CHAPTER 7.
THE MICROECONOMICS OF ECONOMIC EVALUATION OF HEALTH CARE PROGRAMMES II

1. Introduction

This is the second of two chapters on the microeconomic foundations of techniques for the economic evaluation of health care programmes. This chapter is organised as follows:

Chapter 7

1. Introduction
2. Recent work on the theoretical foundations of cost effectiveness analysis
3. The decision rules of cost effectiveness analysis
4. Cost effectiveness analysis and budget maximisation
5. The relationship between cost effectiveness analysis and cost benefit analysis

The most common approach to carrying out economic evaluations of health care programmes is cost effectiveness analysis (CEA), where the costs are expressed in monetary units and the health effects are expressed in non-monetary units such as life-years gained or QALYs gained. We saw in the last lecture that CEA is based on the maximisation of health effects for a given budget. A fundamental weakness of CEA is that its economic foundation in welfare economic theory is unclear. The classical tool of economic evaluation based on welfare economic theory is cost benefit analysis (CBA), where both costs and health effects are measured in the same units (usually money). The main purpose of this lecture is to examine the relationship between CEA and CBA, and to clarify the conditions under which both are 'equivalent', in the sense that they both lead to the same resource allocation decisions. However, we begin by briefly outlining some recent advances in clarifying the theoretical foundations of CEA.

2. Recent work on the theoretical foundations of cost effectiveness analysis

A series of recent papers in the Journal of Health Economics have explored the theoretical foundations of CEA from 'first principles'. The Garber and Phelps (1997) paper in your reading list attempts to ground CEA in von Neumann-Morgenstern utility theory. They write:

> 'Despite the widespread use of CE analysis, we are unaware of any published formal justification of the technique on the basis of first principles. The intuitive appeal of the logic of CE (minimising the cost of producing a given level of health, or corresponndingly, maximsing the achievable level of health for a given budget) *sounds like* a familiar economic problem, and for the most part practitioners have *assumed* that CE analysis could be a tool for utility maximisation. Yet, even with this broad level of agreement (unsupported by any formal proof of the conclusion), a number of thorny problems remain in the CE literature,...' (p. 2)

Garber and Phelps develop a theoretical framework which they use to 'derive' a cost effectiveness criterion to guide resource allocation decisions. They also use their theoretical framework to explore various controversies in CEA, including the problem of finding an 'optimal' cutoff CE ratio. Recall that, when apprasing a particular intervention using CEA, it is common practice to calculate its CE ratio and compare it to those of other interventions. However, under some strong assumptions, an 'optimal' cutoff CE ratio exists, and is equal to

the social marginal WTP for units of the health effect (see below). Garber and Phelps show how their model can be used to estimate this 'optimal' CE ratio (assuming it exists).

Another controversy in CEA concerns the problem of deciding which costs should be included in the numerator of the CE ratio, and which ones should be excluded. In particular, there is a question about whether *future* costs should be included. For example, suppose we implement a health care programme that saves someone's life today, but the person becomes ill in future and requires further health care; these 'future costs' would not have been incurred if we had not implemented the health care programme, and had simply let the person die ! Should we include these future costs in our CEA of the programme that saves the person's life ? One approach, called the 'decision maker' approach to CEA (see below) says that only the costs which are relevant to the particular decision maker should be included in the numerator. Thus, since the future costs will probably be borne by someone else, this approach suggests that we should not include them in our CEA of the programme that saves someone's life today. However, welfare economics says that *all* costs should be included in the numerator, no matter who they fall on.

Garber and Phelps addressed this issue using their newly developed theoretical framework and found that, in a sense, it makes no difference whether or not future costs are included. The paper by Meltzer (1997) in your reading list argues that this result was only obtained because of the strong assumptions made by the Garber and Phelps model. (These assumptions are so strong, in fact, that if they were true CEA and CBA would always give identical results !). Meltzer writes:

> 'One reason for the persistent differences in opinion about how to treat future costs is that there has been no solid theoretical basis for these arguments. Given this void in the literature, Garber and Phelps make a major contribution by proposing that the cost effectiveness methodology be evaluated by its consistency with expected utility theory. Surprisingly, they conclude that it does not matter whether future costs are included because the relative rankings of procedures will be preserved in either case, as long as future costs are treated consistently.....Despite the attractiveness of this theoretical result, the formulation of lifetime resource allocation used by Garber and Phelps contains strong restrictions on the substitution of income across time and potential outcomes'. (p. 36)

Meltzer shows that if a more general model is used, which relaxes some of the key assumptions of the Garber and Phelps model, then CEA is consistent with utility-maximising behaviour only if *all* future costs are included in the analysis, whether medical or non-medical. (This is the same conclusion one arrives at from a welfare economic perspective). For example, the utility of a hamburger eaten by someone 20 years after undergoing a life-saving intervention is a 'benefit' of that intervention; according to Meltzer's findings, the cost of the hamburger should also be included in the CEA of that life-saving intervention!

Finally, a Journal of Health Economics paper by Bleichrodt (1997), also in your reading list, argues that CEA and CBA are only concerned with the 'efficiency' criterion, and that issues of 'equity' have been neglected. He proposes some theoretically based methods for incorporating equity considerations into CEA.

The general consensus that seems to be emerging is that CEA as it is conventionally applied is simply not consistent with welfare economic theory, or even with utility-maximsing behaviour as per the von Neumann-Morgenstern approach. We will see below that CEA and CBA are only 'equivalent' under very strong and very unrealistic assumptions.

3. The decision rules of cost effectiveness analysis

We saw towards the end of the last lecture that cost effectiveness analysis is based on the maximisation of an effectiveness unit (e.g. QALYs) subject to a resource constraint. For simplicity, we will assume for the remainder of this lecture that the effectiveness unit to be maximised is QALYs gained. Imagine that there are a number of independent 'clusters' of possible health care programmes. *Within* each cluster, the programmes are mutually exclusive (e.g. competing treatments for the same illness). *Between* the clusters, the programmes are independent of each other, in the sense that implementing a programme in one cluster does not affect the cost or effectiveness of a programme in any other cluster.

As we saw in the last lecture, the mutually exclusive programmes *within* each cluster should first be ordered according to effectiveness, and then the incremental cost effectiveness ratio (ICER) for each successively more effective programme should be calculated. If any of these ICERs turns out to be less than the previous one in the sequence of increasingly effective mutually exclusive programmes, then the less effective one is ruled out as dominated, and it should never be implemented irrespective of the amount of resources available. This algorithm results in a sequence of programmes with increasing ICERs *within* each cluster.

There are now two optimising decision rules: either
(1). Specify the budget limit, and through a process similar to the one discussed towards the end of the last lecture select programmes from the independent clusters (only one from each cluster) so as to maximse the number of QALYs gained subject to the budget limit; or
(2). Specify a 'cutoff' CE ratio, and choose from each cluster the programme with the highest ICER which is equal to or below the 'cutoff'.

Note that, as we also mentioned towards the end of the last lecture, the maximisation of QALYs gained for a given budget using cost effectiveness ratios is based on the assumption that the scale of a programme can be reduced without changing the ICER (i.e. the ICER is independent of the scale of the programme). We usually refer to this assumption as constant returns to scale (CRS). One obvious case where this assumption will be violated is when programmes are indivisible, since then the scale of the programme cannot be reduced without changing the ICER. In this case, the above optimisation rules no longer work, and other techniques such as non-linear programming or integer programming will have to be used. We do not discuss this any further in this course, and will assume from now on that all programmes are prefectly divisible.

In order to use CEA for decision making, either the budget has to be decided that should be used to maximise QALYs gained, or the price per QALY gained has to be decided to determine which programmes to implement. Based on the budget maximisation decision rule, it is common to include only health care costs (e.g. doctor's labour time, cost of drugs and equipment, etc.) in the CEA. We will now briefly examine this approach to CEA.

4. Cost effectiveness analysis and budget maximisation

The argument behind CEA has often been a so-called 'decision maker' approach to economic evaluation. According to this approach the aim of economic evaluation is to maximise whatever the decision maker wants to maximise, and to only include the costs and benefits that the decision maker finds relevant. This leads to situations where only those costs that fall on the budget of a specific decision maker are included in that decision maker's analysis. In

health care, as stated above, this approach implies that only health care costs should be included in a CEA that is seeking to maximise QALYs subject to a health care budget. The decision maker approach to economic evaluation and CEA can be criticised on many grounds (see e.g. the Meltzer (1997) paper discussed above). From our point of view, the most serious criticism is that it is not compatible with welfare economics, which requires that all costs to society be taken into account. It is therefore obvious that a CEA with only health care costs included will be inconsistent with CBA, since CBA is based on including all costs and benefits irrespective of who they fall on.

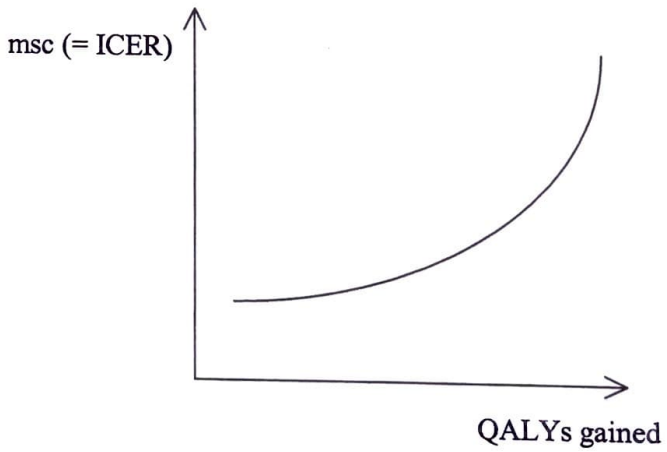
The consensus in the economics literature is that the decision maker approach to CEA is theoretically unjustifiable. The phrase you will commonly encounter is that 'it can lead to serious problems of sub-optimisation'. Our attention should therefore be focused on investigating the relationship between CEA and CBA in cases where both types of analyses include all costs and health effects (benefits) irrespective of who they fall on. CEA based on the decision maker approach is a 'non-starter' as far as comparisons with CBA are concerned.

5. The relationship between cost effectiveness analysis and cost benefit analysis

5.1. The case of constant WTP per QALY gained

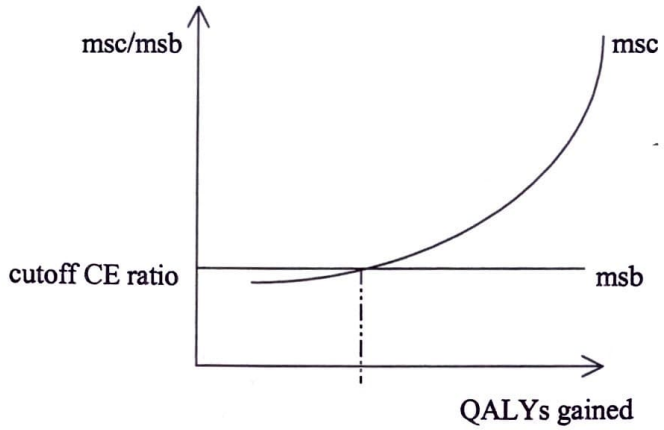
We now focus on the second type of problem above, where a 'cutoff' ratio is specified, and used to pick out from each cluster that programme which has the highest ICER that is equal to or less than the 'cutoff'. In this case, it can easily be shown that CEA and CBA will always give the same results when
(1) all societal costs and benefits are included in the CEA; and
(2) the marginal social benefit of a QALY in monetary terms (i.e. society's WTP for a QALY) is constant for all sizes of the change in QALYs, the same for all individuals under all circumstances, and is equal to the cutoff CE ratio.

To see this, recall that the mutually exclusive programmes within each cluster are in order of rising effectiveness and ICER (we have excluded all dominated programmes). Plotting the ICERs in a given cluster against QALYs gained gives us an upward sloping curve, which can be interpreted as the marginal social cost (msc) curve for producing QALYs in that cluster:



The msc curve is drawn as a smooth function of QALYs, based on the simplifying assumption that there are very many programmes in a cluster, each representing a small increase in the

ICER (i.e. msc) as the number of QALYs gained rises. We can add a horizontal msb curve to represent each individual's (and society's) constant marginal WTP for QALYs:

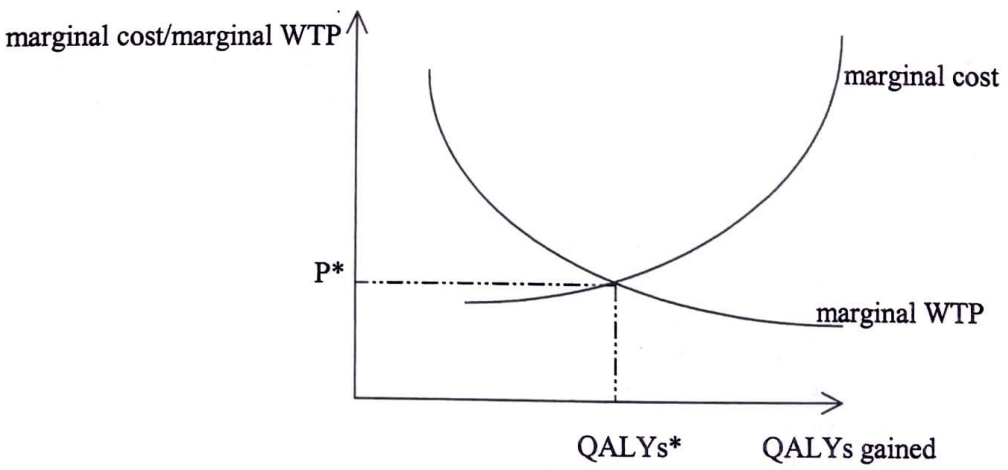


Note that the msb is assumed to be the same for all individuals. If we now set the cutoff CE ratio equal to society's marginal WTP for a QALY (i.e. the horizontal msb curve) CEA will lead us to select the programme at which the msb and msc curves intersect. At the point of intersection we have msb = msc, so we have the same procedure as in CBA: we increase the amount of QALYs until msb = msc for every cluster. Thus, with constant WTP per QALY gained, and identical WTP for all individuals, CEA and CBA yield the same result, provided that the WTP per QALY gained is used as the cutoff ratio in CEA. This means, however, that we need information about the social WTP per QALY gained to provide CEA with a theoretically justifiable decision rule.

5.2. The case of diminishing WTP per QALY gained, and individual heterogeneity

The assumptions of constant WTP per QALY gained and identical WTP for all individuals do not seem very realistic, and it is therefore important to investigate the relationship between CEA and CBA if these assumptions are relaxed. We will therefore consider the case where the WTP per QALY can vary (both with the number of QALYs, and across individuals), but CEA is based on a constant 'cutoff' ratio which is some approximation of the societal WTP.

We start by considering a single individual. The individual's marginal cost curve for producing QALYs in a given cluster (which is the same for all individuals), and the individual's marginal WTP for QALYs are shown in the following diagram:



As before, the marginal cost curve for a cluster can be interpreted as a number of increasingly more effective mutually exclusive programmes available to the individual, with increasing ICERs. The curve is drawn smooth based on the assumption that every programme leads to a marginal cost increase. The optimum for the individual is the point where the marginal WTP for QALYs gained equals the marginal cost for QALYs gained (denoted by P* in the figure). The marginal cost curve shows the marginal ICER of different health programmes, and if we use P* as the price per QALY gained we could implement increasingly more effective programmes until the marginal ICER equals this price. For the individual, CEA based on the price P* will thus yield the same result as using CBA (CBA being defined as the point where the marginal cost and marginal WTP curves intersect).

Even in this simple case where we are only considering a single individual, there is a serious problem which in general will lead to divergence between CEA and CBA. The problem is that, with a diminishing marginal WTP, the msc curve is not necessarily the same for both CEA and CBA ! CEA and CBA do not necessarily agree on which programmes are dominated and which are not in a given cluster, so we cannot generally draw a unique msc curve that applies to both. This problem does not arise with constant marginal WTP for QALYs.

To fix ideas, consider the following simple example involving a CRS technology. Suppose a drug can be given in one of two doses: a 'half-dose' produces 1 QALY at a cost of £50,000, and a 'full-dose' produces 2 QALYs at a cost of £100,000. The ICER of the half-dose is £50,000, and the ICER of the full dose relative to the half-dose is also £50,000. Since the full-dose is more effective than the half-dose, but costs the same per QALY, CEA says that the full-dose dominates the half-dose, and we should eliminate the half-dose from further consideration. But suppose we have a diminishing marginal WTP for QALYs, so that the marginal WTP for QALYs when we have 1 QALY is £50,000, and the marginal WTP for QALYs when we have 2 QALYs is only £40,000. Then CBA prefers the half-dose !

This would not have happened with a constant marginal WTP for QALYs. For example, suppose the marginal WTP for QALYs is £50,000, irrespective of the number of QALYs gained. Then the marginal WTP for QALYs when we have 2 QALYs is £50,000 which is the same as the marginal cost (i.e. the ICER) for the full-dose. Since CBA tells us to increase the number of QALYs until the marginal WTP equals the marginal cost, we would choose the full-dose rather than the half-dose (since it gives us more QALYs), but this is same result we got from the CEA above.

An additional problem arises if we now introduce the possibility that there are many individuals. It is easy to show in this case that CEA and CBA can lead to very different decisions, because there is no reason to believe that the marginal WTP will be the same for different individuals.

To fix ideas again, consider the following simple numerical example. Suppose we are considering implementing two different independent health programmes. One programme in one patient group costs £100,000 and yields 5 QALYs i.e. an ICER of £20,000. Another programme in another patient group costs £100,000 and yields 4 QALYs i.e. an ICER of £25,000. Suppose the price per QALY we use in our CEA is £22,500. According to the CEA, the first programme should be implemented but not the second. But suppose the marginal WTP for QALYs happens to be £15,000 in the first patient group, and £30,000 in the second

patient group. Then a CBA would give the result that the second programme should be implemented rather than the first.

In summary, when we have a diminishing marginal WTP for QALYs for each individual, and varying marginal WTP for QALYs across individuals, CEA and CBA will generally lead to different answers. The only situation in which they will *always* give the same answers is when the marginal WTP for QALYs is constant (i.e. not varying with the number of QALYs), the same for all individuals, and when the 'cutoff' CE ratio is set equal to this constant social marginal WTP.