CHAPTER 6.
THE MICROECONOMICS OF ECONOMIC EVALUATION OF HEALTH CARE PROGRAMMES I

1. Introduction

The 'bread and butter' work of health economists is economic evaluation of health care programmes. This involves measuring the costs and 'benefits' of a given health care programme, and comparing the two in some way to determine whether or not the programme is 'worth the cost'. This is a much more difficult and potentially contentious process than it sounds. For example, the US spends about $253 million per year on heart transplants, adding about 1600 years to the lives of heart patients who would otherwise die prematurely. But at an average cost of $158,000 per year of life saved, is this an efficient way to use society's scarce resources? In practice, the answer to this question depends on our basis of comparison, or 'cutoff' figure for the value of a life year saved; if the $158,000 cost per life year is greater than this, we can say that it is too high, and that heart transplants are not 'worth the cost'; if it is less, we might conclude that heart transplants are a relatively efficient way to produce extra life years.

If we believe in welfare economic theory, our basis of comparison should be some estimate of the 'marginal social willingness to pay' for life years. For example, economists have estimated that people in the US tend to make trade-offs between survival and money at the rate of about $5 million per life saved. They have also estimated that about 15 discounted years of life are saved on average when a premature death is averted. Our estimate of the social willingness to pay for a life year is therefore $5 million/15 yrs. = $333,333. This is well above the $158,000 cost per life year saved by heart transplants, which suggests that heart transplants *are* 'worth the cost'. Another way to say this is that society values the 1600 extra life-years produced by heart transplants at $333,333×1600 = $533.33 million, which exceeds the $253 million society has to pay. When we use estimates of 'marginal social willingness to pay' to compare costs and benefits in monetary terms like this, we are doing what is known as a cost benefit analysis.

However, we know that it is very difficult in practice to determine the validity of estimates of social willingness to pay. The above estimate of $333,333 per life year could easily be completely wrong. There is no simple way in which we can objectively verify it, because there is no 'market' for life years so we cannot see how much value society really does attach to them. Because of this, many health economists are skeptical about using willingness to pay measures, and prefer to use values of the cost per life year in other contexts as bases for comparison. For example, by spending $182,000 every year for sickle cell screening and treatment for black newborns in the US, 769 years are added to their lives, at a cost of only $236 for each year of life saved. This makes the $158,000 per life year saved by heart transplants seem rather expensive. Perhaps society could save more life years for the same amount of money by reducing the number of heart transplants, and using the money to finance other health care programmes. Note that we have used a comparison figure for the value of a life year ($236) that is not the social willingess to pay. When we do this, we are doing what is known as a cost effectiveness analysis.

The problem is that, in our simple example above, the two approaches seem to disagree. Cost benefit analysis (which is based on welfare economic theory) suggests that heart transplants are quite cost effective, but we have doubts about our estimate of marginal social willingness to pay. On the other hand, cost effectiveness analysis (which does not have as good a theoretical foundation) suggests that there are other health care programmes that could save more life years than heart transplants, for the same amount of money. Which of these two is 'right'? This is essentially the dilemma that is currently being faced by health economists around the world. Some argue that we should stick with the welfare economic approach, and just try to improve our methods of estimating social willingness to pay for things. Others argue that the time has come to abandon traditional welfare economics altogether.

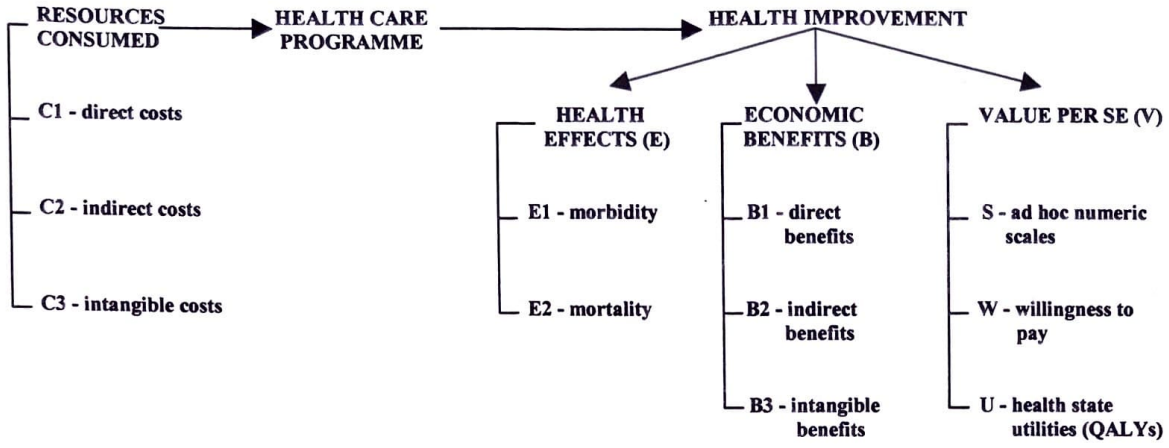
In this chapter and the next, we will investigate in detail the two approaches to economic evaluation outlined above, and their compatibility with 'traditional' microeconomic theory. In particular, we will come away with a clearer idea of the circumstances in which cost benefit analysis and cost effectiveness analysis are 'equivalent' (in the sense that they both lead to the same resource allocation decisions). We will see that these are quite restricted. In general, as the above example suggests, there will be cases in which they lead to quite different decisions. The chapters are organised as follows:

Chapter 6
1. Introduction
2. The economic evaluation of health care programmes
3. Willingness to pay
4. Quality-Adjusted Life Years (QALYs)
5. Theoretical foundations of cost benefit analysis
6. The basic cost effectiveness analysis model

Chapter 7
1. Introduction
2. Recent work on the theoretical foundations of cost effectiveness analysis
3. The decision rule of cost effectiveness analysis
4. Cost effectiveness analysis and budget maximisation
5. The relationship between cost effectiveness analysis and cost benefit analysis

2. The economic evaluation of health care programmes

A good (but somewhat dated) article from the health economics literature on this topic is a review paper by Torrance, W., 1986, *Measurement of health state utilities for economic appraisal*, Journal of Health Economics 5, pages 1-30. It is good beause it provides a nice overview of the main ways in which one can conduct an economic evaluation of a health care programme. As we have seen, this typically involves comparing the resources consumed by the programme with the health improvement generated by the programme. Torrance provides the following diagrammatic representation of the possible ways in which one can measure these costs and benefits:



There are three types of costs, reflecting the resources consumed by a programme:

C1 - these are direct costs, such as the cost of hiring the labour time of doctors, the cost of drugs and equipment used, etc.

C2 - these are indirect costs, such as the cost to the economy of individuals taking time off work in order to undergo the treatment.

C3 - these are intangible costs, expressed as the monetary value of the disutility (e.g. pain, discomfort, etc.) associated with a particular type of treatment, such as chemotherapy in the treatment of cancer.

There is also a range of possible measures of the health improvement (i.e. the 'output') of the health programme. These fall into three basic categories:

Health effects (E)
E1 - these are measures of output in terms of reduced morbidity (in the health economics jargon, 'morbidity' is just another word for 'illness'), expressed in terms of the illness itself. For example, cases of flu prevented, disability-days prevented, hospitalisation-days prevented.

E2 - these are measures of output in terms of reduced mortality e.g. number of lives saved by a particular programme, or the number of extra life years produced by a programme.

Economic benefits (B)
B1 - these are the direct economic benefits arising from the health improvement produced by a health programme, such as the reduced health care costs resulting from making people healthier and therefore less dependent on the health care system in the future.

B2 - these are the indirect economic benefits, such as the increased output in the economy as a result of reducing days taken off work due to illness, or making workers healthier and therefore more productive at work.

B3 - these are the intangible economic benefits, expressed as the monetary value of reduced disutility due to illness (e.g. pain, discomfort etc.) as a result of implementing the programme.

Value of health improvement per se, regardless of economic consequences (V)

S - these are measures on ad hoc numeric scales (e.g. 'on a scale from 1 to 10') of the value of the health improvement itself e.g. on a scale from 1 to 10 (where 1 is 'worthless' and 10 is 'indispensable') how highly would you value your health improvement as a result of having an operation on your back to reduce back ache?

W - these are measures of willingness to pay (WTP) for the benefit from a programme e.g. how much would you be WTP for the benefit of an operation on your back to reduce back ache?

U - these are measures of quality-adjusted life years (QALYs) produced by the programme. They are based on measurements of the utility associated with particular health states on a scale from 0 to 1, where 0 is the utility of death, and 1 is the utility of perfect health. If a treatment raises utility from 0.2 to 0.4 for a period of 10 years, then it has produced $0.2 \times 10 =$ 2QALYs. Alternatively, a treatment (say for breast cancer) might extend life for 5 years, in each of which utility would be 0.4. Then again the treatment produces $0.4 \times 5 =$ 2QALYs. Thus, although the 'outputs' of the two treatments are very different, they are worth the same in terms of QALYs.

Different formulas are used to combine the different cost and benefit measures in the three main methods of economic evaluation used today:

(1). Cost effectiveness analysis - This uses a cost/effectiveness ratio to measure the cost per unit of health improvement associated with a particular programme. For example, the cost in the numerator of the formula might be defined as the *net economic cost to society* in pounds sterling (C1 + C2 - B1 - B2), and the denominator might be defined as *reduced mortality* in terms of extra life years (E2). Then the cost/effectiveness ratio would measure the cost in pounds sterling per extra life year produced by the programme. The main disadvantage of this method is that it cannot be used to compare two health programmes whose effectiveness measures are expressed in different units. For example, it cannot be used to compare a health programme whose effectiveness is measured in terms of extra life years, against a health programme whose effectiveness is measured in terms of cases of flu prevented.

(2). Cost benefit analysis - This measures the *net social benefit* (NSB) of a health programme as

$$NSB = B1 + B2 + W - C1 - C2$$

The key feature of this is that everything (including the benefits) are measured in monetary terms. From a theoretical viewpoint, it would be nice to be able to include the intangibles B3 and C3 in this measure, but Torrance points out that these are usually too difficult to estimate in practice. The decision rule is simply 'adopt the programme if its NSB is positive, do not implement it if its NSB is negative'. Cost benefit analysis based on the NSB overcomes some of the disadvantages of cost effectiveness analysis based on the cost/effectiveness ratio. However, Torrance argues that it places too much emphasis on labour market activity (via the term B2), and is therefore likely to underestimate the true production gain to society through effects on people such as housewives (and househusbands !), who do not earn wages, but who nevertheless do a valuable job for society in looking after children etc. The basic problem

with cost benefit analysis is that it is difficult to arrive at a verifiably reliable figure for the monetary value of health care benefits.

(3). Cost utility analysis - This is just another form of cost effectiveness analysis (i.e. it is based on a cost/effectiveness ratio), but this time *effectiveness* is measured in QALYs. This overcomes the units-of-measurement problem associated with conventional cost effectiveness analysis, because QALYs are supposed to provide a 'standard' measurement unit for the effectiveness of different health programmes. Usually, cost utility analysis is treated as just another kind of cost effectiveness analysis, and we will refer to it as 'cost effectiveness analysis' from now on.

In the rest of his paper, Torrance considers in some detail the use of utilities and QALYs in cost utility analysis. We will cover most of his points later in the chapter. First, we focus more closely on one of the ways of 'valuing health per se' mentioned above: willingness to pay.

3. Willingness to pay

3.1. Definition

From the viewpoint of welfare economic theory, there is only one 'right' way to measure the benefits an individual gets from a health care programme: benefit is defined as ***the individual's maximum willingness to pay (WTP) for the health care programme when supplied with information as complete as it can be.*** WTP, in turn, represents ***the maximum amount of other goods, measured in monetary terms, that an individual would be willing to sacrifice to obtain the benefits from the health care programme.*** Note, then, that for any particular health gain the WTP figure may vary across individuals in society to the extent that their preferences, tastes or utility functions vary. It is perfectly legitimate for different monetary values to be attached to identical changes in health by different people if their utility functions differ.

The total benefit from a health care programme is then often defined as ***the sum of the willingnesses to pay of all persons whose welfare is affected by the health care programme.*** Note that this should include so-called 'caring externalities'. For example, if my welfare is affected by a programme that affects my parents' health, my WTP should be included (along with theirs) in defining the benefits from a health care programme that applies to them. Altruistic concerns about one's fellow human beings can also generate a WTP by others; these should be added in.

One point deserves particular emphasis: in economic theory, the WTP definition of an individual's benefits given above is the only acceptable definition. All other ways of valuing an individual's benefits monetarily should be judged in terms of how closely they approximate this definition. In particular, the 'economic benefits' we defined in the last section (containing direct, indirect and intangible monetary benefits) are theoretically valid only to the extent that they approximate the WTP measure.

Why is this? Simply because WTP takes into account differences in people's preferences. The other measures do not (necessarily). One of the main 'advantages' of cost benefit analysis using WTP is that it inherently takes into account heterogeneity in people's preferences because the WTP figure is allowed to vary across individuals to the extent that they differ in their valuations of the health benefit from a program. In contrast, cost effectiveness analysis

using e.g. life years saved or QALYs in the denominator, does not reflect *differences* in people's preferences: it uses a single comparison figure for the cost/effectiveness ratio that does not reflect heterogeneity in tastes among individuals (unless the comparison figure is measured as the social marginal WTP!). Thus, when there are wide variations in preferences across society, many economists argue that cost benefit analysis is more appropriate (in theory) than cost effectiveness analysis. The question then becomes 'to what extent should we let resource allocation decisions be influenced by individual preferences?'. There is no simple answer to this, but the fact is that using individual preferences as a basis for judging welfare is at the heart of welfare economics; if you ignore individual preferences (e.g. in a cost effectiveness analysis), there seems little point in sticking to the other tenets of welfare economics (e.g. that costs should reflect the 'marginal social cost'). But then, some would argue, we are completely 'at sea'.

This leads on to another criticism that is often levelled at WTP. Some people argue that differences between WTP measures across individuals do not just reflect differences in their preferences, but also differences in their wealth. In particular, they argue, wealthy people will tend to give higher WTP measures than poor people, just because they happen to be wealthier. Therefore, health care programmes which mainly benefit the relatively wealthy will tend to be accepted using cost benefit analysis (because the aggregate WTP will be high), whereas programmes which mainly benefit poor people will tend to be rejected (because the aggregate WTP will be relatively low, assuming that the rich do not have significant 'caring externalities' for the poor !). They argue that this is 'wrong', and that simply adding WTP across individuals in society is therefore an inappropriate basis for allocating resources. Do we agree with this ? It is easy to show that the argument is not valid when we can implement projects for the rich without affecting the poor in any way. However, the argument does have some merit in cases where we can only implement a project for the rich if we do not implement a competing one for the poor.

To illustrate the case where the argument is not valid, suppose we have one possible health care project for the rich and one possible project for the poor, and that we could implement both of them if the benefits exceeded the costs in each case. However, we find that the project for the rich passes the cost benefit test, whereas the project for the poor doesn't, so we only implement the one for the rich. Is this a bad thing ? If you think about it, the answer is 'no' even if WTP *is* positively correlated with wealth. As long as poor people don't end up paying for the rich people's project, the project will make some people (the rich) better off without making anyone else worse off. So why not do it ? Now consider the project for the poor that ended up being rejected because the WTP was not high enough. There is no good reason why it should be implemented ! If we really want to provide benefit for the poor, a more efficient approach would be to use the money that would have been spent on their project to make a direct money transfer to them. Making this money transfer to the poor would benefit them more than the project itself, because they value the project less than the money it costs. In this case, providing the project for the poor, as opposed to making a direct money transfer, is worse both for the poor people (because they would rather have the money) and for the rest of the community (because the community would be making a net welfare loss on the project). If the community has mechanisms in place for making money income transfers to the poor as it sees fit, there is no justification for implementing the project for the poor; if you really want to benefit them, just give them the money ! Thus, cost benefit analysis using WTP has led to the 'right' decision in this case: it has served its purpose by telling us that there are more efficient ways of making the poor better off than by implementing their project. (The issue of

whether or not the community actually decides to make the money transfer to the poor after rejecting their project is a political one that has nothing to do with the WTP method itself).

To illustrate the case where the 'unfairness' argument does have some merit, suppose we have one possible health care project for the rich and one possible project for the poor, but this time we cannot afford to implement both. We have to choose one or the other. We find that the rich people's project passes the cost benefit test (because the aggregate WTP is quite high), but the poor people's project fails the cost benefit test (because the aggregate WTP is quite low), so we choose to implement the project for the rich and do not implement the project for the poor. Is *this* a bad thing? This case is different from the one above, because undertaking the programme that benefits higher-income persons precludes a programme, possibly one saving more lives, that benefits lower-income persons. This time, we cannot make a money transfer to the poor equal to the cost of their rejected project (by assumption, we cannot afford it). Thus, poor people will simply have lost out, and only because they are poor (and so had relatively low WTP measures).

We will be able to study the issue of the relationship between wealth levels and WTP in more detail in the section on the theoretical foundations of cost benefit analysis below.

3.2. Measuring WTP

The most direct method to determine WTP is a questionnaire approach that asks people how willing they are to pay for a particular health benefit. In essence, people are asked to make a trade-off between money and the health benefit in question. This approach is subject to a simple but fundamental criticism: hypothetical trade-offs are difficult (or even impossible) to 'validate' i.e. it is difficult to verify that what people say they would be willing to pay for a health benefit is actually the most that they would pay if they were really confronted with having to pay for it. Very simply, no-one can prove or disprove that WTP measures are valid (a measure is said to be 'valid' if it *really* measures what it is supposed to). This problem will probably always make people uneasy about basing important resource allocation decisions on *direct* WTP measures.

However, all is not lost! There are indirect methods of estimating WTP that are 'self-validating'. These are called hedonic approaches, and are designed to enable the valuation of products that are not traded in a market (so their market price cannot be directly observed). In general terms, hedonic approaches involve specifying a set of key characteristics that a particular product has. The implicit value of these characteristics is then inferred from market data on products that *are* sold in markets, and that have similar characteristics. For example, if a product that either cannot be or has not yet been marketed reduces mortality (this is its 'key characteristic'), one might infer the money value of life by observing different wages paid in jobs that differ in terms of their expected mortality. The market traded product in this case, that has the same key characteristic of reducing mortality, is 'labour time' in a job that has reduced mortality rates compared to others. This should have a lower wage rate associated with it than labour time in a job with a higher mortality rate, because workers would have to be compensated in the latter job for facing a higher risk of death. The degree to which they are compensated by the market may be taken to be a measure of the social monetary value attached to life. This could then be used to attach a 'hedonic' value to the non-traded product that reduces mortality. In principle, then, the hedonic approach looks at every product as a bundle of characteristics, and the 'hedonic' value of those characteristics is what must be estimated. The hedonic approach has not yet been applied extensively in connection with

medical services, but research is currently under way. It has the strong advantage that it is 'self-validating': values are based on actual market behaviour, not on artificial answers to hypothetical questions. As with any approach to estimating WTP, the hedonic approach also has its difficulties, but further research might help to overcome them.

4. Quality-Adjusted Life Years

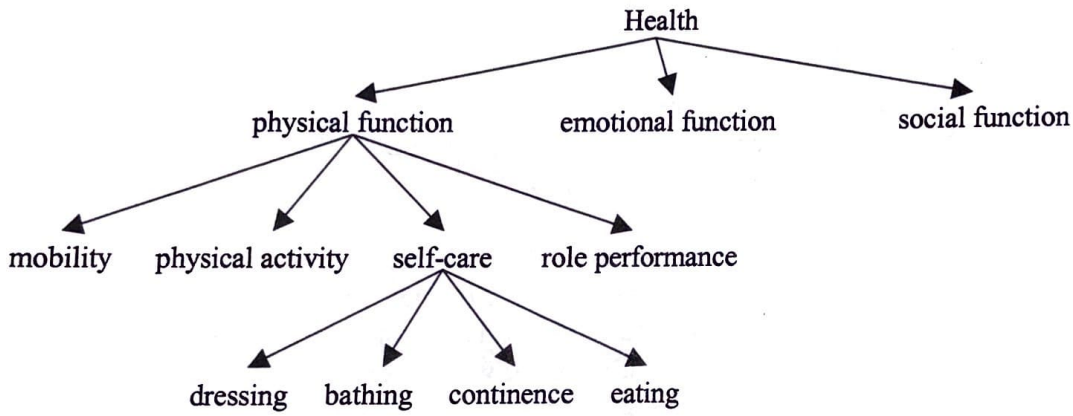
4.1. Health state definitions

In our brief look at QALYs in Section 2 above, we saw that these are based on measurements of the utilities (on a scale from 0 to 1) associated with different possible health states. The utility values reflect the quality of the health states, and allow morbidity and mortality improvements to be combined into a single weighted measure (i.e. QALYs gained). For example, if a programme improves the health of individual A from a 0.5 utility to a 0.75 utility for one year, and extends the life of individual B for one year in a 0.5 utility state, the total QALYs gained for that year would be 0.25 for individual A plus 0.5 for individual B, which equals 0.75. A's contribution to this total came from a morbidity improvement, whereas B's came from an improvement in terms of mortality. The assignment of utilities to these different types of health improvement has allowed us to aggregate their 'value', and express it in terms of a single unit (QALYs). In this section we will consider in detail some of the main approaches used by health economists today to estimate these utilities. First, we need to be clear about what is meant by a 'health state'.

Each unique possible health outcome for the health care programme(s) under evaluation must be defined as a 'health state' for the purposes of utility measurement. Depending on the study, there may be only a few health states or there may be many hundreds. The prevailing view is that each of these health states should be described in 'functional' as opposed to 'clinical' terms. That is, the description of a particular health state (e.g. mild diabetes) should focus on how well or poorly a person in this health state is able to function (e.g. go to work, relax, socialise, eat, dress, bathe, etc.), rather than on the clinical diagnosis or laboratory test results. A comprehensive description of a health state should include a statement on the level of physical functioning (e.g. 'being able to get around the house without help, but having some limitations in physical ability to lift, walk, run, jump or bend'), the level of emotional functioning (e.g. 'being anxious or depressed some or a good bit of the time'), and the level of social functioning (e.g. 'having an average number of friends and contacts with others'). These functional aspects are referred to as 'dimensions'. Studies have shown that the utility associated with a health state is also affected by its duration. Thus, it is important to state the duration of the health state either as part of the description itself, or as part of the measurement process. It is also important to make sure that the utility of a health state is unconfounded by utilities of other states that may follow. Thus, it is important that the 'prognosis' for the health state should not be left unspecified or vague for each subject to interpret differently. Usually this is handled by making no mention of prognosis in the health state description itself, and then specifying a clear and certain prognosis as part of the measurement process (e.g. '3 years in this health state, followed by a return to normal health').

There are three ways to 'describe' a health state to a subject. (1). One can use the subject's own health as the health state to be measured. Thus, to measure the utility associated with mild diabetes, one might simply ask a mild diabetic! This obviates the need to prepare a description of the health state for use by the subject. However, descriptions would still be

required for the anchor states of the utility scale - typically, healthy and dead. Moreover, a description of the patient's health state may still be required so that others can interpret the results. (2). One can use a 'holistic' description of the state. This involves asking subjects who are not in the health state to assess it based on a description. It should be made clear either in the description or in the measurement procedure itself whether this health state is to be considered by the subject as applying to him/herself or to someone else. In addition, the age of onset of the particular health state should be specified to the subject. The descriptions of the health states may vary enormously in level of detail; it is important, however, not to overload the cognitive abilities of the subject with very long and detailed descriptions. In these cases, subjects simply latch on to a few key phrases and ignore the rest. (3). One can use a health state classification system. Instead of describing each specific health state of interest individually, the 'classification system' approach involves defining health status in terms of a number of attributes, possibly hierarchically nested. We can build up a 'tree' of these attributes as follows:



At the highest level of the tree, we break down 'health' into what we consider to be the three main dimensions: physical function, emotional function, social function. At the next stage, each of these in turn is broken down into a set of main attributes. For example, physical function in the diagram above is broken down into moblity, physical activity, self-care and role performance (role performance essentially means 'ability to do one's job'). In turn, each of these is broken down into a further set of attributes. For example, self-care above is broken down into dressing, bathing, continence and eating. Finally, at the lowest level, these might be broken down further. The lowest level attributes in the hierarchical structure are divided into levels that represent step-wise decreases in function on that particular attribute. For example, the attribute 'dressing' might be divided into three levels: (i) able to dress onself normally (ii) able to dress oneself with difficulty or with the use of mechanical aids (iii) requires assistance of another person in dressing.

Within each attribute the function levels must be mutually exclusive and exhaustive, so that at any point in time each individual can be classified on each attribute into one and only one function level. Each different combination of levels, one from each attribute, represents a unique health state. Thus, a classification system can contain an enormous number of health states. To measure the utility of each of these health states one at a time would be an infeasibly large task. However, if the attributes satisfy certain independence properties, something called 'multi-attribute utility theory' can be used to dramatically reduce the amount of work.

4.2. Utility scales

Now that we have some idea of what 'health states' are, we next briefly consider what the measurement scales are on which we can attach utilities to them. Measurement scales are classified, from weakest to strongest, as dichotomous (e.g. gender), categorical (e.g. religion, ethnicity), ordinal (e.g. finishing 'place' in a competition), interval (e.g. temperature on ^{0}F or ^{0}C) and ratio (e.g. weight, length), with the last two also being called 'cardinal'. Dichotomous, categorical and ordinal scales are self-explanatory.

An interval scale is one in which both the zero point and the size of the measurement unit are arbitrary. Such a scale has the property that interval lengths (i.e. differences between scale values) can be compared in a meaningful way, but ratios of scale values cannot. For example, if A is 5^0C (41^0F), B is 10^0C (50^0F), and C is 20^0C (68^0F), it is correct to state that the difference between A and C is one and a half times the difference between B and C, and three times as great as the difference between A and B, but it is incorrect that the actual value B is twice as great as A, etc.

A ratio scale is one in which the zero point is clearly defined and only the unit of measurement is arbitrary (e.g. length in inches, feet or yards). Ratio scales have all the properties of interval scales with the additional property that ratios can be compared.

Utilities can be measured on an ordinal scale or a cardinal scale. Ordinal utilities are simply a rank ordering of the health states or outcomes in order of their preference, with ties allowed. Ordinal utilities are the easiest to obtain and the least demanding in terms of their underlying assumptions, and are therefore the preferred measure when they are adequate. Unfortunately, they are rarely adequate. Cardinal utilities are a set of numbers assigned to the health states or outcomes such that the number represents the strength of the preference on a cardinal scale. The cardinal scale may be interval or ratio depending on the measurement method used. However, interval scales are adequate for use in cost utility analysis and so there has been little concern with the development of techniques to produce ratio scales of utility. All of the measurement methods described below produce interval scales of utility. An interval scale has the property that any two points can be assigned arbitrary values in order to define the scale. In measuring preferences for health states, it has become customary to arbirarily assign the values of 0 and 1.0 to the reference states 'dead' and 'healthy' respectively.

Utility measurement techniques determine a utility for each subject. For economic appraisal these individual utilities must be aggregated into a collective social utility. The question of aggregation and its validity has been addressed by many authors. For example, in the case of ordinal utilities, Kenneth Arrow has shown that there is no aggregation technique that satisfies some 'reasonable' criteria. On the other hand, for cardinal utilities, different sets of 'reasonable' assumptions lead to different results: some authors argue that aggregation is valid, while others argue that it isn't. The reality is, of course, that comparisons of individual preferences are common in practice - indeed, in order to make social decisions and in the very process of making those decisions, individual preferences *must* be and *are* compared. The question, then, is not whether to make such comparisons but how to make them.

In cost utility analysis the aggregation across subjects is achieved by measuring all individual utilities on the common 0-1 dead-healthy scale, and taking the arithmetic mean. The central basis for this method is that the difference in utility between being dead and being healthy is

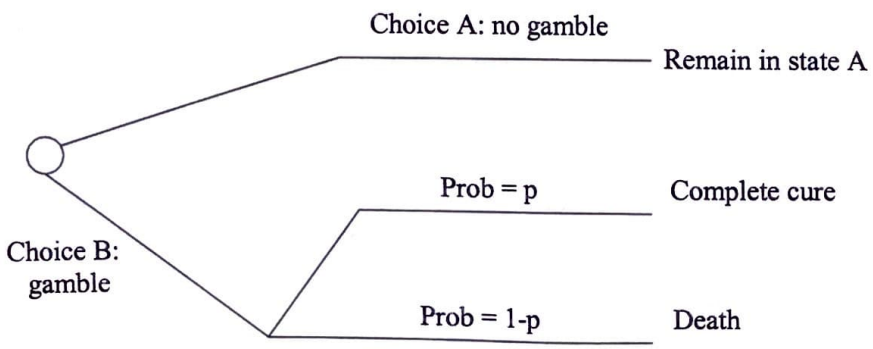
set equal across people. In this way the method is 'egalitarian' with respect to health, in the sense that each individual's health is counted equally.

4.3. Utility measurement

Fundamentally, utility measurement simply consists of presenting a subject with descriptions of several health states and eliciting directly or indirectly the subjects' relative preferences for the states. As discussed above, each description of a health state should be functionally orientated and comprehensive. Further, the description of the state or the measurement process should specify the age of onset of the state, the duration of the state, the exact prognosis for what follows the state, and whether or not the state applies to the subject himself or to someone else. In addition, the utility measurement should be unconfounded by the subject's economic wellbeing. Thus it is important to assure the subject that all treatment and all outcomes will be costless to him and to his family - that is, the subject is to assume full-coverage health insurance and salary continuation insurance.

Ordinal preferences are simple to measure. One merely asks the subject to rank order the health states in order of their preference with ties allowed. Normally, the health states should be of the same duration, same age of onset, and same prognosis - otherwise the results are difficult to interpret. Although ordinal preferences are simple to measure, they pose problems in aggregation across subjects as shown by Kenneth Arrow's 'Impossibility Theorem'. They are also unsuitable for cost utility analysis. Thus, economic evaluations of health care programmes are based on cardinal utilities.

Cardinal preferences can be measured by a number of different techniques. The four major ones are rating scale, standard gamble, time trade-off and person trade-off. Rating scales provide simple techniques for assigning numerical values to objects. A typical rating scale consists of a line on a page with clearly defined endpoints. The most preferred health state is placed at one end of the line and the least preferred at the other end. The remaining health states are placed on the line between these two, in order of their preference, and such that the intervals between the placements correspond to the differences in preference as perceived by the subject. In the usual case where death is judged to be the worst state and placed at the 0 end of the rating scale, the preference value for each of the other states is simply the scale value associated with its placement. The method becomes slightly more complicated if death is not judged to be the worst state but is placed at some intermediate point on the scale. The basic idea is the same, though. The standard gamble is the classical method of measuring cardinal preferences. It is based directly on the fundamental axioms of von Neumann-Morgenstern utility theory. As illustrated in the diagram below, the standard gamble offers a choice between two alternatives: living in health state A with certainty or taking a gamble on treatment for which the outcome is uncertain:



The respondent is told that the treatment (choice B) will lead to perfect health with probability p or immediate death with probability 1-p. The health state described in A is intermediate between wellness and death. The probability p is varied until the subject is indifferent between choices A and B. This value of p is then taken to be the utility of state A. In the time trade-off method, the subject is offered a choice between living in perfect health for a defined period of time (e.g. 1 year) or a greater amount of time in an alternative state that is less desirable. By varying the time in the sub-optimal state, an indifference point may be found from which a utility can be assigned to the sub-optimal state. For example, a subject may rate being in a weelchair for two years as equivalent to perfect wellness for 1 year. Then the utility of being in a weelchair might be rated as 1/2. Finally, the person trade-off method allows comparisons of the numbers of people helped in different states. For example, respondents might be asked how many persons in state B must be helped to provide a benefit equivalent to helping one person in state A. If the answer is 100, then the utility of state B is taken to be 100 times the utility of state A.

5. Theoretical foundations of cost benefit analysis

Cost benefit analysis is based on the welfare economic model we set out in Chapter 3. There we used a social welfare function $W = W(u^A, u^B)$ to illustrate the discussion. We will use this same function to explain the theoretical basis of cost benefit analysis here. Totally differentiating the social welfare function W gives

$$dW = (MW^A{}_u)\cdot du^A + (MW^B{}_u)\cdot du^B$$

where (using the same notation as in Chapter 3) $MW^A{}_u$ denotes the increase in W produced by a one unit increase in u^A (i.e. the marginal social benefit of a one unit increase in u^A), and $MW^B{}_u$ denotes the increase in W produced by a one unit increase in u^B (i.e. the marginal social benefit of a one unit increase in u^B). We said in Chapter 3 that assessing the desirability of particular policies using cost benefit analysis involves comparing different economic states in terms of the social welfare gains and costs associated with those states. If we are already at the social optimum, then any policy which increases welfare by benefitting some people must reduce welfare by the same amount by hurting other people (otherwise we could not have been at the optimum), so in this case $dW = 0$. But in the real world, we often have to decide whether or not to implement particular policies in situations where we are not at the social optimum. In this case, we would get either $dW > 0$ or $dW < 0$. A particular policy might benefit some people (for whom $du > 0$) and hurt others (for whom $du < 0$), but as long as $dW > 0$, society as a whole would benefit, and we might decide on this basis to go ahead and implement the policy. Conversely, if $dW < 0$, we might abandon the policy, even if some individuals in society would benefit from it.

However, the assessment of whether or not $dW > 0$ is not operational as we have described it above, because everything is in terms of utility. For practical policy purposes, we saw earlier that we need to measure the changes in individual welfare not in units of utility but in units of money. The general approach to doing this is as follows. Suppose we are doing a cost benefit analysis of a project which will change individual A's utility by du^A, and individual B's utility by du^B. Let λ^A denote individual A's marginal utility of wealth, and let λ^B denote individual B's marginal utility of wealth. Thus, λ^A measures the increase in individual A's utility as a result of a one-unit increase in his/her wealth, and λ^B measures the increase in individual B's

utility as a result of a one-unit increase in his/her wealth. Now multiply and divide each term on the right-hand side of the expression for dW above by the marginal utility of wealth to the individual concerned, to get

$$dW = (MW^A_u)\cdot \lambda^A \cdot (du^A/\lambda^A) + (MW^B_u)\cdot \lambda^B \cdot (du^B/\lambda^B)$$

For each individual, we now have the following two terms:

1. du^A/λ^A measures A's change in utility as a result of the project, divided by the change in utility that would have been produced by one more unit of wealth. It thus indicates how many units of wealth (i.e. how much money) would have produced the same change in utility as the one A experienced as a result of the project. In other words, it indicates how much money A would be willing to pay for the change in his/her utility brought about by the project. Similarly, du^B/λ^B measures B's willingness to pay for his/her change in utility as a result of the project. These are the WTP figures we were talking about earlier. We write $WTP^A = du^A/\lambda^A$ and $WTP^B = du^B/\lambda^B$.
2. $(MW^A_u)\cdot \lambda^A$ measures the social value of an extra unit of wealth accruing to A, and similarly $(MW^B_u)\cdot \lambda^B$ measures the social value of an extra unit of wealth accruing to B. We can think of these as weights attached by society to each individual's WTP, and write $Weight^A = (MW^A_u)\cdot \lambda^A$ and $Weight^B = (MW^B_u)\cdot \lambda^B$.

Using this notation, we can rewrite the expression for dW above as

$$dW = (Weight^A)\cdot(WTP^A) + (Weight^B)\cdot(WTP^B)$$

For the project in question, we then simply tabulate the WTP for each of the individuals affected, and then assign our own 'ethical' weights to each individual's WTP (or leave this to the politicians). Suppose the values are as follows:

$Weight^A = 1$ $\qquad$ $WTP^A = 200$

$Weight^B = 3$ $\qquad$ $WTP^B = -100$

Then the project would be undesirable, since $dW = 1(200) + 3(-100) = -100 < 0$. But suppose we proposed implementing the project, and then making A pay 100 to B. For this new proposal of the project *plus* compensation, the values would be

$Weight^A = 1$ $\qquad$ $WTP^A = 100$

$Weight^B = 3$ $\qquad$ $WTP^B = 0$

and this does pass the test since $dW = 1(100) + 3(0) = 100 > 0$. In other words, the project plus compensation yields a Pareto improvement.

In Section 3, we discussed the 'unfairness' of basing resource allocation decisions on cost benefit analyses using simple aggregations of WTP across individuals. We said that this can lead to systematic bias against poor people if WTP is positively correlated with wealth. How can this positive correlation happen ? We were implicitly talking about a model like the one above but with $Weight^A = 1$ and $Weight^B = 1$, giving

$$dW = (WTP^A) + (WTP^B) = (du^A/\lambda^A) + (du^B/\lambda^B)$$

The problem of 'unfairness' arises in this case because if we make the standard assumption of diminishing marginal utility, then λ^A (A's marginal utility of wealth) will be small if A is wealthy and large if A is poor. Therefore WTP^A will be large if A is wealthy, but small if A is poor (i.e. WTP will be positively correlated with wealth). The same applies to B. This suggests that it may be more appropriate to use a *weighted* sum of individuals' WTP measures in cost benefit analyses (with heavier weights on poor people's WTP measures), rather than a simple unweighted sum.

6. The basic cost effectiveness analysis model

Cost effectiveness analysis is based on a simple optimisation problem. A decision maker faces a 'menu' of health care programmes (i = 1, 2, . . ., N) from which to choose. The decision maker has a cost budget C. The decision maker's objective is to choose a combination of programmes from the menu so as to maximise total benefits, or effectiveness E (e.g. QALYs). Each program uses part of the budget, that is, has cost C_i. Each programme contributes to total benefits, that is, has effectiveness E_i. Any combination of programmes on the menu is feasible, provided that it satisfies the budget constraint. The cost and effectiveness of any programme are independent of which other programmes are adopted. Programmes are not repeatable, but all programmes are divisible with constant returns to scale (CRS) i.e. it costs half as much to produce half the QALYs. The ratio C_i/E_i is the cost effectiveness (CE) ratio of programme i. It can easily be shown that the optimal budget allocation in this simple case is to rank the porgrammes in order of rising CE ratio, and choose from the lowest to the highest until the budget is exhausted.

Example: Suppose a Health Authority has a budget of £10 million. The table below lists the available health care programmes with associated costs, benefits and CE ratios. The programmes have been arranged in order of rising CE ratio.

Program	Benefits (QALYs)	Cost (£)	C/E ratio
A	500	1,000,000	2,000
B	500	2,000,000	4,000
C	200	1,200,000	6,000
D	250	2,000,000	8,000
E	100	1,200,000	12,000
F	50	800,000	16,000
G	100	1,800,000	18,000
H	100	2,200,000	22,000
I	150	4,500,000	30,000
J	100	5,000,000	50,000

The optimal allocation of £10 million is A, B, C, D, E, F, G, up to a cutoff of £18,000 per QALY. If the budget is only £3 million, the optimal allocation is A, B up to a cutoff of £4,000 per QALY.

The above procedure works as long as the projects are independent of each other. If two or more programmes are not independent of each other, but are competing or mutually exclusive (e.g. they represent competing treatments for a given condition, only one of which will be

Program	Benefits (QALYs)	Cost (£)	Incremental C/E ratio
K_0	0	0	0
K_1	10	50,000	5,000
K_2	15	150,000	20,000

used), we must use the concept of an Incremental CE Ratio (ICER) to find the optimal allocation of the budget. The ICER is the ratio of the incremental benefit to the incremental cost of a programme relative to the next most effective one that is also cheaper. We 'throw out' a programme if it is both less effective and more expensive than any other. Such programmes are said to be 'dominated'. To illustrate the use of ICERs, suppose that in addition to programmes A-J in the first table above, the Health Authority can also fund one of two mutually exclusive options or neither. The benefits, costs and ICERs of these are shown in the second table above. For K_2, the ICER is calculated as [150,000-50,000]/[15-10] = 20,000. Given the £10 million budget, which of K_0, K_1 or K_2 is best ?

The correct procedure is to use the £18,000 per QALY cutoff from the first table. We partially or wholly implement any option from the second table whose ICER is less than £18,000 (at the expense of G in the first table). Thus, we should fund K_1 (and reduce spending on G in the first table to £1,750,000), but we should not fund K_2 since G is better.

Notice that without the concept of ICERs, it would have been tempting to choose K_2 from the second table, since programme K_2 has a CE ratio of £10,000, and £10,000 is less than £18,000. However this would have been wrong, since it ignores the availability of option K_1, which can give two-thirds of the benefit of K_2 at only one-third of the cost.

There is a pitfall in incremental cost effectiveness analysis, as illustrated by the following example. Suppose that an advocate for K_2, noticing the result of the analysis described above, points out that another option, call it $K_{1.5}$, has been left out. Suppose $K_{1.5}$ costs £120,000 and saves 12 QALYs. The advocate points out that the ICER for K_2 relative to $K_{1.5}$ is [150,000-120,000]/[12-10] = £15,000 per QALY. So, he says, K_2 is really cost effective after all, because its ICER is less than £18,000 per QALY.

The argument is fallacious because $K_{1.5}$ is an inappropriate basis for comparison. This can be seen by calculating its own ICER relative to the next less costly option, K_1, which is £35,000 per QALY. Since $K_{1.5}$ has a higher ICER than K_2 (i.e. it is a more 'expensive' way to produce QALYs than K_2), but at the same time is less effective than K_2, it will never be 'cost effective' and should be eliminated from consideration.

This pitfall implies that the choice of a basis for comparison in calculating an ICER can be crucial. Any option can be made to look cost effective if it is compared to a sufficiently cost ineffective alternative! The optimal decision rule is to consider only those options whose ICERs are lower than every competing option which has greater effectiveness.