Discounting for Health Effects in Cost Benefit and Cost Effectiveness Analysis

Hugh Gravelle[*] Dave Smith[**]

***Abstract*:** When health effects can be valued in monetary terms, as in CBA, they should be discounted at the same rate as costs. If health effects are measured in quantities (eg QALYs), as in CEA, and the value of health effects is increasing over time, discounting the volume of health effects at a lower rate than costs is a valid method of taking account of the increase in the future value of health effects. We show that the Keeler-Cretin paradox, often used as an argument against discounting health effects at a lower rate than costs, has no relevance for the choice of discount rate in CEA. We present individualistic and welfare models to argue that the rate of growth of the value of health effects g_v is positive. The welfare model suggests that g_v is a weighted average of the rate of growth the value of the direct effect of health on utility, the growth rate of income, and the growth rate of income times the elasticity of the marginal utility of income.

Keywords: discounting, economic evaluation, value of health
JEL Code: I18, H43

National Primary Care Research and Development Centre, Centre for Health Economics, University of York, Heslington, York YO10 5DD; email: hg8@york.ac.uk. We are grateful for comments from participants in the CHE economic evaluation seminar and the Alan Williams *Black Bull* seminar. Support from the European Commission under contract F14P-CT96-0056 and from the Department of Health to the National Primary Care Research and Development is acknowledged. The views expressed are not necessarily those of the funders.
[**] Centre for Health Economics; email dhs2@york.ac.uk

1 Introduction

There is an ongoing methodological debate about the appropriate way to take account of future health effects in evaluations. The majority view, most recently and comprehensively expounded in Lipscomb, Weinstein and Torrence (1995), is that benefits and costs should be discounted at the same rate. The view dominates the recommendations on discounting by government agencies, regulatory bodies, learned journals and leading health economics texts (Smith and Gravelle, 2000). A smaller body of literature favours a lower rate for health effects than for costs. The most influential example is Parsonage and Neuburger (1992), written by two UK government economists and later reflected in the UK Department of Health recommendations for evaluation of health affecting interventions (Department of Health, 1996).

We suggest in section 2 that at least some of the differences between the two schools of thought arise from different implicit assumptions about the decision context. We show that cost benefit analysis (CBA) of interventions affecting health requires procedures which directly or indirectly are equivalent to discounting the *value* of future health effects at the same rate as costs (Cropper and Portney, 1990; Cropper and Sussman, 1990; Jones-Lee and Loomes, 1995). In cost effectiveness analysis (CEA), where health effects are measured in volume rather than value terms, one valid method of allowing for growth in the value of future health effects is to discount the *volume* of future health effects at $r_h = r_c - g_v$, where r_c is the discount rate applied to costs and g_v the rate of growth of the value of health. An equivalent procedure in CEA is to adjust the volume of health effects by g_v and to discount at the same rate as costs. Thus, providing the context is correctly specified and account properly taken of the changing value of health the two views can be reconciled. Unfortunately, the preponderance of official and semi-official recommendations for CEA is to use the same discount rate for costs and benefits and not to adjust the volume of health effects to allow for the growth in their value.

One barrier to reconciliation of the two views on discounting is the paradox set out in Keeler and Cretin (1983). They show that under CEA certain types of worthwhile projects will be indefinitely postponed unless the same discount rate is used for costs and health effects. Previous responses to the Keeler-Cretin paradox have argued that such projects are very peculiar and never occur in practice or that the paradox does not arise if constraints on funding in any period are recognised (Parsonage and Neuburger, 1992; van Hout, 1998). But the discounting procedure used in assessing projects should give the correct answer irrespective of the project. The Keeler-Cretin paradox points to a logical problem with using different discount rates for costs and health effects in CEA and cannot be dismissed on empirical or practical grounds. We demonstrate in section 3 that the Keeler-Cretin paradox reveals a fundamental difficulty with CEA, though not with CBA. It does not arise because of the use of different discount rates for costs and health effects. The paradox is simply irrelevant to the choice of discount rate for health effects.

The crucial issue is whether the value of health effects is constant over time. A number of authors (Parsonage and Neuburger, 1992; Viscusi, 1995; van Hout, 1998;

Brouwer, van Hout and Rutten, 2000) have suggested that the value of health grows over time and that as a consequence the discount rate on health effects should be less than the discount rate on costs. Lipscomb, Weinstein and Torrence (1995) are the most influential proponents of the majority view that costs and health effects should be discounted at the same rate in CEA. They recognise the possibility that the value of health may be increasing over time and its implications for discounting health effects. But they conclude that "the case for such global adjustments in CEA conducted from a societal perspective has yet to be fully made, in our judgement." (Lipscomb, Weinstein and Torrence, 1995, page 234.) Their reluctance to accept the implications of a positive growth rate in the value of health may in part be due to the absence in the literature to date of arguments based on explicit models with conventional assumptions.

Accordingly, we set out in section 4 two simple models to underpin the more informal arguments which suggest that the value of health grows over time. The first is based on a behavioural model of individual choice of health affecting activities. The second uses the social welfare framework familiar from discussions of the choice of the social discount rate (Layard and Glaister, 1994, Introduction). The framework has been used to argue that the discount rate to be applied to future income changes should be $r_c = \rho + g\varepsilon$, where ρ is the rate of discount applied to future utility, g is the growth rate in income and ε is the elasticity of the marginal utility of income.

We extend the social welfare framework to incorporate health which is valued in its own right and may affect income. We show that the rate of growth of the value of health (g_v) is a weighted average of the rate of growth of the direct utility effect of health (k), the rate of growth of income g, and the rate of growth of income times the elasticity of marginal utility of income ($g\varepsilon$). The weights depend on the extent to which the income loss from ill health is borne by the individual or is covered by insurance.

If health has no effect on income and the utility effect of health is constant over time, the value of future health effects in terms of future income grows at the rate at which at the marginal utility of income falls over time: $g_v = g\varepsilon$. As a consequence the only reason for discounting future health effects is that future utility is intrinsically less valuable and the discount rate on health effects (r_h) should be the rate at which future utility is discounted: $r_h = r_c - g_v = \rho$. In another special case, when health affect income but had no direct utility effect, g_v is equal to the growth rate in income g and the discount rate on health effects is $r_h = \rho + (\varepsilon - 1)g < r_c$.

Typically it is suggested (Arrow, 1995) that ρ is about 1%, ε around 2 and g is 2% to 2.5%, yielding discount rates on costs of 5% to 6%. The discount rate on health effects in the two special cases is about 1% when health has no effect on income and about 3% to 3.5% when it only affects income.

2 Discounting for decision making: two equivalent procedures

Decisions about interventions with consequences for time streams of costs (reductions in income) and health requires a set of judgements about the relative values of health

and income at different dates. Consider a two period example of an intervention which changes present and future costs by Δc_0 and Δc_1 and the quantities of present and future health by Δh_0 and Δh_1. We can summarise value judgements or social preferences over income and health streams in a social welfare function $W(y_0, h_0, y_1, h_1)$ where y_t, h_t are income and health in period t (Jones-Lee and Loomes, 1995). The welfare function embodies judgements which determine the rate at which we are willing to sacrifice one good (health or income at some date) for another.

The marginal social valuation of health in period t in terms of period t income is the rate at which we are willing to give up period t income in exchange for period t health. It is the (negative of) the social marginal rate substitution between income and health in period t:

$$v_t \equiv \frac{W_{ht}}{W_{yt}} \tag{1}$$

where $W_{ht} = \partial W / \partial h_t$ is the social marginal welfare from an increase in health in period t and W_{yt} is the marginal social welfare from income in period t. Similarly for the marginal value of future income in terms of current income

$$\frac{1}{1+r_c} \equiv \frac{W_{y1}}{W_{y0}} \tag{2}$$

and the marginal value of future health in terms of current health

$$\frac{1}{1+r_h} \equiv \frac{W_{h1}}{W_{h0}} \tag{3}$$

The value judgements embodied in W define the social rate of discount on income or costs r_c in terms of the willingness to sacrifice current income for future income and the social rate of discount on health r_h in terms of the willingness to sacrifice current health for future health. In general the marginal social welfare from changes in health or income in any period depend on both income and health in that period and possibly on income and health in other periods. Until we specify both the form of the welfare function and the levels of health and income we do not know whether r_c is greater or less than r_h.[1]

Figure 1 is illustrates these definitions. Although there are four marginal valuations in the two period case $(v_0, v_1, (1+r_c)^{-1}, (1+r_h)^{-1})$ they are not independent: once three of them are specified the other is determined. Consistency requires that the marginal value of one good (health or income) in terms of another is the same whatever the route by which they are compared.

2.1 Cost benefit analysis

To decide whether an intervention is worthwhile the marginal valuations are used to convert all the consequences into equivalent amounts of a common unit of account (income or health at some date). Conventionally the unit of account is income in the

[1] Note that interest rates are dimensionless, so that it makes sense to compare the magnitudes of r_c and r_h. v_t has the dimension of income per unit of health, so that v_1 and v_0 can also be compared.

present period. Δc_1, Δh_0 and Δh_1 must be converted into equivalent changes in Δc_0 which are summed to give the present value of the intervention.

Such *cost benefit analysis* (CBA) is not the prevalent form of evaluation in health economics because of the difficulty in valuing health effects. However it is instructive to start with an outline of discounting in CBA because it has an explicit welfare theoretic foundation.

The present value of the intervention can be derived in two equivalent ways. The direct procedure values health effects in each period in terms of income of that period and then discounts the future value at the rate of interest on income r_c. The present value of the intervention under the direct procedure is

$$v_1 \Delta h_1 \frac{1}{(1+r_c)} + v_0 \Delta h_0 - \Delta c_1 \frac{1}{(1+r_c)} - \Delta c_0 \qquad (4)$$

The indirect method of calculating the present value of the intervention differs from the direct procedure in its treatment of Δh_1. It converts the change in future health into an equivalent change in current health and then applies the value of current health in terms of current income. The present value of the intervention with the indirect procedure is

$$v_0 \Delta h_1 \frac{1}{(1+r_h)} + v_0 \Delta h_0 - \Delta c_1 \frac{1}{(1+r_c)} - \Delta c_0 \qquad (5)$$

Since the two procedures are equivalent (4) and (5) must be equal, so that

$$\frac{v_1}{1+r_c} = \frac{v_0}{1+r_h} \quad \text{or} \quad \frac{1+r_h}{1+r_c} = \frac{v_0}{v_1} \qquad (6)$$

The discount rates on health and costs are the same ($r_h = r_c$) only if the value of health in a period in terms of income in that period is the same in both periods ($v_1 = v_0$). If the value placed on health grows over time ($v_0 < v_1$) then there must be a lower discount rate on health effects than on income or costs ($r_h < r_c$) and vice versa. Defining $g_v = (v_1 - v_0)/v_0$ as the growth rate of the value of health, (6) can be rearranged to get

$$r_h = \frac{v_0}{v_1}(1+r_c) - 1 = \frac{v_0}{v_0(1+g_v)}(1+r_c) - 1 = \frac{1+r_c}{1+g_v} - 1 \approx r_c - g_v \qquad (7)$$

Some of the disagreements over discounting of health effects may arise from a failure to spell out

- whether one is referring to discounting of the *value* of health effects ($v_1 \Delta h_1$) in a future period in terms of the income of that period or to the *quantity* of future health effects (Δh_1)

- what is being assumed about the rate of growth of the value of health effects (g_v).

When the value of health effects is discounted the rate of discount for income r_c should be used. If the volume of health effects is discounted the rate of discount for health effects r_h is correct. The discount rate on the quantity of health effects is less than the discount rate on costs ($r_h < r_c$) if the growth rate in the value of health is positive ($g_v > 0$).

The two procedures require exactly the same information and judgements about the marginal valuations of future cost and health effects in terms of present income. The first procedure, valuing heath in a period in terms of the income of that period and then applying the discount rate appropriate for incomes is perhaps more intuitive. It is also in line with the recommendations of Feldstein (1972). Feldstein suggested that when an intervention has complicated consequences because of its knock on effects on future investment, all the effects of an intervention be expressed in terms of consumption changes which are then discounted at the rate of discount appropriate for consumption.

2.2 Discounting in CEA

In cost effectiveness analysis (CEA) the investigator is limited to quantifying the health effects and does not place a monetary value on them. The aim is derive an *incremental cost effectiveness ratio* (ICER) for the intervention, defined as the discounted present value of incremental costs divided by the discounted sum of incremental health effects.

When projects are mutually exclusive and questions of the scale or divisibility of projects can be ignored, interventions with lower ICERs are preferred to those with higher ICERs. Interventions should be undertaken when their ICER is less than some critical value λ:

$$\frac{\Delta c_1(1+r_c)^{-1} + \Delta c_0}{\Delta h_1(1+r_h^o)^{-1} + \Delta h_0} \leq \lambda \qquad (8)$$

The crucial issue, over which most of the debate in the health economics literature on discounting of health effects has focused, is the discount rate r_h^o to be applied to health effects in CEA to calculate the ICER.

The CEA criterion is used when there is insufficient information on the value of health effects to conduct a CBA. It seems reasonable to require that the CEA criterion would yield the same decisions as CBA if there was information on the value of health effects.[2] CBA would accept projects whose discounted present value, given by (4) or

[2] The welfare foundations of CEA are discussed in Garber (2000) where it is argued that under certain circumstances CEA can lead to welfare maximising decisions. See also Phelps and Mushlin (1991). A contrary view has been expressed by Donaldson (1998) who argues that CEA and CBA are attempts to answer different questions, rather than attempts to answer the same question with different amounts of information. He suggests that CBA is concerned with allocative efficiency and CEA with technical efficiency.

(5), is positive. Rearranging the ICER decision rule (8) the project is accepted under CEA if

$$\Delta h_1 \frac{\lambda}{(1+r_h^o)} + \lambda \Delta h_0 - \Delta c_1 \frac{1}{(1+r_c)} - \Delta c_0 \geq 0 \tag{9}$$

Using (6) and (4) or (5), the ICER criterion is equivalent to the CBA decision rule if and only if

$$\lambda = v_0 \tag{10}$$

$$r_h^o = \frac{v_0}{v_1}(1+r_c) - 1 = r_h \approx r_c - g_v \tag{11}$$

Hence in cost effectiveness analysis health effects should be discounted at the rate $r_h^o = r_h \approx r_c - g_v$. The same discount rate should be applied to health effects in CEA as in the indirect procedure under CBA.[3] We argue in section 4 that the value of future health in terms of future income grows over time ($g_v > 0$), so that future health effects should be discounted at a lower rate than costs if no adjustment is made to the volume of health effects to reflect their growing value over time.

The alternative, direct, way to take account of the changing value of future health effects in CEA is to adjust the quantity of effects. The "real" quantity of future health effects can be defined as $\Delta \hat{h}_1 = \theta_1 \Delta h_1$, where θ_1 is an adjustment factor to allow for the change in the value of future effects. The CEA rule with the same discount rate applied to costs and to the "real" quantity of health effects is to accept the project if:

$$\theta_1 \Delta h_1 \frac{\lambda}{(1+r_c)} + \lambda \Delta h_0 - \Delta c_1 \frac{1}{(1+r_c)} - \Delta c_0 \geq 0 \tag{12}$$

which is equivalent to the CBA rule if

$$\lambda = v_0 \tag{13}$$

$$\theta_1 = (1+g_v) \tag{14}$$

The implications of growth in the value of health for CEA are recognised in the literature (Lipscomb, Weinstein and Torrance, 1996; Parsonage and Neuburger, 1992; Viscusi, 1995; van Hout, 1998) but have made no impact on CEA practice (Smith and Gravelle, 2000). Viscusi (1995) and Parsonage and Neuburger (1992) suggest adjusting the discount rate to allow for the growth in the value of health effects.

[3] In a "consistency" argument frequently cited in the debate, Weinstein and Stasson (1977) assume that $g_v = 0$ and then show, by comparing two projects directly against each other and indirectly via a sequence of equivalent project, that if the two comparisons are to yield the same result then r_h^o must equal r_c. Expression (11) explains why their argument is logically correct when $g_v = 0$ but has no relevance to the case where $g_v > 0$.

Lipscomb, Weinstein and Torrance (1995) favour direct adjustment of the volume of health effects. There are no logical grounds for preferring one approach to the other. The direct adjustment has the advantage of dealing with issue of the growth in the value of health explicitly and separating it from the issue of the rate of discount to be applied in CEA.

If the value of health is growing over time some method of allowing for it in CEA must be found. It is simply incorrect to use the same discount rate for health and cost effects if the value of health is growing. Unfortunately most of the official recommendations do not take account of the possibility that g_v is positive and suggest that the same discount rate be used for costs and health effects (Smith and Gravelle, 2000).

2.4 Inter and intra-generational discounting

In discussion of whether health effects occurring at date t+1 should be given the same weight as health effects occurring at date t, it is important to be clear about whether one is comparing the effects on individuals who will be aged a years at both dates (inter-generational effects) or individuals who will be aged a at date t and a+1 at date t+1 (intra-generational effects). The value v_t of the health effects of an intervention may depend on the age of the individuals affected as well as the date at which they occur.

A number of authors have suggested a method discounting future health effects which distinguishes timing and generational aspects (Lipscomb, 1989; Cropper and Sussman, 1990). The effects on individuals aged a at date t are discounted back to the birth date of the cohort at date t-a and then the discounted value at date t-a are discounted back to the decision data 0.

If the same discount rate is applied at both stages the procedure is equivalent to standard approaches. The procedure allows the possibility of using different discount rates in the two stages. For example, we might be willing to respect individuals' intertemporal preferences as regards changes in their health or income and use discount rates derived from studies of their behaviour to discount income and cost changes affecting them. But we may feel that they undervalue the welfare of future generations in their intertemporal decisions (Sen, 1967) and wish to use different interest rates when discounting their present values (at cohort birth dates) back to the present date. The two stage procedure provides a neat method of reconciling respect for individual preferences over decisions which affect them directly with a social concern for inter-generational equity.

Recognition of the distinction between inter and intra generational discounting does not alter the conclusions about the relationship between the discount rate for health effects and costs. Future health effects at date t accruing to individuals aged a should be taken into account by valuing them in terms of the income of aged a individuals at date t and then discounted back to date t-a at the rate used for income of aged a individuals at date t. The discount rate applied to the wealth of the cohort born at date a-t can then be applied to calculate the present value of the health effects.

3 Keeler-Cretin paradox

Keeler and Cretin (1983) have made a much cited argument for $r_h^o = r_c$ in CEA . They consider the following timing problem. A single period project can be undertaken once only. The costs Δc and health effects Δh are the same whatever the period in which the project is undertaken. The decision problem is to choose now the period in which the project will undertaken. If the project is undertaken in period t the discounted present value of the cost is $\Delta c/(1+r_c)^t$ and the discounted health effect is $\Delta h/(1+r_h^o)^t$. The ICER for the project undertaken at date t is

$$\frac{\Delta c}{(1+r_c)^t} \Big/ \frac{\Delta h}{(1+r_h)^t} = \frac{\Delta c}{\Delta h}\left(\frac{1+r_h^o}{1+r_c}\right)^t \tag{15}$$

which decreases with t if $r_h^o < r_c$. If $r_h^o < r_c$ the ICER indicates that the project becomes more worthwhile the longer it is delayed. With an infinite time horizon the CEA criterion will lead to the project being deferred indefinitely even if it has a very favourable cost effectiveness ratio ($\Delta c/\Delta h$) if undertaken in the present period.

Keeler and Cretin (1983) argue that, because the decision maker is "paralysed" if $r_h^o < r_c$ in such projects, it is correct to set $r_h^o = r_c$ in CEA. With $r_h^o = r_c$ the decision maker is indifferent as to the timing of the project using the ICER criterion and would be willing to pick a start date at random.

We disagree: the reason why Keeler-Cretin projects present difficulties under the CEA criterion is that the CEA decision rule is inherently incomplete and cannot cope with issues of the timing of decisions. The suggested solution of using the same discount rate on health and cost effects in CEA fails to address the underlying problem, which is in the CEA rule, not the rate of discount. Keeler-Cretin projects do not present a problem when the CBA decision rule is used and have no implications for the choice of discount rate to be used in CEA.

Suppose that the Keeler-Cretin project is worth doing in period 0 under the CBA rule: $v_0\Delta h - \Delta c > 0$. It is better under the CBA criterion to defer the project from period t to period $t+1$ if

$$\frac{v_{t+1}\Delta h - \Delta c}{(1+r_c)^{t+1}} > \frac{v_t \Delta h - \Delta c}{(1+r_c)^t} \tag{16}$$

which is equivalent, since $v_t = v_0(1+g_v)^t$, to

$$D_t \equiv r_c \Delta c - v_0 \Delta h\left(1+g_v\right)^t\left(r_c - g_v\right) > 0 \tag{17}$$

The behaviour of D_t depends on the growth rate in the value of health effects. There are three possible ranges of g_v with different implications for how D_t varies over time and therefore for the optimal timing of the Keeler-Cretin project:

(i) $0 \le g_v \le \hat{g}$. For small enough growth rates D_t is negative for all t: the project should be done immediately in period 0.[4] For example, if g_v is zero $D_t = r_c\Delta c - v_0\Delta h r_c < 0$ or, equivalently, with $v_0 = v_1 = v$, $(v\Delta h - \Delta c)/(1+r_c) < v\Delta h - \Delta c$

(ii) $\hat{g} < g_v < r_c$. For intermediate growth rates D_t is positive for small t and then becomes negative: the present value of the project at first increases with t and then decreases. Hence it is optimal to delay the project but not indefinitely. For example if $\Delta c = 100, \Delta h = 70, v_0 = 2$, r_c = 0.06, g_v = 0.025, then the project should be delayed until t = 8.

(iii) $r_c \le g_v$. When the growth rate of the value of health exceeds the discount rate on costs D_t is positive for all t: the present value of the project increases the longer it is delayed.

In case (iii) where $r_c \le g_v$ the decision maker would be "paralysed" under the CBA decision rule since the present value of the project increases the later it is undertaken. Case (iii) seems highly implausible since it implies that one should be willing to sacrifice an arbitrarily large amount of current income to achieve a perpetual increase in health from date in the future, no matter how distant and no matter how small the increase in health.[5]

The CEA rule with the Keeler-Cretin recommendation that the decision maker use a discount rate on health of $r_h^o = r_c$ leads to the decision maker being indifferent as to the start date for Keeler-Cretin projects. The Keeler-Cretin recommendation leads to correct decision, in the sense of maximising net present value only if $g_v = 0$.

If the decision maker confronted with Keeler-Cretin projects uses a CEA rule but with $r_h^o = r_h = r_c - g_v$ she will be indifferent as to the start date when $g_v = 0$ since the discounted cost effectiveness ratio will be constant with respect to the start date. The correct decision is to undertake the project immediately. When $g_v > 0$ she will be led to defer the decision indefinitely which is incorrect except in the highly implausible case in which $r_c \le g_v$. Hence the CEA rule will lead to decisions which are sub optimal in the sense of not maximising the discounted present value of the project.

CEA leads to incorrect decisions with Keeler-Cretin projects, irrespective of the choice of discount rate. The Keeler-Cretin paradox points to a difficulty with CEA for certain rather unusual projects but is irrelevant for the debate about the appropriate rate of discount for health effects and cost effects.[6] We conclude that this paradox is deceased.

---

[4] The critical value at which the CBA rule is indifferent between starting in period 0 and in period 1 is $\hat{g} = r_c(v_0\Delta h - \Delta c)/v_0\Delta h \in (0, r_c)$.

[5] With $r_c \le g_v$, r_h is zero or negative and the present value of the future perpetual increase in health is infinite

[6] The difficulty with the CEA rule is analogous to the problem with comparing the internal rate of return with some target rate of return as a means of taking investment decisions. The internal rate of return rule can lead to correct decisions (ie those which maximise net present value) only in a restricted class of projects (Hirshleifer, 1970, 51-56).

4 Is the value of health constant over time?

The growth rate in the value of health effects g_v is crucial for the choice of discount rate. We outline two models, one individualistic and one societal, to argue that the value of health grows over time.

4.1 Behavioural model

Individuals can alter their health, or their probability distributions over health states, through their lifestyles (Burgess and Propper, 1998) and occupational choices (Viscusi and Moore, 1989; Moore and Viscusi, 1990). They can trade current against future income via capital markets and they can trade income when ill for income when well via insurance markets. We examine derive an expression for the value of health v_t in terms of the individual's preferences, health technology and market prices and then consider how it changes over time.

4.1.1 Value of health

Since the literature has been summarised before (Johansson, 1995) we can be brief. Consider a very simple example of an individual with income *y* who can buy health care *x* which improves health $h(x)$ at a price *p*. (The same conclusions hold for any activity which affects health and which may directly affect utility.) Utility is $u(y - px, h(x), x)$ and the first order condition for an optimal choice of consumption of the health affecting good is $u_h h' + u_x = pu_y$. Dividing through by the marginal utility of income gives the marginal willingness to pay for health as

$$v = \frac{p}{h'(x)} - \frac{m_y^x}{h'(x)} \tag{18}$$

Since a unit increase in *x* increases health by $h'(x)$, a unit increase in health permit a reduction in consumption of *x* by $1/h'(x)$ whilst keeping health constant. An additional amount $p/h'(x)$ is freed to spend on other goods. There is also the effect of the reduction in *x* on his utility. $m_y^x = -u_x / u_y$ is the marginal willingness to pay for an additional unit of *x* ignoring its effect on health. If consuming health care directly reduces utility then $m_y^x < 0$ and the valuation of health in terms of current income is increased.

The price of the commodity *p* is observable but, even if the health care good had no direct effect on utility ($m_y^x = 0$), the marginal effect of *x* on health must be known to calculate *v*. If the health care good also directly affects utility, the potentially observable $p/h'(x)$ will under or over state the marginal value of the health change.

4.1.2 Growth in value of health

Attempts to value health by revealed or stated preference techniques have yielded a wide range of estimates. Such difficulties in measuring the value of health are one justification for cost effectiveness analysis. However, whilst CEA does not require that health be valued, it does require an estimate of the growth in the value health.

Wide variations in the estimates of the value of health render estimates of its growth rate even more problematic.

Consider the expression (18) for the valuation of health by an individual who consumes health care.[7] Suppose that health care has no direct effect on utility ($m_y^x = 0$) so that $v = p / h'(x)$. Even in this simple case *v* will change over time for three reasons: changes in the price of *x*, changes in amount of *x* consumed and shifts in the health production function due to technical progress in health care. Technical progress can be allowed for by writing the health production function as *h*(*x*,*s*). Technical progress over time is associated with an increase in the shift variable *s* which increase both health and the marginal productivity of health. The optimal amount of care consumed will vary over time with the productivity of care, its price and income.

The growth rate in the value of health, when $m_y^x = 0$ can be written as[8]

$$g_v = g_p - e_x^{h'}\left(e_p^x g_p + e_y^x g + e_s^x g_s\right) - e_s^{h'} g_s \qquad (19)$$

where g_p is the rate of growth in health care price, *g* the rate of growth of income, g_s is the rate of technical progress which shifts the health production function, $e_x^{h'}$ is the elasticity of the marginal productivity of health care with respect to *x*, $e_s^{h'}$ is the elasticity of the marginal productivity of health care with respect to the technical progress shift factor *s*, and e_p^x, e_y^x, e_s^x are elasticities of the consumption of health care with respect to its price, income and technology. The rate of growth in the value of health depends on three factors: the growth rate in the price of care, the rate of growth of health care consumption and its effect on the marginal productivity of care and the rate of technical progress and its effect on the marginal productivity of care.

The growth rate of the price of care is plausibly positive: $g_p > 0$. The sign of the second term in (19) is ambiguous. Although the marginal productivity of health care is diminishing ($h'' < 0$) the terms in the bracket in the second term are plausibly of different signs: consumption of care declines with its price, and increases with income. The effect of technical progress on the demand for care is ambiguous since the increase in the marginal productivity of care tends to increase demand and the increase in health to reduce it. The evidence suggests that the consumption of health care increases over time (Blomqvist and Carter, 1997), so that the second factor (increases in health care reducing its marginal productivity) tends to increase the value of $v = p/h_x(x,s)$.

[7] In what follows we investigate the change over time in the value of a given health effect in terms of income of the period in which the effect occurs. We are considering the value of a change occurring to individuals who are the same age at different dates, not the same individual at different ages.

[8] Differentiate $\ln v = \ln p - \ln h_x(x, s)$ with respect to *t*.

Because technical progress increases $h_x(x,s)$, the third factor tends to reduce g_v which could be therefore be negative or positive. We need assumptions about the magnitudes of the terms in (19) as well as their signs.

A positive income elasticity of demand for health or health improving goods is neither sufficient nor necessary for the value of health to increase over time. However, if only income changes over time (19) reduces to

$$g_v = -e_x^{h'} e_y^x g > 0 \tag{20}$$

so that the growth rate in the value of health depends on the rate of growth of income ($g > 0$), the income elasticity of demand for health care which, from the increasing shares of income spent on health as income increases is greater than unity ($e_y^x > 1$), and the elasticity of the marginal productivity of health $e_x^{h'} < 0$.

Determining the precise value of the growth rate in the value of health is clearly difficult, even in the simple case we examined here, but we believe that it is likely to be positive.

4.2 Welfare model

We can reach a similar conclusion using an entirely different approach in which we specify a social welfare function and use it derive the value of changes in future health in terms of future income and hence to derive the growth rate of the value of health. The approach is instructive because it is an extension of a well known framework for discussion of the rate of social time preference for consumption or income when only income enters the social welfare function and there is no uncertainty (Layard and Glaister, 1994).[9]

Suppose that all individuals live for one period, are identical except for the period in which they live and there are an equal number of individuals in each period. The results are not materially different but are more complicated to derive when individuals live for more than one period (Gravelle, 2000). The social welfare function can be written in per capita terms as

$$W = \sum_{t=0}^{\infty} \beta^t EU_t \tag{21}$$

where the expected utility of the representative individual is

$$U_t = \pi_t u_{ht}(y_{ht}) + (1-\pi_t) u_{dt}(y_{dt}) \tag{22}$$

The pure discount factor on utility $\beta = 1/(1 + \rho)$ allows for the possibility that a change in the utility of a generation counts for less solely because it arises at a later date.

[9] van Hout (1998) also uses a welfare framework to derive an expression for the discount rate on health but adopts a social welfare function which is non linear in the health variable, so that it cannot be interpreted as per capita expected utility, and does not allow for increase in per capita income arising from an increase in health.

State *h* is the healthy state in that the individual is better off in state *h* other things (income) being equal: $u_{ht}(y) > u_{dt}(y)$. We can interpret π_t as the probability of the health state in a world where health outcomes are independent or as the proportion of the population who are healthy. Endowed income when diseased is y_t and when healthy is $y_t + \ell_t$ where ℓ_t is the effect of ill health on income. We assume that endowed income in both states grows at the rate *g*.

To maintain comparability with the literature on the discount rate under certainty, assume that the utility functions have constant elasticity of marginal utility

$$u_{ht} = \frac{y_{ht}^{1-\varepsilon}}{1-\varepsilon} + K_t; \qquad u_{dt} = \phi \frac{y_{dt}^{1-\varepsilon}}{1-\varepsilon} \tag{23}$$

where the elasticity of the marginal utility of income is $-\varepsilon$ and $\varepsilon > 1$ to ensure that utility is bounded above. When ϕ = 1 the marginal utility of income is not directly dependent on the state: $u'_{ht}(y) = u'_{dt}(y)$, but there is a direct utility loss of K_t from being in the unhealthy state. With $0 < \phi < 1$ marginal utility is smaller when ill than when healthy and the case in which the state *d* is death, rather than diseased could be allowed for with $\phi = 0$.[10]

We use the welfare function to derive the discount rate on income (costs), the value of health in terms of income and the discount rate on health. We first illustrate the procedure for a simple case in which there is insurance and the marginal utility of income is not state dependent, so that the optimal insurance scheme is full cover against income losses from ill health. A planner who has access to an actuarially fair insurance scheme in each period chooses y_{ht}, y_{dt} to maximises *W* subject to the insurance pool breaking even in each period.[11] The Lagrangean for the problem is

$$L = W + \sum_t \sigma_t [\pi_t (y_t + \ell_t - y_{ht}) + (1-\pi_t)(y_t - y_{dt})] \tag{24}$$

The first order conditions are

$$\beta^t \pi_t u'_{ht}(y_{ht}) - \sigma_t \pi_t = 0; \qquad \beta^t (1-\pi_t) u'_{dt}(y_{dt}) - \sigma_t (1-\pi_t) = 0 \tag{25}$$

Given the assumption that the marginal utility of income is state independent, the optimal insurance scheme gives the insured an income in each state equal to expected income:

$$y^*_{ht} = y^*_{dt} = y^*_t = y_t + \pi \ell_t \tag{26}$$

The social value of additional income y_t is $\sigma_t = \beta^t E u'_{it} = \beta^t (y^*_t)^{-\varepsilon}$ where Eu'_{it} is expected marginal utility in period *t*.

The marginal social value of an additional unit of period *t*+1 income in terms of period *t* income (the income discount factor) is, using the envelope theorem on the Lagrangean (24)

[10] When $\phi < 1$ we assume that income is always greater than the level required to ensure that utility when healthy is greater than utility when diseased at the same income level.

[11] Essentially the same results can be derived in more complex settings in which being healthy has a direct effect on marginal utility so that the optimal insurance scheme does not equalise income across states (Gravelle, 2000).

$$\frac{1}{1+r_{tc}} \equiv \frac{\partial W/\partial y_{t+1}}{\partial W/\partial y_t} = \frac{\sigma_{t+1}}{\sigma_t} = \frac{\beta^{t+1} E u'_{it+1}}{\beta^t E u'_{it}} = \frac{\beta (y^*_{t+1})^{-\varepsilon}}{(y^*_t)^{-\varepsilon}} = \beta(1+g)^{-\varepsilon} \qquad (27)$$

where g is the growth rate in income. Using $r_c \approx \ln(1/(1+r_c))$, remembering that $\beta = 1/(1+\rho)$, the discount rate on income is

$$r_c \approx \rho + g\varepsilon \qquad (28)$$

We have the standard result (Layard and Glaister, 1994) that the planner should discount future income relative to current income because it is less valuable. First, future utility is valued less highly *per se* than current utility (ρ). Second, the increase in future utility as income is transferred from the current to the future period is smaller than the reduction in current utility because income grows over time (g) and the marginal utility of falls as income increases (ε).

The same result can be derived for the case in which marginal utility is state dependent. It also holds if we drop the assumption of optimal insurance and reinterpret L as a social welfare function, $\pi_t(y_t + \ell_t - y_{ht}) + (1-\pi_t)(y_t - y_{dt})$ as the expected public sector surplus from a possibly suboptimal insurance scheme, and assume that a certain £1 has the same social value wherever it accrues: $\sigma_t = \beta^t E u'_{it}$.[12]

Using the more general interpretation of (24) the period t value of an increase in health (an increase in the probability of the healthy state) in terms of income in period t is

$$v_t \equiv \frac{\partial L/\partial \pi_t}{\partial L/\partial y_t} = \frac{\beta^t (u_{ht} - u_{dt}) + \sigma_t I_t}{\sigma_t} = \frac{K_t}{E u'_{it}} + \frac{(y_{ht}^{1-\varepsilon} - y_{dt}^{1-\varepsilon})(1-\varepsilon)^{-1}}{E u'_{it}} + I_t \qquad (29)$$

where I_t is the gross amount of compensation the individual is paid if unhealthy (the amount of cover against ill health). An increase is health is valuable for two reasons: First, it raises utility directly (the first term in (29)). The value of the direct increase in utility in terms of current income depends on the size of the utility gain and on the marginal utility of income. At higher levels of income marginal utility is smaller and hence the monetary value of a given increase in health is greater.

Second, there is an increase in expected income because healthy individuals are more productive (the second and third terms in (29)). The value of the increase in expected income depends on the insurance arrangements. If there is full cover insurance ($I_t = \ell_t$) so that individual gets the same income whether healthy or ill all the productivity gain accrues entirely via the last term and is equal to the increase in expected income. If insurance is incomplete ($I_t < \ell_t$) some of the gain accrues to the individuals because they have a greater probability of being in the healthy state where they have a higher income.

[12] To derive the results in the simple form below we also assume that the possibly suboptimal insurance scheme maintains a constant ratio of individual income in the healthy and unhealthy states and that initially the probability of the healthy state is the same in all periods.

Using (29) we get

$$g_v = \frac{d \ln v_t}{dt} = (1-b_t)[a_t k + (1-a_t)g(1-\varepsilon) + g\varepsilon] + b_t g \qquad (30)$$

where $a_t = K_t / [K_t + (y_{ht}^{1-\varepsilon} - y_{dt}^{1-\varepsilon})(1-\varepsilon)]$, $b_t = I_t / v_t$ and k is the rate of growth of K_t the direct utility gain from better health. The marginal value of health (29) at any date depends on four factors: the direct impact of health on utility, the increase in utility from having a higher income when healthy, the marginal expected utility of income, which converts the first two utility effects into income terms, and the expected reduction in the cost of insurance. The growth rate in the value of health is the weighted average of the rate of growth of these factors where the weights in general vary over time.

When there is no productivity gain from better health (so that ℓ and hence I are zero) and health merely has a direct effect on utility then $g_v = k + g\varepsilon$. The direct effect of better health on utility may vary over time because of the effect of changes in public goods or the environment on utility. The value of health in terms of income grows more rapidly the larger the growth rate of income g and the rate at which marginal utility falls with income ε: both increase the willingness to give up income in exchange for health because reductions in income have a smaller utility consequence.

If there is a productivity gain from better health but no direct effect on utility ($K_t = 0$) then $g_v = g$. In the simple case in which there is full cover insurance the increase in expected income accrues entirely to the insurance pool and the growth rate of expected income is g since we assume that the endowed income of the sick grows at the same rate as the endowed income of the healthy. If there is incomplete insurance the difference between utility from income when healthy and income when sick decreases so that expected utility gain from better health *falls*. With income in the unhealthy state proportional to income in the healthy state the expected utility gain falls at the rate $(1-\varepsilon)g$. However, expected marginal utility of income falls even faster at the rate $g\varepsilon$, so that the value of additional utility in terms of income grows at the rate g. Hence whatever the extent of insurance and the sharing of the increases in expected income between the individual and the insurance pool $g_v = g$ when there is no direct effect of health on utility.

Table 1 summarises the implications of alternative assumptions for the growth in the value of health and the rate of discount on health effects. In addition to assumptions about social welfare (embodied in ρ and ε), technology and resources (embodied in g) Table 1 makes clear that the appropriate rate of discount on health effects in CEA also depends on the impact of ill health on individual income and the extent of insurance.

If utility from income is bounded ($\varepsilon > 1$), as is usually assumed, in the limit as income becomes large the effect of health on income becomes unimportant compared to the direct effect of health on utility and g_v tends to $k + g\varepsilon$ and the discount rate on health to $\rho - k$.

The current English Department of Health recommendation (Department of Health, 1996), based on Parsonage and Neuburger (1992), is that health effects be discounted at 1.5% and costs at 6%. A number of authors (Arrow, 1995) have suggested that ρ is around 1% and that ε is about 2. Our results suggest that the recommendation on r_c are broadly correct, but they support the recommendation on r_h only if less obvious further assumptions are made, for example that the value of the direct effect of health on utility is constant and large relative to the value of the effect of health on income.

5 Conclusions

Our conclusions can be summarised thus

- if it is believed that value of future health effects in terms of future income grows over time, the estimated health effect used in the evaluation should be adjusted or a lower discount rate for health effects than for costs.

- the direct method of allowing for the change in the value of health over time (adjusting the health effect) has the merit of being explicit and separating out the issues of the value of health effects in monetary terms and the discount rate to be applied to future income

- for cost benefit analysis all health effects should be valued in the income of the period in which they occur and then discounted back to a present value using the rate of discount appropriate for costs

- for cost effectiveness analysis, where health effects cannot be valued in income of the period, the nominal quantity of health effects should be adjusted to a "real" quantity to reflect the growth in the value of future health effects and the same discount rate be applied to costs and "real" health effects

- evaluations should be explicit about the approach taken to discounting health effects and the reasons underlying it, such as assumptions about the growth in the value of health effects

The suggestion that the discount rate for health effects should be less than the discount rate for costs because the value of health grows over time is not new. However, the facts that the correct discounting procedure for CEA was recognised but then dismissed in the chapter by Lipscomb, Weinstein and Torrence in the influential compendium of best practice in health economic evaluations commissioned by the US Public Health Service (Gold *et. al.*, 1996), the failure to use the correct procedure revealed in published studies, and the incorrect procedures recommended in many official guidelines, suggest that the case needed to be made more firmly. We hope that by using simple explicit models of intertemporal decision making, we have strengthened the case for allowing for the growth in the value of health in economic evaluations. Evaluations based on CEA criteria require estimates of the growth in the value of health and CBA is impossible without estimates of the value of health. Attention should now be turned to the fundamental issue for decision making in health care: the value of health.

References

Arrow, K. J. (1995). "Intergenerational equity and the rate of discount in long term social investment", *mimeo*, paper for International Economics Association, World Congress, December.

Blomqvist, A.G. and Carter, R.A.L. (1997). "Is health care really a luxury good?", *Journal of Health Economics*¸16, 207-230.

Brouwer, W., van Hout, B. and Rutten, F. (2000). "A fair approach to discounting future effects: taking a societal perspective", *Journal of Health Services Research and Policy*, 5, 114-118.

Burgess, S.M. and Propper, C. (1998). "Early health related behaviours and their impact on later life chances: evidence from the US", *Health Economics*, 7, 381-400

Cropper, M.L., and P.R. Portney. (1990). "Discounting and the evaluation of lifesaving programs", *Journal of Risk and Uncertainty* 3:369-79.

Cropper, M.L. and Sussman, F.G. (1990). "Valuing future risks to life", *Journal of Environmental Economics and Management*, 19, 160-174.

Department of Health. (1996). *Policy Appraisal and Health: A Guide from the Department of Health*. London. GO7/038 3901, February. 1996

Donaldson, C. (1998). "The (near) equivalence of cost-effectiveness and cost-benefit analyses: fact or fallacy", *Pharmocoeconomics*¸ 13, 389-396.

Feldstein, M.S. (1972). "The inadequacy of weighted discount rates", In R. Layard (ed.), *Cost Benefit Analysis*, 140-55. Penguin Books, London.

Garber, A.M. (2000). "Advances in cost-effectiveness analysis of health interventions", in Newhouse, J.P. and Culyer, A.J. (eds.), *Handbook of Health Economics*, North Holland, in press; also as Working Paper 7198, June 1999, *National Bureau of Economic Research*.

Gold, M.R., Siegel, J.E., Russell L.B., and Weinstein, M.C. (eds.) (1996). *Cost-Effectiveness in Health and Medicine*Oxford, Oxford University Press.

Gravelle, H. (2000). "Valuing and discounting future health changes", June.

Hirshleifer, J. (1970). *Investment, Interest and Capital*, Prentice Hall, New Jersey.

Hout, van B. (1998). "Discounting costs and effects differently: a reconsideration", *Health Economics*, 7, 581-594.

Johansson, P.O. (1995). *Evaluating Health Risks: An Economic Approach*, Cambridge University Press.

Jones-Lee, M. W. and Loomes, G. (1995). "Discounting and safety", *Oxford Economic Papers*, 47, 501-512.

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

Layard, R. and Glaister, S. (1994). *Cost-Benefit Analysis*, Second Edition, Cambridge University Press.

Lipscomb, J. (1989). "Time preference for health in cost-effectiveness analysis", *Medical Care* 27:S233-S253.

Lipscomb, J., Weinstein, M.C. and Torrance, G.W. (1996). "Time preference", in *Cost-Effectiveness in Health and Medicine*, Gold, M.R., Siegel, J.E., Russell L.B., and Weinstein, M.C. (eds.), Oxford, Oxford University Press.

Parsonage, M., and H. Neuburger. (1992). "Discounting and health benefits", *Health Economics* 1:71-76.

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

Sen, A.K. (1967). "Isolation, assurance and the social rate of discount", *Quarterly Journal of Economics*, 81, 112-124.

Smith, D and Gravelle, H. (2001). "The practice of discounting economic evaluations of health care interventions",. *International Journal of Technology Assessment in Health Care*, 2001, to appear. [See *Centre for Health Economics, Technical Paper*, No 19, July 2000]

Viscusi, W.K. (1995). "Discounting health effects for medical decisions", in F.A. Sloan (ed.) *Valuing Health Care: Costs, Benefits, and Effectiveness of Pharmaceuticals and Medical Technologies*, New York: Cambridge University Press.

Weinstein, M.C., and W.B. Stason. (1977). "Foundations of cost-effectiveness analysis for health and medical practices", *New England Journal of Medicine* 296:716-21.

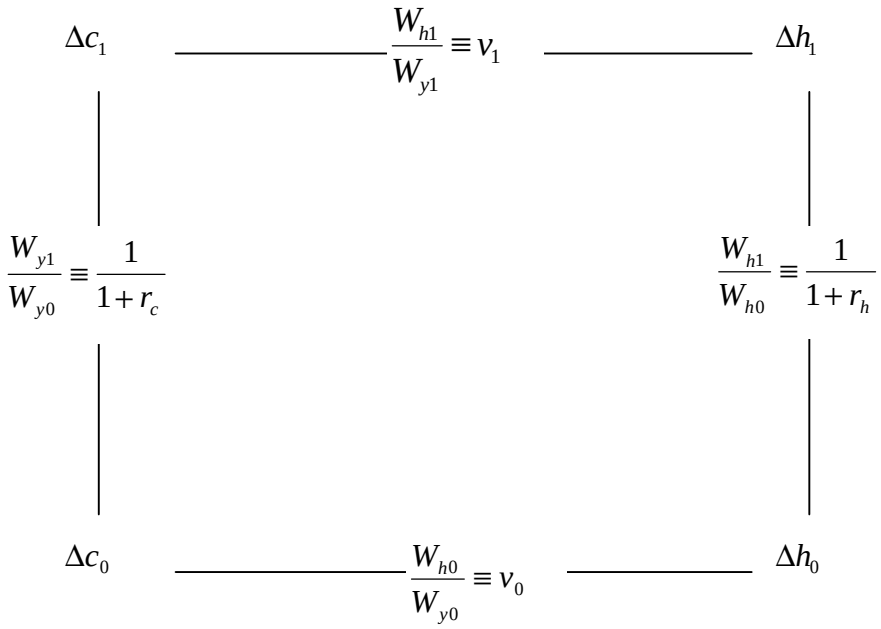

Figure 1. Valuing changes in future health in terms of current cost (or income) directly: $\Delta h_1 v_1 /(1+r_c)$; and indirectly: $[\Delta h_1 /(1+r_h)]v_0$

	Direct effect of health on utility?	
	No	Yes
Health affect income?		
No	$g_v = 0$ $r_h = \rho + g\varepsilon = r_c$	$g_v = k + g\varepsilon$ $r_h = \rho - k$
Yes	$g_v = g$ $r_h = \rho + g(\varepsilon - 1) < r_c$	$g_v = (1-b_t)[a_t k + (1-a_t)g(1-\varepsilon) + g\varepsilon] + b_t g \rightarrow k + g\varepsilon$ $r_h = \rho + g(1-\varepsilon)[(1-b_t)(1-a_t) + b_t] - (1-b_t)a_t k \rightarrow \rho - k$

g: rate of growth of income; ε: elasticity of marginal utility of income; ρ: pure utility discount rate; k: rate of growth of direct effect of health on utility; r_c: discount rate on income (costs); $b_t = I_t / v_t$

$a_t = K_t / [K_t + (y_{ht}^{1-\varepsilon} - y_{dt}^{1-\varepsilon})(1-\varepsilon)]$;. limits apply if utility from income is bounded above ($\varepsilon > 1$).

Table 1. Rate of growth of value of health (g_v) and discount rate on health effects (r_h).