated in such a way that the marginal benefit from a dollar spent on each of them is the same. Although such an allocation would not necessarily be first-best optimal (because the total amount spent on health services might be too large or too small),

it would be second-best optimal in the sense of maximizing welfare subject to the fixed amount of public spending on health care.

Surprisingly, however, information contained in answers to standard gamble questions of the type considered in case A (i.e., those that make possible computation of QALY indexes of the conventional kind) cannot be used to establish second-best optimality in this sense. To see this, consider again equations (9)-(12). Using the conventional approach, the benefits of each kind of health improvement is measured by the first term which can be calculated from a QALY index up to a factor of proportionality ($1/\lambda$). The cost, on the other hand, consists of the sum of the second and third terms (the consumption and public costs, respectively). Thus it is possible to calculate the relative marginal cost per QALY of each intervention, as the ratio of the public plus consumption cost (the sum of the second and third terms) to the marginal QALY gains which are proportional to the first term. Intuitively, it would then seem that reallocation of a fixed health services budget from interventions with a high cost per QALY to those with a lower cost per QALY would necessarily increase welfare.

The reason why this intuition is incorrect is that the assumption of a fixed budget for health spending only refers to the public cost, while the incremental social cost of each type of care should include the consumption cost as well. Thus, even if the total public cost is held fixed as resources are reallocated, the total incremental resource cost (including consumption cost) may increase or decrease. Depending on the absolute marginal value of health services, this effect may increase or decrease welfare. If it decreases welfare (for example, if the result of the reallocation is an increase in the total public plus consumption cost, and the marginal social benefit of each kind of health services spending is negative to begin with), this effect may be

large enough to offset the beneficial effect of a reallocation toward relatively more efficient types of care.[9]

While the conventional cost-per-QALY criterion is inconsistent with CBA in the sense that it cannot be used to maximize expected utility subject to a fixed budget, it *can* be used to establish a second-best optimum in a different sense. Specifically, it can be used to ensure that a given health target (that is, a target value for the expected number of QALYs for a representative individual) is met at the lowest possible cost to society . The second-best efficiency condition for this case is that the total marginal cost (the sum of the public cost and the [private] consumption cost, as defined above) of generating an additional QALY through any of the four types of spending, should be the same for each type.

These two types of second-best efficiency conditions are closely related, and for a given allocation, efficiency in one sense implies efficiency in the other sense. However, if one starts from an inefficient spending pattern, one will arrive at two different second-best efficiency points depending on which of the two cases applies. The reason, of course, is that different things are held constant. If resources are reallocated subject to a fixed budget, the expected number of QALYs for the representative individual will be larger at the efficient equilibrium than it was initially. If they are reallocated subject to a fixed QALY target, the total (public plus consumption) cost of health care will be lower under the new allocation.

The distinction can be important in practice. Consider, for example, the Oregon experiment in which the objective was to find the second-best way of spending a fixed health care budget on different kinds of health services under the Oregon Medicaid plan. If decisions on spending reallocation (in comparison with some initial pattern) were made as in case A (that is,

on the basis of the estimated public plus consumption cost per QALY for each type of intervention), there is no guarantee that the resulting allocation would imply a higher expected utility for a representative individual than in a given initial allocation.[10]

2. Should consumption of survivors be treated as a cost?

From the preceding analysis, it is clear that the anwer to this question is directly related to the distinction between the public cost and the consumption cost of health improvements: The opportunity costs of the incremental consumption of young and old survivors are just the second terms in equations (9) and (11) that measure the profitability of mortality-reducing interventions for young and old individuals. (Incremental consumption costs also exist for health services that improve the quality of life, if the equilibrium consumption and labour supply differs between well and ill individuals; the relevant effects are measured by the second terms in (10) and (12).)

Specifically, the answer is that if the CUA is performed on the basis of case A (using conventionally defined QALYs), the decision-maker should evaluate consumption costs and include them in the calculation when the relative social value of the different health services is established, since, by assumption, the answers to the standard gamble questions used to establish the QALY index do not take these costs into account. Conversely, if the CUA is based on relative valuations established as in case B, the answers to the underlying *ex ante* standard-gamble questions are already supposed to have taken the consumption costs into account, so that the decision-maker should only take account of the public costs.[11]

Since practical applications have generally been based on conventional QALYs (case A), the conclusion regarding case B may not seem important in practice. However, in the next section I will argue that, in reality, it is not clear whether the QALYs used in practical

applications have in fact conformed to case A so that, in some cases, valuations that have included incremental consumption costs may have yielded misleading results.

3. Does CUA imply an age bias?

An implicit argument in the preceding discussion is that the appropriate standard of reference when analyzing the properties of CUA is a version of constrained cost-benefit analysis. The conclusion above was that, when properly used, both forms of CUA considered in this paper result in the same allocation as would result from cost-benefit analysis subject to the relevant constraint (that is, attaining a given target value of a QALY index in case A, and staying within a given aggregate health care budget in case B). Against this standard of reference, therefore, neither implies an "age bias".

As just noted, practical applications of CUA have been based on QALYs as defined in case A. In case A, the answers to standard gamble questions that should be used to establish QALY indexes as conventionally defined are implicitly assumed to take the state-specific levels of consumption and labour supply as constant and independent of the changes in probabilities of death and health status on which they are based. In the next few paragraphs, I will argue that when individuals are asked standard gamble questions in practice, they are unlikely to answer them based on these assumptions, and that, as a result, conventional QALY indexes may lead to misleading conclusions regarding the true relative valuations of different health services. In particular, they may imply biassed estimates of the relative value of health services produced for the young and the old.

To see this, consider the question under what conditions it is reasonable for individuals to take the state-specific levels of consumption and labour supply as given. The reason why the

probabilities q^j and s^j enter the budget constraint (3) is, implicitly, that they affect the state-contingent prices on the basis of which individuals are assumed to make their consumption and labour supply decisions. Intuitively, this would be something real-world people are unlikely to take into account when considering the answer to a standard gamble question, partly because some state-contingent markets simply do not exist, partly because of the implicit assumption that the state-contingent prices in (3) reflect the relevant probability for each individual, rather than for a representative individual. This line of argument might seem to support an interpretation of answers to standard gamble questions as in case A.[12]

However, the absence of certain state-contingent markets also means that in reality, young individuals typically cannot buy income replacement insurance that fully covers them against income losses they may incur if they become ill when they are young. Therefore, young individuals are likely to allow for the expected loss of life-time income when considering the consequences of ill health when young. That is, they are likely to include implicitly a term (possibly large) similar to ΔN^y on the right hand side of (16).

The implication of the preceding argument is that their answer to a standard gamble question involving the valuation of ill health when young (that is, the valuation of interventions that change q^y) may yield a relatively larger value than would be obtained if the question were answered as in case A. In other words, individuals would indicate that they would be willing to accept a relatively large decrease in survival probability in order to compensate for a given decrease in the probability of ill health when young, because they would take into account the expectation that they would have to reduce their consumption, both when they were young and when they were old, if they were in ill health when they were young. As a result, a computation

of quality of life factors on the basis of (13) to (15) would tend to produce an underestimate of $f^{y}(0)$ (and therefore, an overallocation of resources to improve the probability of good health among the young). Similarly, if the discount factor β is computed on the assumption that (15) holds, but if in reality the consumer's perception of the term multiplying Δq^{y} in (15) is closer to the corresponding term in (18), the procedure will tend to produce a relative value of increasing survival in old age that is too low. Thus in practice, applications of CUA based on QALYs as conventionally defined may indeed imply an age bias if the incremental consumption of survivors is treated as a cost of the project.

4. Conclusions

The most important conclusion of this paper is that, provided the quality adjustment factors used in QALY-based CUA have been established from truthful answers to well-defined standard gamble questions, CUA can be used to attain allocations of health resources that are second-best optimal in one of two senses. In what I have referred to as the conventional case, it can be used to minimize the social cost of attaining a given population health target. In the alternative case (when it is assumed that the standard gamble questions are answered taking lifetime net consumption effects into account), CUA can be used to maximize expected utility subject to a given health care budget. However, CUA based on conventional QALYs cannot be used to allocate a fixed health care budget efficiently.

From a practical point of view, an important implication of the analysis is that, in either case, the only difference between CUA and CBA is that the latter presupposes knowledge of exactly one additional parameter (λ in the formal analysis). This parameter can be interpreted as the (inverse of the) representative individual's willingness to pay for a healthy life year. If an

estimate of this willingness to pay can be obtained from evidence based on individual behaviour with respect to any health risk (such as studies estimating the value of a statistical life), it could be combined with information from the kinds of standard gamble questions discussed in this paper to perform a full-fledged CBA for any kind of intervention. Thus the main conclusion is very much in the spirit of those who argue that when properly used, CUA applies essentially the same criteria as CBA, and that the empirical work used to establish appropriate QALY valuations can be seen as complementing that which seeks to establish willingness to pay, and vice versa.[13] However, consistency between the CUA and CBA approaches requires a clearer understanding of the assumptions that individuals are supposed to make when answering the standard gamble questions on which the quality adjustment weights in CUA are based, than has typically been recognized in the literature.

Endnotes

1. An important paper on this subject is Garber and Phelps (1997). Another recent contribution, with an extensive bibliography, is Bleichrodt and Quiggin (1999). In the models in these papers, individuals are identical ex ante. That is, they have the same utility functions and income, and are subject to the same risks of illness and death. Given these assumptions, an efficient allocation is one that maximizes the representative individual's expected utility. Thus, if QALY maximization is equivalent to expected utility maximization, use of the QALY criterion leads to efficiency in the conventional sense.

2. In his lucid survey article of CEA, Weinstein (1995) notes that CEA can be undertaken from many different perspectives, each giving rise to a different definition of efficiency. In the present paper, the focus is on what Weinstein calls the societal perspective, since that is the perspective taken in CBA and in conventional microeconomic analysis of efficiency.

3. This issue is most extensively considered in the important paper by Meltzer (1997). The question of age bias is also closely related to the way future life years are discounted, a subject reviewed in Viscusi (1995).

4. This specification of the life-time budget constraint is similar to that in Meltzer (1997). Implicitly, it presupposes a complete set of state-contingent markets for consumption goods and labour.

5. In general, the values of consumption and labour supply will depend on the survival probabilities and the probabilities of being in a state of ill health, and thus on the amounts spent on the different kinds of health services. However, the envelope theorem ensures that the effects of induced changes in consumption and labour supply on expected lifetime utility are zero.

6. In applied work, what I have termed "consumption costs" are often referred to as "indirect costs".

7. Although I write the state-dependent utility levels here as functions of the health state only, it is understood that they also depend on the (optimally chosen) level of consumption as well.

In the papers by Garber and Phelps (1997) and Bleichrodt and Quiggin (1999), the approach is different from the one used here in that the quality adjustment factors are constants that are independent of variables such as consumption and labour supply. The state-dependent utility levels used in computing expected utility are then written as (discounted) products of these factors and a time-invariant function of consumption. With this formulation, it is possible to define a measure of expected discounted QALYs that is independent of consumption. The papers then ask the question under what conditions maximization of the QALY index is equivalent to the maximization of expected utility, and find that this equivalence only holds if consumption is constant over time.

8. An advantage with choosing interventions that increase the survival probability of a young person as the numeraire is that the willingness to pay for such interventions corresponds to the "value of a statistical life", a concept that has been the subject of extensive empirical work. Since the scale of measurement is arbitrary, such a numeraire could of course also be expressed as a specific number of (discounted) "life years in normal health".

9. This conclusion is consistent with that in Meltzer (1997, pp. 41-2). In his paper, Meltzer criticizes Weinstein whose work on cost-effectiveness has focussed on efficient allocation of a fixed health care budget (see the quote of Weinstein (1986) on p. 35 in Meltzer's paper). The implication of the discussion above is that provided the QALY index has been defined as in case B, a fixed health care budget can indeed be efficiently allocated via CUA.

In his 1995 paper, Weinstein discusses the inclusion of indirect costs and benefits in the numerator of cost-per-QALY ratios, but describes this practice " ... as controversial for economic and ethical reasons. The economic argument is that these gains and losses are already captured by quality-of-life adjustments in the denominator of the ratio" (p. 83). This argument is correct provided the QALY index has been defined as in case B. Thus the results in the present paper suggest that the disagreement between Meltzer and Weinstein in this regard essentially hinges on the definition of the QALY index.

10. A good description of this aspect of the Oregon experiment can be found in Kaplan (1995).

11. In their analysis of this issue, Garber and Phelps (1997) arrive at the puzzling conclusion that cost-effectiveness analysis will yield the same result whether or not "unrelated future costs" are included in the definition of costs. They derive this conclusion by showing that in their model, if the cost-effectiveness ratio for two interventions, *a* and *b*, are the same when these costs are excluded, they will also be the same when they are included.

This result appears to stem from their implicit assumption that *a* and *b* are related so that they must be varied together (see the definition of the term *z* on pp. 12-14 in their paper). If one assumes that *a* and *b* can be varied independently (which must be possible if they are to be evaluated independently) and sets $z = 0$, the result will not hold.

12. In particular, answers to standard gamble questions designed to establish the relative value of mortality reductions and increased expected quality of life among old individuals are also likely to satisfy the conditions relevant to case A: If the standard gamble question is asked of old persons whose savings and labour supply decisions when young have already been made, there is no reason why they should take into account the opportunity cost (in the form of additional savings required when individuals are young) of an old survivor's consumption.

13. Papers that emphasize this approach include Phelps and Mushlin (1991), Johanneson *et al* (1996), and Kenkel (1997).

References

Bleichrodt, H. and J. Quiggin (1999), Life-cycle preferences over consumption and health: When is cost-effectiveness analysis equivalent to cost-benefit analysis? Journal of Health Economics 18, 681-708

Garber, A. M., and C. Phelps (1997), Economic foundations of cost-effectiveness analysis. Journal of Health Economics 16, 1-32

Johanneson, M., B. Jonsson, and G. Karlsson (1996), Outcome measures in economic evaluation. Health Economics 5, 279-96

Kaplan, R. M. (1995), Utility assessment for estimating quality-adjusted life years. Chapter 3, pp. 31-60 in F. Sloan, ed., Valuing health care ...

Kenkel, D. (1997), On valuing morbidity, cost-effectiveness analysis, and being rude. Journal of Health Economics 16, 749-57

Meltzer, D. (1997), Accounting for future costs in medical cost-effectiveness analysis. Journal of Health Economics 16, 33-64

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

Sloan, F., ed. (1995), Valuing health care: Costs, benefits, and effectiveness of pharmaceuticals and other medical technologies. Cambridge University Press.

Viscusi, W.K. (1995), Discounting health effects for medical decisions. Chapter 7, pp. 125-148, in F. Sloan, ed., Valuing health care ...

Weinstein, M. C. (1995), From cost-effectiveness ratios to resource allocation: Where to draw the line? Chapter 5, pp. 77-97 in F. Sloan ed., Valuing health care ...

Weinstein, M. C. (1986), Challenges for cost-effectiveness research. Medical Decision Making 6, 194-8