Techniques of Health Status Measurement using a Health Index

Gillian Holland

Health Economics Research
Department of Economics, Brunel University

Office of Health Economics
12 Whitehall
London SW1A 2DY

April 1985

Techniques of Health Status Measurement using a Health Index

Background

It is widely accepted that resources for the provision of health care are scarce, that is, there are not and never will be enough resources to satisfy either subjective "demands" for health care or externally measured "needs". The use of resources in one health care activity means the opportunity to use those same resources in a competing activity is automatically foregone (the economist´s notion of opportunity cost) (Drummond 1983). It is logical therefore that, for the purpose of resource allocation decisions, health care options should be compared in terms of their relative costs and benefits.

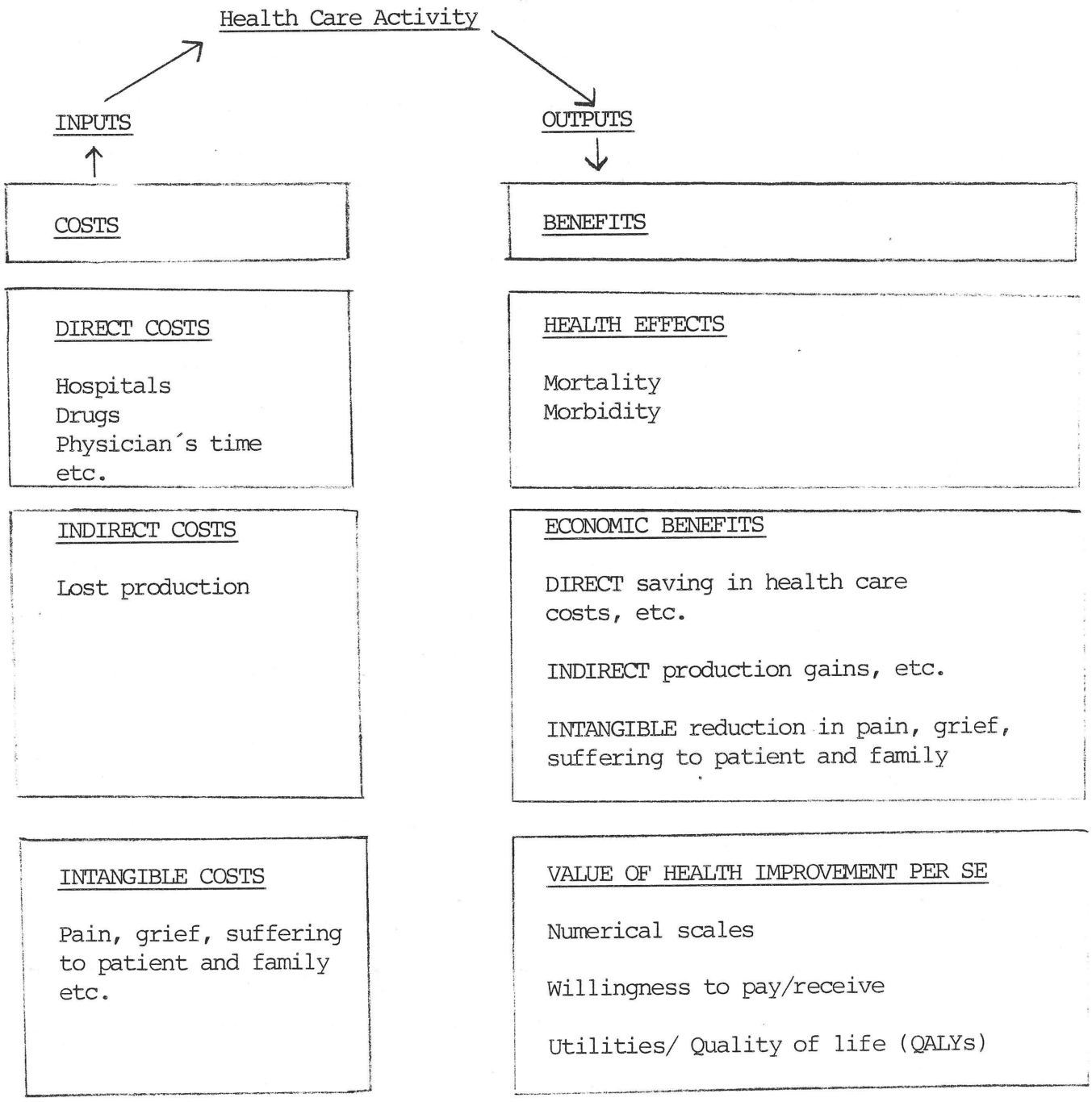
Fig 1 illustrates the components of costs and benefits encountered in a typical health care programme. Economic analysis and its techniques are simply a systematic way of measuring and displaying these in order to enable the decision maker to make a more informed choice.

Taking first the most simple form of economic appraisal, "Cost-analysis" involves the investigation of merely the direct or the direct and the indirect costs. Indeed, in that the benefits or consequences of health care interventions are ignored it may not be considered a form of appraisal at all. However, cost-analysis may be the appropriate method to use for comparing the costs of a new therapy with existing treatment options.

"Cost-benefit analysis" (CBA), which in the past has been applied to many health contexts ranging from mobile coronary care units (Acton 1973) to vaccination against infectious diseases (Adler et al 1983) entails the systematic comparison of as many costs and benefits as possible, with a view

Figure 1

COMPONENTS OF ECONOMIC APPRAISAL



to determining which health care programme maximises the difference between the costs and benefits. In CBA both the costs and benefits are measured in commensurate terms, which usually, though not necessarily, means money terms. This gives rise to one major disadvantage of CBA, namely, that difficulty in measuring and evaluating intangible costs and benefits often results in these components being ignored and thus, to some extent, limits the usefulness of the analysis.

For this reason analysts frequently turn to "Cost-effectiveness analysis" (CEA), which serves to place priorities on alternative expenditures without requiring the monetary value of health outcomes to be assessed. In CEA resource costs are inevitably measured in money terms but the benefits may be expressed in physical units in a number of ways, such as life years gained and number of days free from disease. (Thus the practical difference between CEA and CBA is that the former requires commensuration of outcome measures within the major categories of resource cost and health effectiveness (benefits) whilst the latter requires that costs and benefits all be valued in the same units) (Weinstein 1981). CEA is a useful way of comparing alternative procedures whose effects are measured in the same units, and hence to assess the efficiency with which limited resources are being allocated to achieve the desired benefits. However, it suffers from the disadvantage that it can not be used for comparing disparate alternatives, such as comparing kidney dialysis for renal patients with home care for the frail elderly.

"Cost-utility analysis", (CUA) with which this paper is chiefly concerned, is a form of CEA in which the most generally used measure of effect is a "quality adjusted life year gained". The extension to life, following medical intervention, is adjusted by a series of "utilities" which reflect the relative value of one health state to another. (This outcome measure has more recently become known as the QALY.) In that the QALY incorporates

changes in survival and morbidity into a single measure reflecting trade-offs between them, it is a particularly useful outcome measure when the benefits of health activities can more appropriately be assessed in terms of their impact on the quality, not quantity, of life (Drummond 1984).Similarly, it is useful when the health activity being considered affects both mortality and morbidity and a common unit that combines both effects is desirable. This approach, CUA, is the appropriate technique for comparing programmes with a wide range of outcomes and for comparing one programme to others that have already been evaluated using CUA.

Economic analysis has been widely applied to the health field and there are numerous examples of each form of economic appraisal in the literature. Drummond (1981) has classified over 50 studies which illustrate these differences in approach.

Quantitative valuation of the health improvement per se - health status measurement

In the past much of the emphasis of economic appraisal of health care programmes has been on valuing costs, changes in health services and community resources and economic benefits. That is, because of difficulties in quantification and valuation, changes in health state per se have tended to be omitted. This suggests, quite wrongly, that economic appraisal is synonymous with the assessment of merely the financial aspects of health treatments.

More recently there has been a growing tendency among health care professionals, researchers and economists to recognise the need to develop ways to measure and quantify the change in health status itself resulting from a given health care activity. In pursuing this aim three main approaches have been developed: the first involves the use of ad hoc numerical scales, the second is the willingness to pay/receive approach and the third is through the use of utilities and QALYs.

Focusing on the first of these, the use of ad hoc numerical scales involves assessing the individual on a number of aspects of his/her health, assigning numerical scores to each assessment and adding up the scores.

Grogono and Woodgate (1971) used this approach for their "Index for Measuring Health". They identified 10 dimensions of human functioning which reflected the aspects of life upon which medicine was expected to have an impact. (See fig 2). The scoring system used was to allocate 1, ½ or 0 to each factor according to whether the patient was normal, impaired or incapacitated.

Figure 2

COMPONENTS OF THE GROGONO-WOODGATE INDEX

1. Work
2. Recreation
3. Physical Suffering
4. Mental Suffering
5. Communication
6. Sleep
7. Dependency on others
8. Feeding
9. Excretion
10. Sexual activity

The score at a particular point in time for each patient was taken as the sum of the scores across all ten dimensions and the total was then divided by 10 to yield a health index. The authors suggested their instrument could be used to evaluate the benefits derived from medical treatment for individuals, and to allocate resources in communities for treatment and research.

Although this was an ambitious proposal, this like other such indexes is essentially arbitrary and has several serious methodological problems.* Other examples of this approach to measuring health status include the Harris Index (1971), the Karnofsky Index (1949) and Spitzers QL Index (1981). For application in economic appraisal these indexes could be used as a measure of effect in CEA.

Drawing on the work of Schelling (1968), Mishan (1971) developed the willingness to pay (WTP) approach, which is based firmly in modern welfare economics. That is not to say that the principle is uncontroversial, but it is a clearly well understood philosophical rational. It rests on the idea that individuals´ valuations are reflected in what they would be willing to pay to receive certain benefits or avoid certain costs. (Pay is used here in the sense of what individuals are willing to forego or sacrifice and not just in the monetary sense.)

---

* Culyer (1978) points out that there was no apparent awareness in the study that certain value judgements were being made, and once exposed, these would not be the value judgements that the authors would be likely to make. These are a) the judgement that the 'rate of substitution' of one dimension for another is constant i.e. a half unit increase in one dimension can always be exactly offset by a given decrease in any other dimension. b) the judgement that an increase in one dimension is always exactly offset by an identical decrease in any other dimension and c) the judgement that a move from one index contour to another gives equal increments of health status.

The approach can take several forms. One alternative involves the use of actual market decisions as a basis for making inferences about WTP. For example, in a widely cited study, Thaler and Rosen (1975) looked at wage preminiums paid to persons in hazardous occupations in return for accepting identifiable risks. A second alternative entails the use of survey based inferences of WTP. Acton (1973) used a direct survey procedure to determine how much people would be willing to pay for emergency coronary care services which reduced the probability that a heart attack victim would die as a direct consequence of the heart attack. The use of decision analysis, which provides a set of procedures for explicitly analysing complex decision problems and chooses according to the expected utility principle, constitutes a third alternative.

Rosser and Watts (1972) measured what they described as the "Willingness to Receive", (WTR) as determined by the amount of a court award for monetary compensation for injury. They analysed about 500 awards made by the high courts of Great Britain empirically to determine the relative value of health states based on monetary criteria.

Both WTP and WTR provide a monetary value which can be used in CBA. However, in addition to the objections of principle, many practical problems encountered in this approach have lead to its infrequent use.

The third approach, to measuring health status, pioneered by Torrance, is through the use of utilities and QALYs. It depends on the use of a cardinal scale in which the differences between the individual values along the scale can be compared in a meaningful way. An everyday example of such a scale is the use of degrees centigrade for temperature measurement. Thus Torrance (1984) describes utilities as "cardinal values that are assigned to each health state on a scale that is established by assigning a value of 1.0 to

being healthy and 0.0 to being dead. (This shall now be referred to as the dead-healthy scale.) The utility values reflect the quality of health states and allow morbidity and mortality improvements to be combined into a single weighted measure, QALYs gained." To use his example, if a health care programme improves the health of individual A from a 0.50 utility to a 0.75 utility for one year and extends the life of individual B for one year in a 0.50 utility state, the total QALYs gained for that year would be 0.25 for individual A plus 0.50 for individual B, giving a total of 0.75.

The determination of numerical weights or utility values, as referred to above is the focus of attention in this paper, the contents of which are based on the aforementioned paper presented by Torrance. In tackling this problem the analyst has a choice of three alternative methods: judgement, the use of suitable existing utility values published in the literature, or the use of measurement techniques to measure the values him/herself. Once established these weights or utility values can be used in practice to measure the quality of life either at a point in time or over a period of years for a group of actual patients.

Alternative 1: Judgement

The use of judgement to estimate utility values is undoubtedly the simplest method and has two advantages in that it is relatively quick and cheap. The analyst himself may make a simple estimation or a more formal measurement may be made based on the knowledge of a sample of experts who will allocate different utility values to different states of health.

The unavoidable subjectivity of the judgemental approach, however, makes it necessary to carry out sensitivity analysis in those studies in which this method is adopted. If the analysis shows that the conclusions of the study are relatively robust i.e. relatively insensitive to wide changes in the subjectively assessed utility values, then this approach may be considered adequate. However, if the conclusions are sensitive to changes in the utility values, it would be necessary to obtain more credible values by using an alternative technique.

Alternative 2: Use of utility values taken from the Literature

There are a growing number of studies in which utilities for certain health states have been measured and published. By way of example, for end stage renal failure patients Churchill (1984) published utilities for haemodialysis, continuous ambulatory peritoneal dialysis (CAPD) and transplantation. On a utility scale ranging from 0.00 for death to 1.00 for perfect health, the mean utility for chronic haemodialysis for the 42 patients receiving the treatment at the time of interview was 0.57. Similarly, for the 17 CAPD patients it was also 0.57, and for the 14 transplanted patients the mean had a value of 0.80. Pliskin and colleagues (1980) reported utilities for 2 levels of angina pain - mild and severe. Taking a pain-free year as having a utility value of 1.0, the estimated

value of a year with severe chest pain ranged from 0.42 to 1.00 (with a mean of 0.69 and a standard deviation among estimated values of 0.22) and the estimated value of a year with mild chest pain ranged from 0.74 to 1.00 (with a mean of 0.88 and a standard deviation among respondents of 0.10).

These and other existing values, taken from the literature* may be employed by other researchers. Caution is required however to ensure the health states measured in the original study match those of the new study. In addition, the subjects used in the measurement process in the original study must be appropriate for the new study. And finally, the original study must have used valid methods of measurement.

Alternative 3 - Measurement of the utility values.

A third and more accurate way to acquire utility values is for the analyst to obtain the values him/herself using formalised measurement techniques. Four stages can be identified in such a measurement process and each is considered here in turn:

(i) Identification of health states for which utilities are required

(ii) Preparation of health state descriptions

(iii) Selection of subjects

(iv) Use of utility measurement instrument

Stage (i) Identification of health states

In the first stage each unique possible health outcome which may be

---

* (see for example utilites for loss of speech due to Laryngectomy reported by McNeil et al (1981), utilities for Cancer related states reported by Llewellyn-Thomas et al (1982) Sutherland et al (1983).

encountered in the study should be identified. Inevitably the number of different health states which may be established in this way depends on the nature of the study itself. In a study of neonatal intensive care for very-low-birth-weight infants (Boyle et al 1983) there were 960 distinct possible health states. (A classification of health states was developed to measure the health of survivors according to their physical function (six possible levels), role function (five levels), social and emotional function (four levels), and health problems (eight levels). Thus, there were 6x5x4x8= 960 health states). Whereas by contrast, a demonstration application of a utility maximisation model (Torrance, Sackett, Thomas 1973) involved the measurement of utilities for just 5 health states (home confinement under treatment for tuberculosis, sanatorium confinement under treatment for tuberculosis, home dialysis, hospital-based dialysis and kidney transplant) for use in the evaluation of three health care programmes: a programme for mass chest X-ray and tuberculin testing, a screening programme for the prevention of haemolytic disease of the newborn, and a kidney dialysis and transplantation programme.

Stage (ii) Preparation of health state descriptions

Once each unique possible health outcome has been identified, health state descriptions should be prepared to be presented to the subject and/or used by the analyst. As a starting point, health states should be described in functional as opposed to clinical terms. That is, the description should focus on how easy or difficult it is for a person in a particular health state to be able to function. A statement on the level of physical, emotional and social functioning is required. And, since the utility of a specific health state is affected by its duration and prognosis, these should also be specified either in the description itself or as part of the measurement process. For chronic states, the prognosis should be stated as

no change until death and for temporary states it should be stated as no change until the end of the temporary duration specified, at which point the person returns to normal health. Finally, the description should include the age of onset for the state and specify whether or not the state has to be thought of as applying to the subject himself or to someone else.

Following identification of the health states and preparation of health state descriptions, the analyst has three possibilities for describing a health state to a subject.

When the relevant health states for utility measurement are simply those of the patients themselves involved in the study, the individuals can provide a utility measurement for their own health state. At first sight it would seem unnecessary in this case to provide a health state description; however, to enable others to interpret the results health state descriptions may still be required.

This approach (i.e. the use of patient´s own health state) was adopted in the aforementioned study by Churchill et al (1984). Torrance forecasts a considerable future for this approach in clinical trials. Here the quality of life, as measured by utility scores can be determined on each subject in each of the experimental and control groups at baseline and at each follow up point. And/or by asking patients in the study to compare their state of health now with that on entry to the study, changes in utility scores can be measured directly.

However in the case of a subject who is not in a particular health state, he/she must be asked to assess a given state based on description. For example, consultants and graduate students in nursing and health administration were used to assess different health states in an analysis of

a phenylketonuria (PKU) screening programme, (Bush, Chen and Patrick 1973). Similarly McNeil (1981) investigated the attitudes of 37 healthy volunteers, interviewing 12 firefighters and 25 middle and upper management executives to determine their preferences for longevity as against impairment of speech through cancer surgery.

The level of detail in the health state description varies greatly from one study to the next. In the study above relating to speech impairment subjects were presented with a written "scenario" to obtain their attitudes towards the absence of normal speech for various periods of survival In addition, a tape recording was played to respondents, to illustrate the speech capabilities of two patients who had undergone the operation, laryngectomy. By comparison, Patrick and his colleagues (1973) used descriptions that included merely a few key words or phrases which highlighted the chief characteristics of the health states.

Torrance (1984) reports that comparison among different approaches suggests that sometimes utility values differ depending on the level of detail and sometimes they do not. Other investigations have focused on the problem of bias in the answer, as determined by the way in which the health state is described. Torrance´s advice for measuring utilities on the general public is to use abbreviated descriptions to avoid cognitive overload, to supplement those with prior more detailed explanations of the key phrases used in the abbreviated descriptions, and to avoid the framing bias by wording the question in a balanced (positive and negative) manner.

The third possibility for describing health states, which is the

appropriate approach when large numbers of health states are involved, is to use a "Health State Classification System" (HSCS) that incorporates all states of interest. A HSCS is based on the concept that health status can be defined in terms of a number of attributes. Each attribute is divided into a number of mutually exclusive and collectively exhaustive levels.

The specific combination of levels, one from each attribute is taken to represent a unique health state. In this way a HSCS may generate a very large number of health states. For example, if there are ten different levels for each of three attributes, one thousand discrete health states will be defined.

Different health state classification systems have been developed by various analysts for various uses. Bush and his colleagues developed a system for general use with four attributes (mobility, physical function, social function and symptom problem complex) (Kaplan et al 1976), while Rosser (1976) developed a system with just two attributes disability and distress for application to inpatients. Wolfson and colleagues (1982) developed a system for application to stroke patients with 10 attributes (dressing, bathing, continence, eating, transfer, wheelchair, ambulation, understanding, speech and mental status) and Torrance and his colleagues (1982) developed a system for general use with four attributes, (physical function, role function, social emotional function and health problem).

Stage (iii) Selection of Subjects

The selection of subjects or individuals, whose utilities are to be measured is a controversial issue. Different studies have used different types of people. Some have investigated patients´ preferences (Churchill et al 1984) on the grounds that they can best appreciate the implications of particular health states, others have used a random sample of the population on the premise that society´s preferences should count as society´s resources

are being allocated, and others have investigated the preferences of health professionals on the grounds that they are more knowledgeable.

On deciding whom should be asked, the purpose and viewpoint of the study inevitably plays an important role. Thus the patients themselves are the appropriate subjects to ask regarding the utility of their condition in clinical trials. Similarly, informed members of the public are appropriate subjects in a study conducted from the societal viewpoint. "Informed" implies, however, that the subject has a good knowledge of what the specified health state is like. This immediately raises the question of how to describe a dysfunctional health state to a healthy individual who has no prior experience of the particular state? To some extent the problem is overcome by careful design (style and content) of the health state description and through the use of reliable and valid (to be described later) measurement techniques. Emerging evidence also suggests that different groups do not generally produce very different results (Kaplan & Bush 1982, Sackett & Torrance 1978) and hence the problem may not be unduly significant.

Stage (iv) Use of Utility Measurements.

Before considering some of the measurement techniques developed to date, it is useful to go back to distinguish between ordinal, cardinal and ratio scales.

An ordinal scale is simply a rank ordering of health states, in order of their preference with ties allowed, and is sufficient merely for answering questions of the sort "How does the outcome of intervention A compare with the outcome of intervention B?"

Cardinal scales, may be interval or ratio. Measurement on an interval scale

implies that the zero point and the numbers assigned to the entities are arbitrary, save that they order them (as in ordinal measurement) and keep the ratio of the intervals between them the same. This kind of measure is akin to that used for measuring temperature in ^{o}F or ^{o}C and is required to answer questions of the type, "How much more effective is A than B?" However individual scores - like individual temperature measurements - cannot be added up.

With a ratio scale the origin is not arbitrary (i.e. zero means none) and only the unit of measurement is arbitrary (eg. length in milimetres, centimetres or metres). A ratio scale provides values which can be added up (as distances can), and which indicate meaningful ratios between measurements. They provide answers to questions of the form, "Proportionately how much better is A than B?"

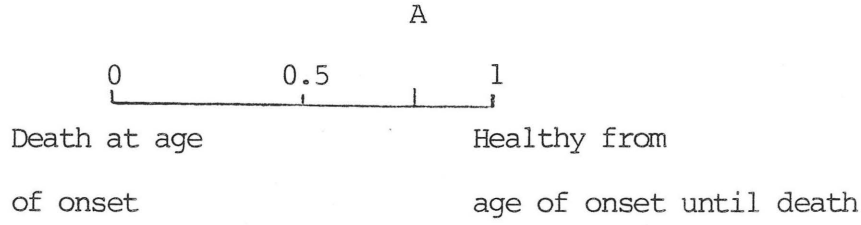
An ordinal scale is clearly the simplest to obtain but it is rarely adequate for use in economic appraisal. In recent years most activity has focused on the development of techniques to produce interval scales and each of the measurement techniques considered here produce interval scales of utility. The rating scale technique, the standard gamble technique and the time trade off technique are described in the next section.

The Rating Scale

A typical rating scale consists of a line drawn on a page with clearly defined endpoints such as "Death, least desirable" at one end and "Healthy, most desirable" at the other. The remaining health states are then located on the line between these two in order of their preference such that the intervals between the placements correspond to the differences in preference between the health states, as perceived by the subject. This is the interval scaling principle.

The rating scale is suitable for measuring preferences for both chronic and temporary health states. Chronic states should be described to the subject as permanent from age of onset until death, with the age of onset and death given. All chronic states with the same age of onset and death are then grouped together and measured relative to each other. Chronic states with different ages of onset and/or death can be measured by using several groups. Each group must have two additional chronic states as reference states for the scale added to it. These might be "healthy (from age of onset until death)" and "death at age of onset". The scale is then measured from 0 assigned to the worst health state of the group and 1 assigned to the best (See Figure 3). The subject is asked to select the best and worst health states from the group and then locate the other states on the scale relative to each other, according to the interval scaling principle described above.

Fig 3



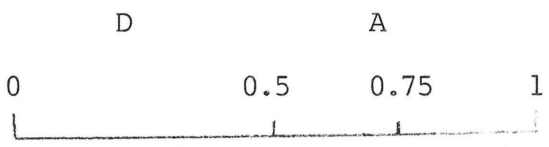


Figure 4

If death is considered the worst state of health and is placed at 0 on the scale, the preference value for each of the other states is simply the scale value associated with its placement. Suppose A represents a given chronic state, as shown in Fig 3, then the preference value can be read from the scale, which in this case is 0.8. However, it may be the case that death D is not considered the worst state and hence is repositioned as depicted in Fig 4 reflecting the subject who prefers to be dead than to be in certain specified chronic states. In this case the preference value for chronic state A must be re-calculated so that a new position for A relative to D can be established on the scale. This may be obtained by applying the formula

$$\frac{X - D}{1 - D}$$

Where X denotes the scale placement of the health state. This will give a measure of the ratio of the preference value to the new scale value. Thus, a preference value of 0.8 may, in this case, be translated into:

$$\frac{0.8 - 0.2}{1 - 0.2}$$

resulting in a placement value on the scale of 0.75 (Fig 4).

When preferences for temporary health states are measured on a rating scale, the states are described to the subject as being of a specified duration after which the person returns to normal health. All temporary health states of the same duration and with the same age of onset are grouped together and measured relative to each other. Temporary health states with different durations and/or ages of onset can be measured using multiple groups.

Each group requires the additional state 'healthy' to be added to it. The subject is asked to locate the best health state (which presumably would be healthy) at one end of the scale and the worst temporary health state at the other. The remaining temporary states are then located on the scale by the subject using the aforementioned interval scale principle.

This procedure is sufficient if the programmes being evaluated involve only morbidity and not mortality and in circumstances when it is not necessary to compare the findings to programmes that do involve mortality. If, however, this is not the case and mortality is encountered then the interval preference values for temporary states must be transformed on to the standard 0-1 health preference scale. To achieve this the worst temporary health state is redefined as a chronic state of the same duration and its preference value is measured by the method described for chronic states. Through the use of a positive linear transformation, that is, increment by a unit value of 1, the values for the remaining states can then be transformed on to the standard 0-1 health preference scale. (This procedure is akin to that of converting degrees fahrenheit to degrees centigrade.)

Standard Gamble

The "Standard Gamble technique", based on the work of Von Neumann and Morgenstern (1953) is used widely as a general measure for utilities and preferences. In recent years, it has been used in the field of health to measure preferences for different health states.

Using this technique subjects are asked to choose between a gamble, with a desirable outcome, with risk P, and a less desirable outcome, with risk 1-P, and a certain option of intermediate desirability. The subject is asked what probability of getting the desirable or less desirable outcome will make him indifferent between the gamble and the certainty.

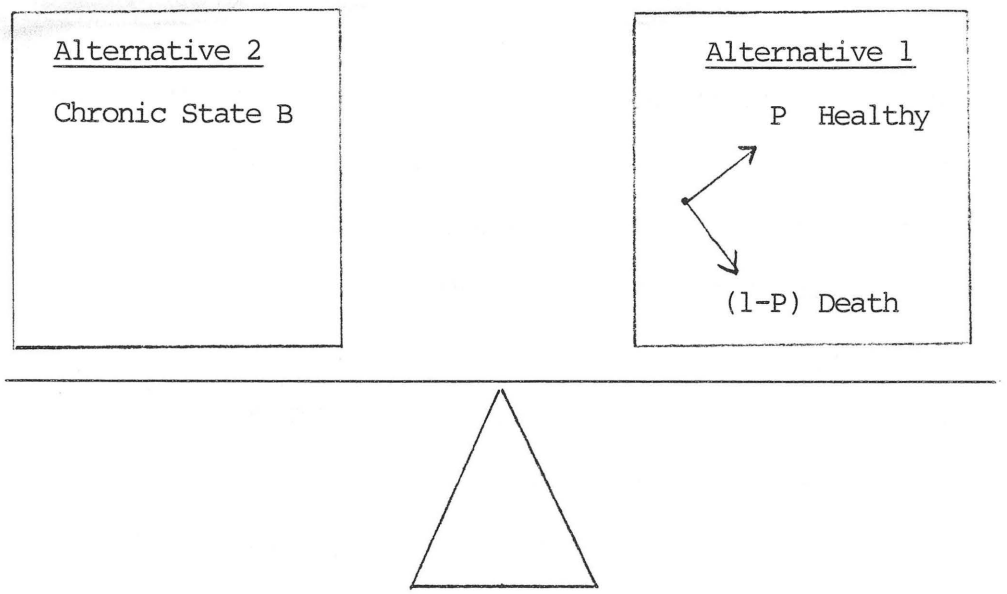
By way of illustration, subjects may be presented with the question:

"Suppose you have a choice between living t years in health state A, or taking a gamble between a P-chance of t years in perfect health (H) and a 1-P chance of t years in state B (which might be coma or some other extreme reference state). What probability, P, would make you indifferent between the sure thing and the gamble?"

The value of P corresponding to the best outcome, perfect health, is 1, and the value of P corresponding to the worst outcome, B, is 0. On answering the question the subject provides a number, P that can be used as the weight assigned to health state A.

The "standard gamble" technique, can be used in the health field to measure preferences for both chronic and temporary health states. Fig 3 illustrates the method for measuring chronic states preferred to death.

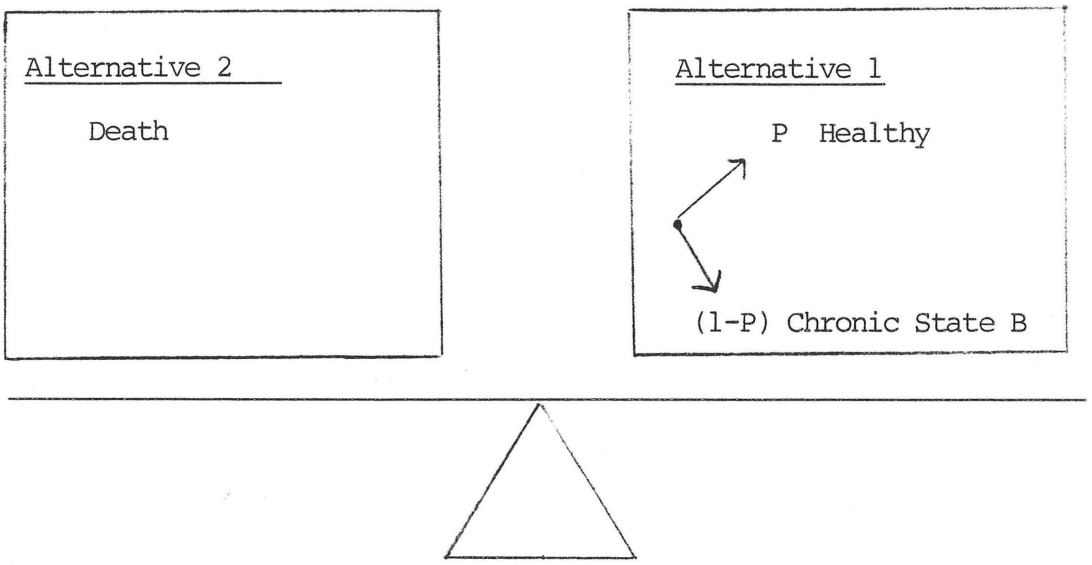
Fig 5



The subject faces 2 alternatives. Alternative 1 is a treatment with two possible outcomes: at a probability ´P´ the patient will return to normal health and live for an additional ´t´ years, or at a probability (1 - P) the patient will die immediately. Alternative 2 has the certain outcome of chronic state B for life (t years). Probability P is varied until the subject is indifferent between the two alternatives. At which point the preference value for chronic state B, (h_B), is simply P, ($h_B = P$).

For measuring chronic states considered worse than death the standard gamble method must be slightly modified. This is illustrated in Fig 6.

Fig 6



Here the gamble alternative (alternative 1) leads to outcomes healthy, at probability P, or chronic state B at probability (1 - P). The certain alternative leads to death.

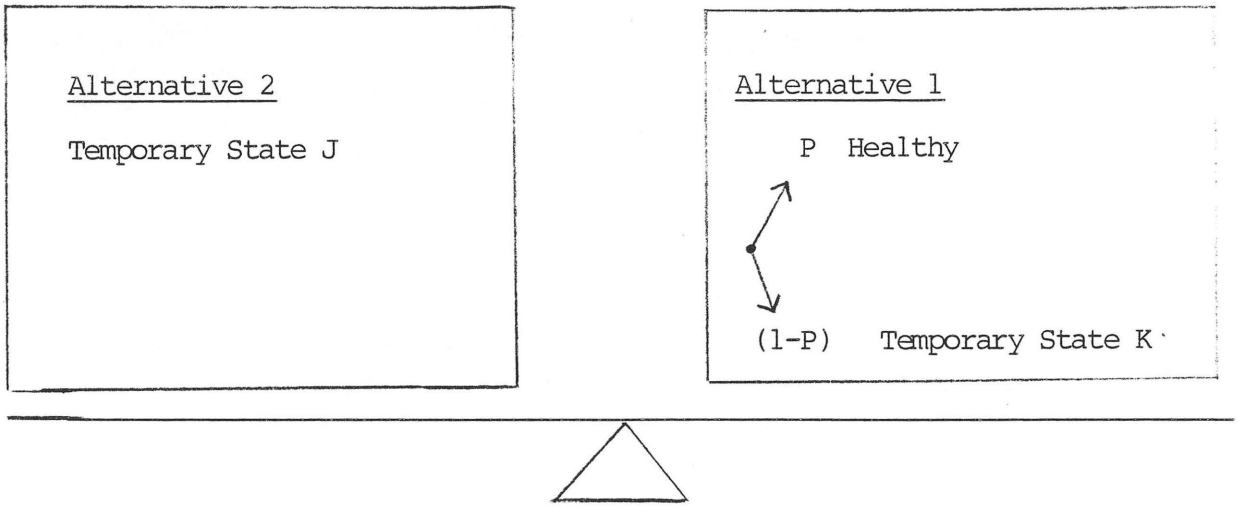
Torrance outlines one way in which this choice may be presented to the subject. Let the subject imagine he is faced with a rapidly progressing terminal disease, which if left untreated will quickly lead to death. A treatment is available however with the probability P of returning the patient to full health, and probability (1 - P) of leaving the subject irreversibly in chronic state B. As before, probability P is varied until the subject is indifferent between the uncertain and the certain alternatives. At this point the preference value for chronic state B is given by the formula:

$$\frac{h(D) - ph(H)}{1 - p}$$

Where h(D) denotes the preference value for death and h(H) the preference value for healthy.

Fig 7 illustrates the standard gamble approach to measuring preferences for temporary health states.

Fig 7



As before, the subject faces two alternatives. Alternative 1 is a treatment with two possible outcomes: at probability P the patient returns to normal health, and at probability (1 - P) the patient suffers from the worst temporary health state, K. Alternative 2 has the certain outcome of an intermediate temporary health state, J. The subject selects probability, P, at which point he is indifferent between the two alternatives. In this way the intermediate temporary health state (J) is measured relative to the best state (healthy) and the worst temporary health state (K).

Using the procedure outlined above, the preference value for temporary health state J is given by the formula,

$$h_J = P + (1 - P)h_K.$$

When mortality is not involved, h_K, the preference value for the worst temporary health state, can be set equal to 0 and hence the preference value for temporary health state J is simply, $h_J = P$. However, when mortality is a consideration and it is desirable to relate these values to the 0-1 dead healthy scale, state K must be redefined as a short duration chronic state, followed by death, and be measured on the 0-1 scale using the technique outlined for chronic states. This, in turn, provides a value for h_K which can then be be used in the formula to enable the value h_J to be calculated.

Time Trade-off

The "time trade-off" technique, pioneered by Torrance, is similar to the standard gamble technique in that it is based on paired comparison and allows the analyst to derive preference values implicitly, based on the subjects' responses to decision situations. It differs, however, in that no probabilities are involved.

The subject is presented with 2 alternatives and asked to select the most preferred. Alternative 1 offers the subject a particular outcome for a specified length of time followed by death, and alternative 2 offers a different outcome for a different length of time. The time is varied until the respondent is indifferent between the 2 alternatives.

As with the standard gamble and rating scale techniques, this approach can be used to measure preferences for both chronic and temporary health states. Fig 8 illustrates the application of the time trade-off technique for chronic states preferred to death.

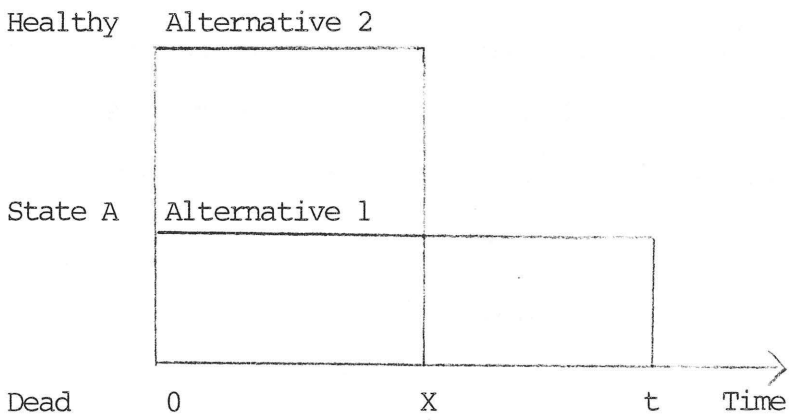


Fig 8

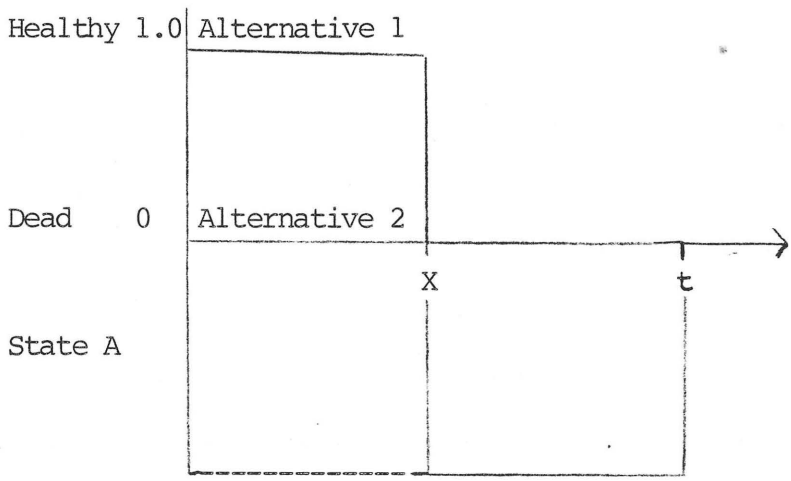
Alternative 1 is chronic state A for time t (i.e. the life expectancy of an individual with the chronic condition) followed by death, and alternative 2 is healthy for time X, where X t, followed by death. Time X is varied

until the subject is indifferent between the two alternatives at which point the preference value for chronic state A (hA), is given by

$$\frac{X}{t}$$

Fig 9 illustrates the procedure for the determination of preference values for chronic states dispreferred to death.

Fig 9



Here the subject is asked to determine the time X such that he/she is indifferent between alternative 1, which represents healthy for time X (where X t) followed by chronic state A until time t, followed by death, and alternative 2, which is to die immediately after birth. At the point of indifference the preference value for chronic state A (hA) is given by the formula:

$$\frac{X}{X - t}$$

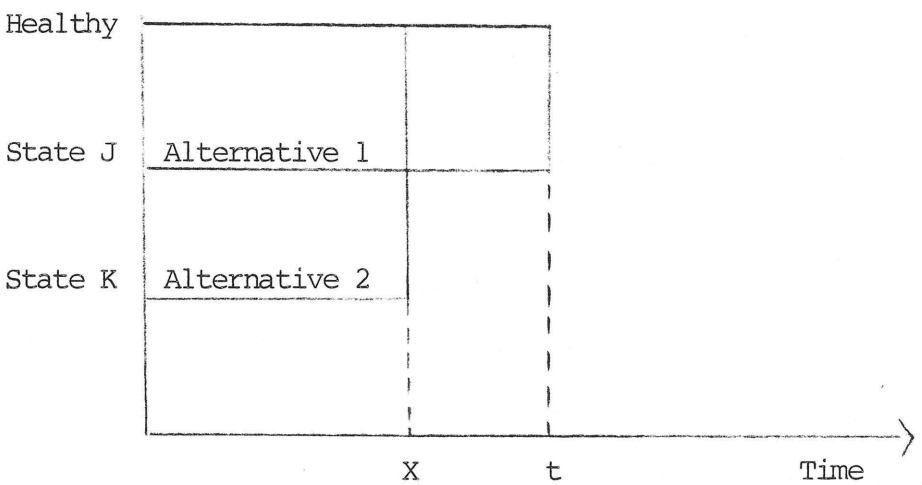
which is derived by equating the two alternatives,

$$1.0X + hA\ (t - X) = 0$$

and solving for hA. *

The application of the time trade-off technique to measure preferences for temporary health states is illustrated in fig 10. The intermediate temporary health state (J) is measured relative to the best state (healthy) and the worst temporary health state (K).

Fig 10



The subject has a choice of 2 alternatives: alternative 1 is intermediate temporary health state J for time t (the time duration specified for temporary states) followed by healthy, and alternative 2 is temporary state K for time X (where X t) followed by healthy. The time X is varied until the respondent is indifferent between the 2 alternatives, at which point the preference value for temporary state J (h_J) is given by

---

* Torrance (1984) points out that in practice one difficulty encountered in this procedure is that although it imposes an upper limit of 1.0 on states preferred to death, it imposes no comparable lower limit on states dispreferred to death. One solution to this is to scale the preference values of those states considered worse than death, so that the worst possible state is assigned a preference value of -1.0.

$$\frac{1 - (1 - h_K)\, X}{t} .$$

If $h_K = 0$, i.e., the worst temporary health state is set equal to 0, h_J equals

$$\frac{1-X}{t}$$

If the preference values for the temporary health states are to be transferred to the 0 - 1, dead healthy scale, then the worst temporary health state must be redefined as a short duration chronic state and measured by the method previously described for chronic states.

As suggested, the rating scale, standard gamble and time trade-off techniques can all be applied to produce interval scales of utility. However, the measurement of utilities or preferences for health is clearly a complex and controversial task. Debate continues over the most appropriate use of those techniques considered here and some investigators have opted for alternative approaches. Of particular interest is a method used by Rosser and Kind (1978) in which subjects were asked to provide a ratio of undesirability of pairs of health states so as to produce a ratio scale of utility. A similar technique is the 'equivalence technique' whereby subjects are asked to identify their point of indifference between keeping alive a group of people in a "standard state" of perfect health and a larger group, whose size is defined by the subject, of less well people.

Are the utilities valid?

The utility values or numerical weights assigned to different health states should, according to Torrance (1976) be non-arbitrary, community based, scientifically measured values reflecting the relative desirability of various states of health. This requires the availability of a reliable and valid measurement instrument(s) which can be used on the general public to quantify the preferences for the relevant states of health.

With this in mind Torrance (1976) carried out an empirical investigation of three of the more commonly administered measurement techniques: he assessed the category method (an application of the rating scale) the standard gamble technique and the time trade-off technique for their feasibility, reliability, validity and comparability. Each health state selected for use in the study was described in a scenario outlining the physical, emotional and social characteristics of the state, and three groups acted as judges, a stratified sample of the population of Hamilton, Ontario, graduates from McMaster University and patients involved in a local home dialysis programme.

The feasibility of each technique was determined by its acceptability to the judges, its ease of use for the interviewers and its cost. Taking the first of these criteria, the subject´s willingness to go through with the interview in all three cases, reflected their acceptability for use on the general public. However, there were noticeable differences in the ease with which the subjects found the techniques - the time trade-off technique proving to be the easiest, the standard gamble questions proving somewhat more difficult and the category scaling proving most difficult.

The professional interviewers found all three techniques easy to learn and straightforward to administer, although the use of a probability wheel was considered essential to enable the administration of the standard gamble technique. (A probability wheel is an adjustable disc with two sectors, each a different colour, constructed so that the relative size of the two sectors can be easily changed to reflect the relative probability of the alternative outcomes.)

Turning to the cost encountered in the application of the three techniques, the standard gamble and time trade-off approaches are inherently expensive, being both time-consuming and requiring an interview for administration. The category method, by comparison, is relatively cheap in that it is less time-consuming and has the potential for being used in the form of a mailed questionnaire.

Focusing on the reliability of the measurement techniques, if a utility can be measured more than once and produce identical results, the measurement technique is said to be reliable. In this study "internal reliability" was tested by using replicated measurements* and "test-retest reliability" was tested by retesting one group of subjects one year later.

When investigating "internal reliability" the question arises of whether the change of the measurement is sufficient to disguise the replication from the subject and yet at the same time insufficient so as not to affect the characteristic being measured. Since no subjects complained of questions being repeated the first objective appeared to be satisfied. However, statistical analysis of the differences between the original measurements

---

* This was not possible for the category scaling technique and hence there were no internal reliability measures for this method.

and the replications, indicated that in this study with the time trade-off technique "the replicated measurement contained a content change such that the modified question was measuring a slightly different phenomenon". Furthermore, it was suggested that, had the sample sizes for the standard gamble been larger, the same conclusion would probably have been achieved.

One year test-retest reliability gave a coefficient of 0.53 for the standard gamble, 0.62 for time trade-off and 0.49 for the category technique. Although the time trade-off technique can be seen to have the highest coefficient of test - retest reliability, the difference is not significant at the 0.05 confidence level. In Churchill´s (1984) study, a 6 week test - retest correlation coefficient produced values ranging from 0.628 to 0.802. This might indicate that people´s preferences shift over time.

Turning to validity, if the measurement technique actually measures what it claims to measure, in this case the utility or strength of a subject´s preference for certain health states, it is said to be valid. "Criterion validity", in which a new measure is assessed against a ´well-accepted´ measure, was applied in this study with the standard gamble technique taken to represent the latter.

The criterion validity of the time trade-off technique, as determined by the coefficient of validity (i.e. the product moment correlation coefficient between the measure under investigation and the criterion measure) was concluded to be "satisfactory". On the other hand the criterion validity of the category method was, found to be "significantly poorer", and when recalculated using the time trade-off as the "well-accepted" measure, the results were not significantly improved. This seems to suggest that at least for the category method, criterion validity is unsatisfactory.

Finally, the comparability of the three techniques was assessed in terms of whether or not they produce the same values, and if not whether the values

derived are related in some systematic way so as to enable conversion curves to be constructed. When addressing this question to the measurement of population mean values, the time trade-off technique appeared to give equivalent results to the standard gamble technique, with a relationship between the two measures of, standard gamble = time trade-off. However, when the question was addressed to the measurement of individual values the relationship was "not so clear, but it seems likely that the same function may hold".

The category scaling technique produced significantly different values for both individual and population mean values from those derived by either of the other two techniques. That said however, there were systematic differences, for population mean values, between measures obtained by category scaling and those obtained by the time trade-off.

All this suggests that the time trade-off technique is the best of the three methods tested for measuring preferences for health states, with the standard gamble technique coming a close second. This study, and others like it, also serves a useful purpose in highlighting some of the inherent problems and uncertainties encountered in preference measurement. By way of example, differences in demographic characteristics such as age, sex, religion etc, cannot fully account for the not insignificant differences in individual´s health state preferences. Sackett and Torrance (1978), found a standard deviation between scores of about 0.30 for individual preferences among the public for a single health state. This notwithstanding, the differences are less apparent among more homogenous subjects with a good knowledge of the health state. In application of the time trade-off technique (Torrance 1976, 1984) 29 home dialysis patients rating the home dialysis scenario resulted in a standard deviation of 0.18 compared to 0.28 for the general public.*

---

* This problem of differences between individual preferences can largely be overcome by taking the mean value of a large group of subjects.

Applications and Discussion.

The evaluation of health states by "psychometric methods" is an exciting, innovative feature of current research on health indicators. As derived here, the values have the interval scale property which makes them useful for evaluative research and for projecting and comparing the benefits of alternative health programmes. It will doubtless be some time before measurement techniques have been developed which satisfy more fully criteria for reliability, validity, comparability and generalisability of social preferences, (Patrick, Bush, Chen 1973) and the indices that they produce are accepted as valid inputs to decision making. However, if further research is successful in developing health status indices, which are acceptable to decision makers, then clearly they will be powerful tools for all aspects of health care policy making.

It was less than two decades ago when, in one of the earliest recorded applications Klarman and his colleagues (1968) introduced the concept of "quality adjusted" life years gained in a cost effectiveness analysis of different treatments for renal failure. It was assumed that one year of life gained by transplant was equivalent to 1.25 years gained by dialysis, reflecting the higher quality of life under transplantation. Since then, there has been a rapid inrease in the literature concerned with measuring the quality of life and research has progressed a long way.

Rosser (1983) provides a historical review of health indicators that claim to be direct assessments of a population´s health. Under the heading, ´The Phase of Cardinal Measurement´ the classic paper of Bush and his colleague (Fanshel and Bush 1970) is discussed. They made a significant contribution using psychometric scaling techniques to develop a health index (formerly the Function Status Index and lately the Index of Wellbeing), which has

since been modified and utilised in several applications including a tuberculosis prevention and treatment programme in New York (Fanshel and Bush 1970), a phenylketonuria (PKU) screening programme (Bush et al 1973) and a large household survey (Kaplan 1976). Reynolds and colleagues (1974) also claimed to have applied a modified version of the index in a survey of two counties in Alabama.

Card's group in Glasgow focused primarily on the measurement of utilities of states of illness for the purpose of formalising clinical decisions, by way of incorporating the utility values into decision models. In particular they studied gastro-intestinal diseases and utilities of head injury; furthermore they anticipated the conversion of utilities into money equivalents for use in CBA (Card 1975) as did Culyer, Lavers, Williams in York (1971, 1972).

At about the same time as Bush began his prolific research, Torrance's group at McMaster University published a cost-utility model (Torrance et al 1972) which has since been further developed and applied to several health care programmes including tuberculosis screening, haemolytic disease, Rhesus disease, renal dialysis and more recently neonatal intensive care of very low birth weight infants. In addition two surveys of the general public to measure health state utilities have been carried out with one being based on a multi attribute health state classification system (as previously described).

Further contributions in this field of work have been made by Rosser, Watts and Kind, who have focused particularly on indicators of hospital performance (Rosser, and Watts 1972, Rosser, 1976). They used two scaling methods, psychometric and behavioural. The former being based on magnitude estimation but including a lengthy interview procedure devised by Gibbs and

Wishlade in their work on crime seriousness, and the latter as already mentioned, was obtained by the analysis of legal awards for non-pecuniary consequences of personal injury and industrial accidents and disease. (This scaling method is significant in that unlike those described it reflects actual behaviour, and values are inferred from an existing resource allocation process.)

Thus research into health status measurement has made considerable progress in a relatively short space of time, and yet there are still a number of controversial and unresolved issues.

To begin with the whole concept of combining the impact of a given health care activity on morbidity and mortality into a single measure (QALYs) gained is still debatable. It needs to be justified methodologically and ethically. It has to be established that the users of the studies fully understand the trade-offs built into the calculations.

Secondly, in measuring utilities the question arises of whose values should count? That is, who should place values on states of health? To provide an answer to this question it must be established whether there are differences in opinion about the severity of illness between individuals and between different socio-economic groups and, if there are differences, can they be aggregated or are they mutually exclusive?

A third, and particularly important, issue concerns the specificity or generalisability of the utility values. Can a universal set of health states utilities be determined and used in all studies or does each study require its own utilities?

Finally one must ask which technique is best to use and whether they are

subject to different biases (such as risk aversion in the case of the standard gamble technique and time preference in the time trade-off technique).

The purpose of this paper has been to expose some of the techniques currently being developed and utilized in the determination of health state utilities for use in economic appraisal. It has been demonstrated that health state preferences can be measured using these techniques, albeit somewhat imprecisely. However, as the impact of health care activities on the quality of life plays an increasingly important role, so the need to evaluate this objective becomes more and more apparent. Whereas the benefits of medicines introduced in the 1940´s and 1960´s were easy to measure, in terms of reduced hospital costs, deaths and sickness absence payments, as depicted in the introduction to the proceedings of the Office of Health Economics meeting on the measurement of social benefits of medicine, "there is now an overwhelming need to quanitfy the benefits of the ´quality of life´ medicines, of the 1980´s " (Teeling Smith 1983). In the allocation of scarce resources available to society, it is irresponsible to omit from economic appraisal, quality of life and other intangible benefits (which receive high priority in the hierachy of objectives of health care providers and consumers) simply because of difficulties in measurement and evaluation.

References

Acton J P, (1973). Evaluating Public Programs to Save Lives: The Case of Heart Attacks. In Rand Corporation Report R-950-RC. Santa Monica, California.

Adler M W, Belsey E M, McCutchan J A and Mindel A, (1983). British Medical Journal, 286, 1621-1624.

Boyle M H, Torrance G W, Sinclair J C and Horwood S P, (1983). New England Journal of Medicine, 308, 1330-7.

Bush J W, Chen M M and Patrick D L, (1973). Health Status Index in Cost Effectiveness: Analysis of PKU Programme. In Health Status Indexes, Proceedings of a Conference conducted by Health Services Research. Ed. Berg R L.

Card W I, (1975). Ciba Foundation Symposium 34 (New Series) Elsevier-Excerpta Medica, Amsterdam.

Churchill D N, Morgan J and Torrance G W, (1984). Peritoneal Dialysis Bulletin, January - March, 20-23.

Culyer A J, (1978). Measuring Health: Lessons for Ontario. University of Toronto Press.

Culyer A J, Lavers R J and Williams A, (1971). Social Trends, 1, 31-42. HMSO.

Culyer A J, Lavers R J and Williams A, (1972). Health Indicators. In Social Indicators and Social Policy. Ed. Shonfield A, Shaw. S. Heinemann, London.

Drummond M, (1981), Studies in Economic Appraisal in Health Care. Oxford University Press.

Drummond M, (1983). Economic Assessment of therapy. In Measuring the Social Benefits of Medicine. Ed. Teeling Smith G. OHE, London.

Drummond M, (1984). Economic Evaluation in the Development and Promotion of Medicines. Paper for Conference entitled New Challenges in Drugs' Development, Promotion and Innovation Strategies held in Paris.

Fanshel S and Bush J W, (1970). Operations Research, 18, 1021-1066.

Gibbs R J, (1972) Home Office Police Planning Organisation Report. No. 10/72

Grogono A W and Woodgate D J, (1971). Lancet, 1024-1026.

Harris A I, Cox E and Smith R W, (1971). Handicapped and Impaired in Great Britain. HMSO, London.

Kaplan R M and Bush J W, (1982). Health Psychology, 1, 61-80.

Kaplan R M, Bush J W and Berry C C, (1976). Health Services Research, 11, 478-507.

Karnofsky D A and Burchenal J H, (1949). The Clinical Evaluation of Chemo-therapeutic Agents in Cancer. In Evaluation of Chemotherapeutic Agents. Ed. Maclead C M. Columbia University Press.

Klarman H E, Francis J O's and Rosenthal G D, (1968). Medical Care, 6, 48-54.

Llewellyn - Thomas M, Sutherland H J, Tibshirani R, Ciampi A, Till J E and Boyd N F, (1982). Medical Decision Making, 2, 449-462.

McNeil B J, Weichselbaum R and Pauker S G, (1981). New England Journal of Medicine, 305, 982-987.

Mishan E, (1971). Journal of Political Economy, 79, 687-705.

Neumann J von and Morgenstern D, (1953) Theory of Games and Economic Behaviour. Third Edition. New York, Wiley.

Patrick D L and Bush J W, Chen M M, (1973). Health Services Research, 8, 228-245.

Pliskin J S, Shepherd D S and Weinstein M C, (1980). Operations Research, 28, 206-224.

Reynolds W J, Rushing W A and Miles D L, (1974). Journal of Health and Social Behaviour, 15, 271-89.

Rosser R M, (1976). Medical Care, 14, Supplement, 138-147.

Rosser R, (1983). Issues of Measurement in the Design of Health Indicators: a review. In Health Indicators. Ed. Culyer A J. Martin Robertson.

Rosser R and Watts V C, (1972). International Journal of Epidemiology, 1, 361-368.

Rosser R M and Kind P (1978). International Journal of Epidemiology, 7, 347-358.

Sackett D L and Torrance G W, (1978). Journal of Chronic Diseases, 31, 697-704.

Schelling T C, (1968). The Life you save may be your own. In Problems in Public Expenditure Analysis. Ed. Samuel B. Washington D C.

Spitzer W O, Dobson J J, Hall J, Chesterman E, Levy J, Shepherd R and Battista R, (1981). Journal of Chronic Diseases, 34, 585-597.

Sutherland H J, Dunn V and Boyd N F, (1983). Medical Decision Making, 3, 477-487.

Teeling Smith G (1983). Measuring the Social Benefits of Medicine. OHE, London.

Thaler R and Rosen S, (1975). The Value of Saving a Life: Evidence from the labour market. In Household Production and Consumption. Ed Terleckyj N. National Bureau of Economic Research.

Torrance G W, (1976). Socio Econ Plan Sci, 10, 129-136.

Torrance G W, (1984). Health Status Measurement for Economic Appraisal. Paper presented to Health Economists' Study Group meeting, Aberdeen.

Torrance G W and Zipursky A, (In Press) Clinics in Perinatalogy.

Torrance G W, Thomas W H and Sackett D L, (1972). Health Services Research 7, 118-133.

Torrance G W, Sackett D L and Thomas H T, (1973). Utility Maximisation Model for Program Evaluation: A Demonstration Application. In Health Status Indexes, Proceedings of a Conference conducted by Health Services Research. Ed. Berg R L.

Torrance G W, Boyle M H and Horwood S P, (1982). Operations Research, 30, 1043-69.

Weinstein M C, (1981). Medical Decision Making, 1, 309-330.

Wolfson A D, Sinclair A J, Bombardier C and McGeer A, (1982). Preference Measurement for Functional Status in Stroke Patients: Inter-Rater and Inter-Technqiue Comparisons. In Values and Long Term Care Eds. Kane R, Kane R. D.C. Health Publishers.