MEASUREMENT OF HEALTH STATE UTILITIES FOR ECONOMIC APPRAISAL

A Review

George W. TORRANCE

McMaster University, Hamilton, Ont., Canada L8S 4M4

Final version received November 1985

Health status measurement for use in economic appraisal of health care programmes is reviewed in this paper, with particular emphasis on utility measurement. A framework for economic appraisal is presented displaying the various components that must be measured, and showing how the three forms of analysis (cost-effectiveness analysis, cost-benefit analysis and cost-utility analysis) relate to the framework and to each other. One of the components in the framework is health status; it can be measured using ad hoc numeric scales, willingness to pay/receive or health state utilities. The determination of health state utilities is reviewed covering the following issues: alternative sources of utilities, health state descriptions, multi-attribute utility approach, determination of appropriate subjects, utility aggregation, and accuracy. Three measurement techniques for health state utilities are described in detail: rating scale, standard gamble, and time trade-off. The use of these methods is described for both chronic and temporary health states and for both health states considered better than death and those considered worse than death.

1. Introduction

The goal of all health services activities and programs is to improve the health of people. Thus, it is not surprising that, over the years, there has been considerable interest and activity in developing methods to measure quantitatively the health status of individuals and populations. The objective is to develop a general measure that would allow changes in health status over time and differences in health status among groups to be identified and quantified. One important application of such a measure is in the economic appraisal of health care programmes for resource allocation decisions. This paper provides a review of health status measurement for use in economic appraisal, with particular emphasis on one approach—the use of utility measurements. The paper begins with a review of the techniques of economic appraisal.

2. Economic appraisal

2.1. Overview

An economic appraisal of a health care programme compares the re-
sources consumed by a programme (the costs) with the health improvement created by the programme (the consequences). Fig. 1 displays the components that can enter into such an appraisal. The resources consumed by the programme can be of two types: direct costs ($C_1$), the costs of physicians' time, hospitals, drugs, and other health care costs; and indirect costs ($C_2$), the cost of lost production, if any, because the patients participate in the programme (e.g., extra time off work to receive the intervention). In addition, a third type of cost ($C_3$), is sometimes identified. This would represent the monetary value of the pain, grief and suffering of the patient and family, if any, due strictly to participation in the programme.

The output created by a health care programme is health improvement. As fig. 1 displays, this can be measured a number of ways. First ($E$), it can be measured in units that are natural to the programme or disease; for example, cases found, cases prevented, disability-days prevented, hospitalization-days prevented, lives saved, or life-years gained. Second ($B$), one can measure the economic benefits associated with the health improvement caused by the programme. Like costs, economic benefits may also be of three types. Direct benefits ($B_1$) are the savings in health care costs because the programme makes people healthier and they consequently use fewer health care resources. Indirect benefits ($B_2$) are the production gains to society because more people are well, or alive, and able to return to work. Intangible benefits ($B_3$) represent the monetary value of the reduction in pain, grief and suffering of the patient and family due to the improved health outcome. Third ($V$), one can measure the value to the patient, family or society of the health
improvement itself, regardless of any economic consequences. Approaches here can be classified into three groups: those that use ad hoc numeric scales (S), those that use monetary quantities based on the individual's willingness to pay or willingness to receive (W), and those that use quality-adjusted life-years (QALY's) based on utility measurements (U).

2.2. Forms of analysis

The components displayed in fig. 1 can be compared in a variety of ways and this leads to the different forms of economic appraisal and their variations. The simplest possibility is a cost analysis, which investigates only C1 or sometimes C1+C2. This is not really a form of economic appraisal at all because consequences are ignored.

2.2.1. Cost-effectiveness analysis

Cost-effectiveness analysis (CEA) determines a cost/effectiveness ratio. The numerator, depending on the study and on the viewpoint taken, may be total cost (C1+C2), net health care cost (C1-B1) or net economic cost to society (C1+C2-B1-B2). The denominator is the measure of health effect most relevant to the program under study. Thus, the results of a CEA are expressed in terms such as pounds per case of disease A prevented, or pounds per life saved, or pounds per life-year gained. CEA is useful in comparing alternative programmes whose effects are measured in the same units – for example, alternative programmes to treat hypertension might be compared in terms of pounds per mm.Hg. blood pressure reduction, or alternative programmes to treat end-stage renal disease could be compared in terms of pounds per life-year gained. However, CEA is not very helpful in assessing a single program, because there is nothing to which one can compare the C/E ratio. Moreover, CEA cannot be used to compare disparate alternatives such as comparing kidney dialysis for renal patients to home care for the frail elderly, since the denominators of the C/E ratios will be in different units. Finally, CEA is not suited to the analysis of programmes that have several types of clinical effects – for example, reductions in both morbidity and mortality.

2.2.2. Cost-benefit analysis

Cost–benefit analysis (CBA) determines the net social benefit of the programme, NSB=B1+B2-C1-C2. Although B3 and C3 are normally recommended for inclusion, this is seldom accomplished because of difficulties measuring these components. The CBA decision rule is straightforward; if NSB>0, the project is cost-beneficial and it should be implemented; if NSB<0, it should not. If there are multiple alternatives of comparable scale, the one with the largest NSB is the best. Potentially, CBA overcomes
the disadvantages of CEA; a decision can be made on a single programme, and disparate effects in the same or different programmes can be compared. However, as countless critics have pointed out [Klarman (1982)], CBA as formulated above is seriously flawed because of the $B2$ term. This term biases the results in favour of programmes for those who work and earn, as opposed to programmes that improve the health of children, housewives and the elderly. There are two aspects to this criticism. One is that the true production gain is underestimated, the other that there is more value to improved health than just production gains. Regarding the first aspect, if $B2$ only incorporates individuals in the labour market, it underestimates the true magnitude of the production gain to society due to improved health. Production gains of housewives, for example, should also be included as part of $B2$ [Klarman (1982)]. However, even with this correction, the CBA formulation is still inadequate for decision making. It provides a measure of the net resource drain of a particular programme on society, but this is only one side of the analysis (costs but not consequences), and is insufficient for decision making. The value of the consequences — that is, the value of the improved health itself — must also be entered into the analysis. For a CBA formulation this value must be measured in monetary units. Willingness to pay ($W$) is considered the theoretically correct approach, although it has proved difficult to implement in practice [Klarman (1982)]. This is the amount that an individual would be willing to pay to purchase the health improvement itself, everything else being equal, if it were available on the market. A closely allied value is the amount that an individual would be willing to receive, everything else being equal, in compensation for the health decrement. Court awards for personal injury cases could be viewed as consisting, in part, of this willingness to receive component. [In a related development Rosser and Watts (1975) used court awards to determine a ratio scale for their set of health states.] Willingness to receive is the selling price, willingness to pay is the purchase price, both measure $W$ the monetary value of the health improvement per se. Given a value for $W$, then, the complete CBA formulation would be

$$NSB = BI + B2 + W - C1 - C2.$$  

Moreover, it is not double counting to include $BI$ and $B2$ along with $W$ as economic benefits in the same formula; recall that $W$ is measured net of any effect on earnings and health care expenditures.

2.2.3. Cost–utility analysis

Cost–utility analysis (CUA) is a special form of cost-effectiveness analysis in which the measure of effect is quality-adjusted life-years (QALY’s) gained. The quality adjustment weight is a utility value which can be measured,
estimated, or sometimes taken from past studies. The advantage of CUA over CEA is that it uses a common unit of measure, QALY's gained, for all programmes and thus allows comparisons across all programmes. See table 1 for a recent example of such a comparison. Moreover, as the bank of programmes analyzed with CUA grows, it becomes increasingly easy to use the technique to analyze single programmes. For example, based on their experience, Kaplan and Bush (1982) suggest that programmes with a C/E ratio of less than $20,000 per QALY gained are 'cost-effective by current standards', those with a ratio of $20,000 to $100,000 per QALY gained are 'possibly controversial, but justifiable by many current examples', and those over $100,000 are 'questionable in comparison with other health care expenditures'. Thus CUA, like CBA, can overcome the disadvantages cited for CEA. However, based on our experience, CUA has the additional advantage compared to CBA of being more compatible with the decision making style of planners and managers in the health care field and, thus, being more acceptable to them. Moreover, CUA explicitly incorporates the quality of life associated with the health outcomes.

Cost-utility analysis is appropriate in the following situations: when quality of life is the important outcome, when quality of life is an important outcome, when the programme under evaluation affects both morbidity and mortality and you wish to have a common unit of outcome that combines both effects, when the programmes being compared have a wide range of different kinds of outcomes, and when you wish to compare a programme to others that have already been evaluated using CUA. Cost-utility analysis is inappropriate or unnecessary when the effectiveness data for final health outcomes is not available, when the effectiveness data show that the programmes being compared are all equally effective, when quality of life is important but it can be captured by a single variable measured in easily understood natural units, or when it is clear that the extra cost of obtaining and using utility values cannot change the results.

In reading the literature beware that cost-utility analysis appears under several different labels. We have recently adopted the CUA nomenclature to distinguish the technique from other approaches. However, in our early work (Torrance (1970)) and in the current work at Harvard University (Weinstein and Stason (1977), Weinstein (1981)) CUA is treated as a special case of CEA. Most of the work from Bush's group at San Diego uses the terminology health status index approach (Fanshel and Bush (1970), Bush et al. (1972), Bush et al. (1973)), although recently they also have used the cost-utility label (Kaplan and Bush (1982)).

2.3. Viewpoint

It is important to specify the viewpoint of an economic appraisal. Possible,
Table 1
Comparative cost-utility results for selected programs.*

<table>
<thead>
<tr>
<th>Program</th>
<th>Reported cost/QALY&lt;sup&gt;b&lt;/sup&gt; gained in U.S. dollars (year)</th>
<th>Adjusted&lt;sup&gt;c&lt;/sup&gt; cost/QALY&lt;sup&gt;b&lt;/sup&gt; gained in U.S. dollars 1983</th>
</tr>
</thead>
<tbody>
<tr>
<td>PKU screening [Bush et al. (1973)]</td>
<td>&lt;0</td>
<td>&lt;0</td>
</tr>
<tr>
<td>Post-partum anti-D [Torrance and Zipursky (1977)]</td>
<td>&lt;0</td>
<td>&lt;0</td>
</tr>
<tr>
<td>Ante-partum anti-D [Torrance and Zipursky (1984)]</td>
<td>1,220</td>
<td>1,220</td>
</tr>
<tr>
<td>Coronary artery bypass surgery for left main coronary artery disease [Weinstein (1981)]</td>
<td>3,500</td>
<td>4,200</td>
</tr>
<tr>
<td>Neonatal intensive care, 1000–1499 g [Boyle et al. (1983)]</td>
<td>2,800</td>
<td>4,500</td>
</tr>
<tr>
<td>T4 (thyroid) screening [Epstein et al. (1981)]</td>
<td>3,600</td>
<td>6,300</td>
</tr>
<tr>
<td>Treatment of severe hypertension (diastolic ≥ 105 mm Hg) in males age 40 [Stason and Weinstein (1977)]</td>
<td>4,850</td>
<td>9,400</td>
</tr>
<tr>
<td>Treatment of mild hypertension (diastolic 95–104 mm Hg) in males age 40 [Stason and Weinstein (1977)]</td>
<td>9,880</td>
<td>19,100</td>
</tr>
<tr>
<td>Estrogen therapy for postmenopausal symptoms in women without a prior hysterectomy [Weinstein (1980)]</td>
<td>18,160</td>
<td>27,000</td>
</tr>
<tr>
<td>Neonatal intensive care, 500–999 g [Boyle et al. (1983)]</td>
<td>19,600</td>
<td>31,800</td>
</tr>
<tr>
<td>Coronary artery bypass surgery for single vessel disease with moderately severe angina [Weinstein (1981)]</td>
<td>30,000</td>
<td>36,300</td>
</tr>
<tr>
<td>School tuberculin testing program [Bush et al. (1972)]</td>
<td>13,000</td>
<td>43,700</td>
</tr>
<tr>
<td>Continuous ambulatory peritoneal dialysis [Churchill et al. (1984a)]</td>
<td>35,100</td>
<td>47,100</td>
</tr>
<tr>
<td>Hospital hemodialysis [Churchill et al. (1984a)]</td>
<td>40,200</td>
<td>54,000</td>
</tr>
</tbody>
</table>

*These studies use similar, but not identical, methods. Generally, costs are net health care costs; however, discount rates and preference weights are not completely consistent. Differences in methods should be considered when comparing the relative cost-utility. For details, see original sources. Table taken from Torrance and Zipursky (1984).

<sup>b</sup>QALY denotes quality-adjusted life-year.

<sup>c</sup>Adjusted to 1983 dollars according to the U.S. Consumer Price Index for Medical Care for all urban consumers. Source: U.S. Bureau of Labor Statistics, Monthly Labor Review.
viewpoints include that of the patient and family, the hospital, the health care sector, the entire health and social services sector including education, and the global or societal viewpoint. The appropriate viewpoint depends upon the question to be answered. For example, a patient choosing among several alternative treatments would be interested in the viewpoint of himself and his family; a health care planning body attempting to allocate its limited health care budget would be interested primarily in the health care viewpoint; a social planner concerned with the overall impact of alternative health care policies and programmes would be interested primarily in the societal viewpoint. In general, the societal viewpoint is the appropriate one for public policy decision making. Although a study will normally have one primary viewpoint, it may also have additional supplementary viewpoints to provide additional information. For example, although home dialysis may be more cost-effective than hospital dialysis from the societal point of view, the reverse may be true from the viewpoint of some patients and their families. This insight could be important in forecasting the likely enrolment for a proposed home dialysis programme.

2.4. Health status measurement

The early work on economic appraisal of health care programmes concentrated on columns C, E and B of fig. 1 (costs, effects and economic benefits). Data from these columns were used in CEA's and/or CBA's. More recently, methodological developments have focused on column V, the quantitative valuation of the health improvement itself (health status measurement).

One approach in column V is the use of ad hoc numeric scales to quantify the health states (box S in fig. 1). This approach involves assessing the individual on a number of aspects of his or her health, assigning numeric scores to each assessment, and adding up the scores, sometimes with weights. Examples of this approach include the Karnofsky index [Karnofsky and Burchenal (1949)], the Harris index [Harris et al. (1971)], the Grogono-Woodgate index [Grogono and Woodgate (1971)], and Spitzer's QL-index [Spitzer et al. (1981)]. For application in economic appraisal these indexes could be used as a measure of effect in a cost-effectiveness analysis. However, all such indexes are essentially arbitrary and have serious methodological problems [Culyer (1978), Hutchinson et al. (1979)].

A second approach to measuring the value of health improvement is the willingness to pay/receive approach (box W in fig. 1) first proposed by Drèze (1962). In discussing this approach Schelling (1968) drew a clear distinction between the value of livelihood (B2) and the value of life (W). One of Schelling's students, Acton, applied the approach in a study of life-saving techniques for victims of sudden heart attacks [Acton (1973)]. Thompson et
al. (1982, 1984) used the technique on arthritis patients. Rosser and Watts (1975) measured the willingness to receive as determined by court awards for disabilities in order to develop a ratio scale for health states. The willingness to pay/receive approaches determine a monetary value $W$ which can be used in a cost–benefit formulation as described earlier. However, practical difficulties in measuring $W$ have discouraged widespread use of the approach [Klarman (1982), Muller and Reutzel (1984)].

The third approach to measuring the value of health improvement is through the use of utilities and quality-adjusted life-years (box $U$ in fig. 1). This approach, associated with cost–utility analysis, is reviewed in considerable detail in the next section of the paper.

3. Health state utilities

The basic idea of health state utilities and quality-adjusted life-years is quite simple. Utilities are 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. The utility values reflect the quality of the health states and allow morbidity and mortality improvements to be combined into a single weighted measure, QALY's gained. For example, if a 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 QALY’s gained for that year would be 0.25 for individual A plus 0.50 for individual B for a total of 0.75.

3.1. Sources of utility values

Analysts have three choices in determining the utility values for use in a study: they can estimate the values using judgment, they can look for suitable published values in the literature, or they can measure the values.

3.1.1. Judgment

The simplest method of obtaining a health state utility is to use judgment to estimate the utility value, or to estimate a range of plausible values. Judgments can be simple estimates made by the analyst or by a few physicians [for examples, see Weinstein (1981)], or they can be formal measurements made on a small non-random convenience sample of physicians or other experts [for examples, see Torrance et al. (1973), Pliskin et al. (1980)]. When judgmental utility values are used, it is essential to undertake extensive sensitivity analysis to determine the robustness of the conclusions to the judgments.

The judgmental approach has been used extensively by the group at Harvard [see, for example, Weinstein (1981)]. It has the advantage of being
quick and inexpensive. Moreover, it may be sufficient, if the sensitivity analysis shows that the conclusions are relatively insensitive to wide changes in the utility values. Thus, the judgmental approach may be a good way to begin. It may also be a useful approach if the decision makers, for example a health planning council, wish to see the impact on the study results of their own judgments about the quality of the health states. However, if the analysis shows that the results are sensitive to utility values, one would want to obtain utilities that are more credible, either from the literature or by measurements.

3.1.2. Literature

In some cases it may be possible to use existing utility values available in the literature. It would be important to determine that the health states used in the measurement match those of your own study, that the subjects used in the measurement are appropriate to your own study, and that the measurement instruments used were credible.

There is a growing literature of studies in which utilities have been measured for a few specific health states. For example Sackett and Torrance (1978) presented utilities for depression, home confinement for tuberculosis, home confinement for an unnamed contagious disease, hospital confinement for tuberculosis, hospital confinement for an unnamed contagious disease, hospital dialysis, home dialysis, kidney transplant, mastectomy for breast cancer, and mastectomy for injury. Churchill et al. (1984b) reported utilities for hospital dialysis and continuous ambulatory peritoneal dialysis. Utilities for loss of speech due to laryngectomy have been reported by McNeil et al. (1981). A group at Princess Margaret Hospital in Toronto have reported utilities for cancer related states [Llewellyn-Thomas et al. (1982, 1984), Sutherland et al. (1983)]. Utilities for different levels of angina pain have been reported in two studies [Pliskin et al. (1980), Read et al. (1984)].

There are two studies in which a classification system covering a fairly wide range of health states has been established, and utility values have been measured for the states. One is a study by Bush and his colleagues reported in Kaplan et al. (1976), in which the measurements were performed using a rating scale technique on the general public in San Diego. The other is a McMaster University study reported in Torrance et al. (1982) in which the measurements were performed on parents of school age children in Hamilton using both rating scale and time trade-off measures within a multi-attribute utility theory framework. The health state classification system from the McMaster study is shown in table 2. The formula for use with the table can be determined from the original publication [Torrance et al. (1982)] or, alternatively, it is described in a simpler form in the book by Drummond et al. (1986).
Table 2
Health state classification system (age ≥ 2 years).

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>physical function: mobility and physical activity*</td>
</tr>
<tr>
<td>Level x1</td>
<td></td>
</tr>
<tr>
<td>1 P1</td>
<td>Being able to get around the house, yard, neighborhood or community WITHOUT HELP from another person; AND having NO limitation in physical ability to lift, walk, run, jump or bend.</td>
</tr>
<tr>
<td>2 P2</td>
<td>Being able to get around the house, yard, neighborhood or community WITHOUT HELP from another person; AND having SOME limitations in physical ability to lift, walk, run, jump or bend.</td>
</tr>
<tr>
<td>3 P3</td>
<td>Being able to get around the house, yard, neighborhood or community WITHOUT HELP from another person; AND NEEDING mechanical aids to walk or get around.</td>
</tr>
<tr>
<td>4 P4</td>
<td>NEEDING HELP from another person in order to get around the house, yard, neighborhood or community; AND having SOME limitations in physical ability to lift, walk, run, jump or bend.</td>
</tr>
<tr>
<td>5 P5</td>
<td>NEEDING HELP from another person in order to get around the house, yard, neighborhood or community; AND NEEDING mechanical aids to walk or get around.</td>
</tr>
<tr>
<td>6 P6</td>
<td>NEEDING HELP from another person in order to get around the house, yard, neighborhood or community; AND NOT being able to use or control the arms and legs.</td>
</tr>
<tr>
<td>X2</td>
<td>role function: self-care and role activity*</td>
</tr>
<tr>
<td>Level x2</td>
<td></td>
</tr>
<tr>
<td>1 R1</td>
<td>Being able to eat, dress, bathe, and go to the toilet WITHOUT HELP; AND having NO limitations when playing, going to school, working or in other activities.</td>
</tr>
<tr>
<td>2 R2</td>
<td>Being able to eat, dress, bathe and go to the toilet WITHOUT HELP; AND having SOME limitations when working, going to school, playing or in other activities.</td>
</tr>
<tr>
<td>3 R3</td>
<td>Being able to eat, dress, bathe and go to the toilet WITHOUT HELP; AND NOT being able to play, attend school or work.</td>
</tr>
<tr>
<td>4 R4</td>
<td>NEEDING HELP to eat, dress, bathe or go to the toilet; AND having SOME limitations when working, going to school, playing or in other activities.</td>
</tr>
<tr>
<td>5 R5</td>
<td>NEEDING HELP to eat, dress, bathe or go to the toilet; AND NOT being able to play, attend school or work.</td>
</tr>
<tr>
<td>X3</td>
<td>social-emotional function: emotional well-being and social activity</td>
</tr>
<tr>
<td>Level x3</td>
<td></td>
</tr>
<tr>
<td>1 S1</td>
<td>Being happy and relaxed most or all of the time, AND having an average number of friends and contacts with others.</td>
</tr>
<tr>
<td>2 S2</td>
<td>Being happy and relaxed most or all of the time, AND having very few friends and little contact with others.</td>
</tr>
<tr>
<td>3 S3</td>
<td>Being anxious or depressed some or a good bit of time, AND having an average number of friends and contacts with others.</td>
</tr>
<tr>
<td>4 S4</td>
<td>Being anxious or depressed some or a good bit of time, AND having very few friends and little contact with others.</td>
</tr>
</tbody>
</table>
Table 2 (continued)

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>X4</td>
<td>health problem$^b$</td>
</tr>
<tr>
<td>Level $x_4$</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>H1 Having no health problem.</td>
</tr>
<tr>
<td>2</td>
<td>H2 Having a minor physical deformity or disfigurement such as scars on the face.</td>
</tr>
<tr>
<td>3</td>
<td>H3 Needing a hearing aid.</td>
</tr>
<tr>
<td>4</td>
<td>H4 Having a medical problem which causes pain or discomfort for a few days in a row every two months.</td>
</tr>
<tr>
<td>5</td>
<td>H5 Needing to go to a special school because of trouble learning or remembering things.</td>
</tr>
<tr>
<td>6</td>
<td>H6 Having trouble seeing even when wearing glasses.</td>
</tr>
<tr>
<td>7</td>
<td>H7 Having trouble being understood by others.</td>
</tr>
<tr>
<td>8</td>
<td>H8 Being blind OR deaf OR not able to speak.</td>
</tr>
</tbody>
</table>

$^a$Multiple choices within each description are applied to individuals as appropriate for their age. For example, a 3-year-old child is not expected to be able to get around the community without help from another person.

$^b$Individuals with more than one health problem are classified according to the problem they consider the most serious.

3.1.3. Measurement

 Generally the most accurate way to obtain utility values for your own study is to measure them yourself. This involves the identification of health states for which utilities are required, the preparation of health state descriptions, the selection of subjects, and the use of a utility measurement instrument. These steps are described in detail below.

3.2. Health state descriptions

 Each unique possible health outcome for the programme under evaluation and for the comparison programme must be defined as a health state for utility measurement. Depending on the study there may be only a few health states or there may be many hundreds. For example, a study of kidney dialysis and transplantation required only four states because there were only four distinct health outcomes - kidney transplant, hospital dialysis, home dialysis and death [Torrance et al. (1973)]. On the other hand, a study of neonatal intensive care required 960 states because there was a vast array of possible outcomes [Torrance et al. (1982), Boyle et al. (1983)].

 The health states should be described in functional, as opposed to clinical, terms. That is, the description should focus on how well or poorly a person in this health state is able to function, rather than on his clinical diagnosis or his laboratory test results. A comprehensive description should include a statement on the level of physical functioning, the level of emotional functioning and the level of social functioning. In addition, other aspects are sometimes added to the health state description that are important for a
utility judgment but do not fit neatly into the physical, emotional or social attributes. For example, Torrance et al. (1982) included the ‘health problem’ (e.g., blind) in their descriptions.

The utility for a health state is affected by its duration [Torrance et al. (1972), Sackett and Torrance (1978)]. Thus, it is important to specify the duration of the health state either as part of the description itself or as part of the measurement process.

The utility of a health state should be unconfounded by utilities for other states that may or may not follow this one. Thus, it is important that the prognosis for the health state 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. In measuring the utility of a chronic condition, the prognosis should be stated as no change other than the normal aging process until death with the age of death specified. In the case of a temporary condition, it should be no change until the end of the temporary duration specified at which point the person returns to normal health.

There are three possibilities for describing a health state to a subject. The first is to use the subject’s own health as the health state to be measured, the second is to use a holistic description of the state and the third is to use a health state classification system.

In some applications the relevant health states for which utilities are required are simply the health states of the patients in the study and, furthermore, the patients are available and appropriate as subjects for the utility measurement task. In that case each patient can provide a utility measure for his own health state, and 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 patients’ health state may still be required so that others can interpret the results.

Using patients to measure the utility for their own health states has not been widely undertaken in the past, but appears to be on the increase. Churchill and colleagues have been studying the quality of life in end-stage renal disease using this approach. They have reported the results of their initial pilot work [Churchill et al. (1984b)], and are now involved in a much larger study with many more patients. The approach seems particularly well suited for use in clinical trials. Here, the quality of life, as measured by the utility score, can be determined on each patient in each of the experimental and the control groups at baseline and at each follow-up point. In addition, or alternatively, utility change scores can be measured directly at follow-up by asking patients to compare their condition now with their condition on entry to the study. This latter approach is currently being used in a large
international multi-centre double-blind randomized clinical trial for a new arthritis drug. The intention is to evaluate the drug, in part, on the basis of a cost-utility analysis. In addition, a number of other studies are either underway or planned at McMaster to measure the utility of patients under treatment for spinal cord injury, cancer, chronic obstructive lung disease, heart attack, and genetic counselling using chorion biopsy or amniocentesis. We see a considerable future for this approach as economic appraisals are increasingly incorporated directly into clinical trials.

### 3.2.1. Holistic health state descriptions

Frequently in programme appraisal patients are not used to measure utilities. Rather, subjects who are not in the health state are asked 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. For example, in our work we have always instructed the subject to consider that the health state applies to himself [Sackett and Torrance (1978), Torrance et al. (1982)], while some other investigators have specified to the subject that the condition applies to someone else [Pauker et al. (1981), Rosser and Kind (1978)]. 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. Some investigators have used low fidelity descriptions that include only a few key words or phrases for each of the major characteristics of the health state [Patrick et al. (1973), Kaplan et al. (1976)]. Others have used medium fidelity descriptions consisting of a few paragraphs of narrative [Torrance et al. (1973), Sackett and Torrance (1978)]. Still others have investigated the use of high fidelity descriptions including video tapes [Cadman and Goldsmith (1982)] and audio tapes [McNeil et al. (1981), Boyd et al. (1982)]. Comparisons among these different approaches suggests that sometimes the utility values differ depending upon the level of detail (Boyd) and sometimes they do not (Cadman). Furthermore, although the longer more comprehensive descriptions may appear to have greater face validity, they may, in fact, simply overload the cognitive abilities of the subject so that he or she merely latches onto a few key phrases and ignores the rest. Finally, it is well known that the way in which a health state is described and the way in which the question is framed can systematically bias the result [Kahneman and Tversky (1982), Hershey et al. (1982)]. Given the uncertainties discussed above the best current advice in measuring utilities on the general public is probably 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.
3.2.2. Health state classification system

An alternative approach to describing each specific health state of interest is to define a health state classification system that encompasses all states of interest. One example of such a system is shown in table 1, from Torrance et al. (1982); another example can be found in Kaplan et al. (1976), while a general review of this approach is available in Boyle and Torrance (1984).

The general idea of a multi-attribute health state classification system is shown in fig. 2. It is based on the concept that health status can be defined in terms of a number of attributes, possibly hierarchically nested. In theory, the system of attributes can be as broad and as deep as necessary in order to encompass the level of detail required in the study. The lowest level attributes in the hierarchical structure are divided into levels that represent step-wise decrements in function on that particular attribute. For example, attribute ‘dressing’ might be divided into the three levels: (i) able to dress oneself 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 contains an enormous number of health states. To measure the utility of each of these states one at a time would be an infeasibly large task. However, if the attributes satisfy certain independence properties, multi-attribute utility theory [Keeney and Raiffa (1976)] can be used to reduce dramatically the amount of work. An example of this approach has been reported by Torrance et al. (1982).

![Fig. 2. Example of attributes for a health state classification system.](image-url)
3.3. Subjects

The subjects are the individuals whose utilities are to be measured. Various studies have used various types of individuals for this purpose. A random sample of the general public has often been used [Kaplan et al. (1976), Sackett and Torrance (1978)] on the grounds that it is society's resources that are being allocated to the various health care programmes and so it is society's preferences that should count. Studies have also investigated patients' preferences on the argument that who better can appreciate the true implications of a particular health state than a person who has first hand knowledge [Sackett and Torrance (1978), Wolfson et al. (1982), Churchill et al. (1984b)]. Others have investigated the preferences of health professionals and health leaders on the grounds that they are the most knowledgeable and/or they are the appointed proxies of the public in the health field [Bush et al. (1973), Wolfson et al. (1982)].

So whom should you ask? The answer can be determined, in part, from the purpose and the viewpoint of the study. Most cost-utility analyses are undertaken to influence public policy decisions and, accordingly, are conducted from the societal viewpoint. In this case, the appropriate utilities are those of an informed member of the general public or community representative. Informed means that the subject truly understands what the health state is like. This is the sticking point. How do you describe, in a complete and yet unbiased manner, a particular dysfunctional health state (for example, kidney dialysis) to a healthy individual who has no experience with the condition? And how do you know when you have done it right? Issues to be considered here include the style and detail of the health state description (discussed earlier), the reliability and validity of different procedures (discussed later) and the extent of differences in measurements made on different types of subjects such as patients, physicians and the general public (discussed below).

Patients are appropriate subjects to ask regarding the utility of their condition in clinical trials as described earlier. This is particularly true when the focus is on comparisons of alternatives within the trial, as it generally is, rather than comparisons beyond the trial. That is, a trial comparing two methods of treating arthritis can properly use patients' utilities as part of the comparison and part of the recommendation. Similarly, a different trial comparing two methods of treating scoliosis could also use patients' utilities. However, it would be dangerous for a policy maker to compare the utilities of the best arthritis treatment with the best scoliosis treatment in making a resource allocation decision. The problem here is that the utilities have been measured on two, possibly quite different, groups. Certainly in this example, the age–sex profile of the two groups of patients would be quite different. Moreover, at least in theory, patients of a particular disorder have an incentive to exaggerate the disutility of their condition in order to enhance
the cost–utility of preventive and treatment programmes aimed at the disorder. On the other hand, the limited evidence available to date [Sackett and Torrance (1978), Wolfson et al. (1982)] suggests that this does not happen.

Health professionals, such as physicians and nurses, have also been used as the source of health state utilities. This has many of the same advantages and disadvantages as the use of patients. It minimizes the problems of describing the states, but at the expense of possible bias due to conflict of interest and due to the special age, sex, and socio-economic status of health professionals.

Who you should ask is only an issue if different groups are known to give different results. With interval scale cardinal utilities this has generally not been the case. Most investigations have found no difference among different groups – age, sex, socio-economic status, ethnic background, religious affiliation, general public, physicians, nurses, patients [Kaplan and Bush (1982), Sackett and Torrance (1978), Wolfson et al. (1982)]; a few have found small differences [Sackett and Torrance (1978)]; none have found large differences.

3.4. Utility scales

Measurement scales are classified, from weakest to strongest, as categorical (e.g., religion, ethnicity), ordinal (e.g., finishing ‘place’ in a competition), interval (e.g., temperature on °F or °C) and ratio (e.g., weight, length), with the last two also being called cardinal. 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 intervals (differences between scale values) can be compared in a meaningful fashion but ratios of scale values cannot. For example, if $A$ is 5°C ($41^\circ$F), $B$ is 10°C ($50^\circ$F) and $C$ is 20°C ($68^\circ$F), it is correct to state that the difference between $A$ and $C$ is one and one half times as great as the difference between $B$ and $C$ and three times as great as the difference between $A$ and $B$, but it is incorrect to say that $B$ is twice as great as $A$. 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 meaningfully.

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 simplest to obtain and the least demanding in terms of their underlying assumptions, and therefore are the preferred measure where they are adequate. Unfortunately, they are seldom 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 upon 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 in detail in this paper produce interval scales of utility. One method [Rosser and Kind (1978)], described briefly later in the paper, produces a ratio scale.

A cardinal preference measure is sometimes referred to as ‘value’ and often referred to as ‘utility’. Technically there is a difference between value and utility, depending upon the measurement techniques used and whether or not it incorporated uncertainty [Keeney and Raiffa (1976)]. However, at this stage in the development of applications in health the distinction has not proven to be very important.

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 arbitrarily 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 Arrow (1963) 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 led to different results: for example, Hildreth (1953), Harsanyi (1975) and Keeney (1976) all agree that aggregation is valid, while Kalai and Schmeidler (1977) argue the converse. 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 within the health domain; that is, each individual’s health is counted equally. The mean is the calculation method recommended by Harsanyi and is also the method obtained from the models of Hildreth or Keeney by assigning each individual equal weight.

3.5. 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 earlier each description of a health state should be functionally oriented and comprehensive. Further, the description of the state or the measurement process should specify the age of onset for 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 [Arrow (1963)]. They are also unsuitable for cost-utility analysis. Thus, applications in programme evaluation make use of cardinal measures.

Cardinal preferences can be measured by a number of different techniques. Moreover, the application of some of the techniques differs depending upon whether the health state is chronic or temporary. Three major techniques — rating scale, standard gamble and time trade-off — are described in some detail. Several other techniques — equivalence, ratio scaling — are covered more briefly.

3.5.1. Rating scale

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. Variations on this procedure include category scaling in which a specified number of categories (for example 11 categories) are used. The most preferred health state is placed in category 1, the least preferred in category 11, and the others are placed in the category that seems most appropriate to the subject in order to reflect the strength of the subject’s preference, assuming equal change in preference between adjacent categories. Other variations include visual aids such as a thermometer with a scale from 0 to 100 on a felt background with foam sticks labelled with the health states. The subject places the foam sticks on the scale and adjusts their location until he feels confident that they reflect his true feelings [Torrance et al. (1982)]. In all cases with a rating scale it is important to
stress to the subject that the distances between the placements of health states should correspond with his feelings about the relative differences in preference between the health states. This is important in order to capture the essential nature of an interval scale.

Preferences for chronic states can be measured on a rating scale. The chronic states are described to the subject as irreversible; that is, they are to be considered permanent from age of onset until death. The subject must be provided with the age of onset and the age of death, and these should be the same for all states that are measured together relative to each other in one batch. States with different ages of onset and/or ages of death can be handled by using multiple batches. Two additional chronic states are added to each batch as reference states for the scale - 'healthy (from age of onset to age of death)' and 'death at age of onset'.

The subject is asked to select the best health state of the batch, which presumably would be 'normal healthy life' and the worst state, which may or may not be 'death at age of onset'. He is then asked to locate the other states on the rating scale relative to each other such that the distances between the locations have the interval property described earlier. The rating scale is measured between 0 at one end and 1 at the other end. If 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. If death is not judged to be the worst state but is placed at some intermediate point on the scale, say $d$, the preference values for the states are given by the formula $(x - d)/(1 - d)$, where $x$ is the scale placement of the health state.

Preferences for temporary health states can also be measured on a rating scale. Temporary states are described to the subject as lasting for a specified duration of time at the end of which the person returns to normal health. As with the chronic states, temporary states of the same duration and same age of onset should be batched together for measurement. Each batch should have one additional state 'healthy' added to it. The subject is then asked to place the best state (healthy) at one end of the scale and the worst temporary state at the other end. The remaining temporary states are located on the scale using the interval scaling principle.

If the programmes being evaluated involve only morbidity and not mortality and if there is no need to compare the findings to programs that do involve mortality, the procedure described above for temporary health states is sufficient. However, if this is not the case the interval preference values for the temporary states must be transformed onto the standard 0–1 health preference scale. This can be done by redefining the worst temporary health state as a chronic state of the same duration, and measuring its preference value by the technique described for chronic states. The values for the other temporary health states can then be transformed onto the standard
3.5.2. Standard gamble

The standard gamble is the classical method of measuring cardinal preferences. It is based directly on the fundamental axioms of utility theory, first presented by von Neumann and Morgenstern (1953). The method has been used extensively in the field of decision analysis, and good descriptions of the method are available in books from this field [for example, see Holloway (1979)].

The method can be used to measure preferences for chronic states but the method varies somewhat depending upon whether or not the chronic state is preferred to death. For chronic states preferred to death the method is displayed in fig. 3. The subject is offered two alternatives. Alternative 1 is a treatment with two possible outcomes: either the patient is returned to normal health and lives for an additional $t$ years (probability $p$), or the patient dies immediately (probability $1-p$). Alternative 2 has the certain outcome of chronic state $i$ for life ($t$ years). Probability $p$ is varied until the respondent is indifferent between the two alternatives, at which point the required preference value for state $i$ is simply $p$; that is, $h_i = p$.

Since most subjects cannot readily relate to probabilities, the standard gamble method is often supplemented with the use of visual aids, particularly a probability wheel [Torrance (1976)]. This is an adjustable disk with two sectors, each a different colour, and constructed so that the relative size of the two sectors can be readily changed. The alternatives are displayed to the subject on cards, and the two outcomes of the gamble alternative are colour-keyed to the two sectors of the probability wheel. The subject is told that the chance of each outcome is proportional to the similarly coloured area of the disk.

![Fig. 3. Standard gamble for a chronic health state preferred to death.](image-url)
Fig. 3 shows the basic format of the standard gamble method. However, variations on this basic format are possible. For example, it is not essential that the two outcomes of the gamble alternative be healthy and dead. As long as the two outcomes of the gamble alternative are such that one is preferred to the outcome of the certain alternative and the other is dispreferred to it, the standard gamble method can be used to relate the preference values of the three health states. By application of this principle one can develop various ways to present the questions to the subject in order to determine the preference values. Two systematic variations using this approach have been investigated by Boyd et al. (1982). In one variation they consistently related the health state to be measured in the certain alternative to the one immediately preferred to it and to death. In the other variation they related the health state to be measured in the certain alternative to the one immediately dispreferred to it and to healthy. The latter variation has also been used by Torrance and colleagues [Torrance et al. (1972), Torrance (1976)]. Other variations are also possible but one must be careful to ensure that each health state is measured in a way such that all health state values can be calculated. The principle to keep in mind in determining whether or not to use variations like these is the fact that subjects can best respond to probability questions when the probabilities are in the mid-range, preferably near 0.5, whereas they have difficulty in properly responding to questions with probabilities that are extremely small or large. Findings suggest that probabilities less than 0.1 or greater than 0.9 are difficult for subjects to use [Kahneman and Tversky (1982)].

For chronic states considered worse than death the standard gamble technique must be modified slightly as shown in fig. 4. Here the uncertain alternative leads to outcomes healthy or state 1 with probabilities p and 1−p, whereas the certain outcome is death. One way to present this choice is to ask the subject to imagine that he has a rapidly progressing terminal disease which if left unattended will lead rapidly to death. However, there is a treatment with probability p of curing the subject and probability 1−p of leaving the subject irreversibly in chronic state 1. The probability p is varied until the subject is indifferent between the two alternatives, at which point the preference value for state 1 is given by $h_1 = \frac{-p}{1-p}$.

Preferences for temporary health states can be measured relative to each other using the standard gamble method as shown in fig. 5. Here intermediate states i are measured relative to the best state (healthy) and the worst state (temporary state j). Once again other variations on this basic format can be used following the principle that the certain alternative must be a state whose preference is between that of the two outcomes of the uncertain alternative. In the basic format shown in fig. 5 the formula is $h_i = p + (1-p)h_j$ where i is the state being measured and j is the worst state. If death is not a consideration in the use of the utilities, $h_j$ can be set equal to
zero and the $h_i$ values determined from the formula which then reduces to $h_i = p$. However, if it is desired to relate these values to the 0–1 dead–healthy scale, the worse of the temporary states (state $j$) must be re-defined as a short duration chronic state (followed by death) and measured on the 0–1 scale by the technique described above for chronic states. This gives the value for $h_j$ which can then, in turn, be used in the above formula to find the values for the $h_i$.

3.5.3. Time trade-off

The time trade-off method was developed specifically for use in health care by Torrance et al. (1972). It has been validated against the standard gamble method and has been found to give similar results, although the validation has only been carried out for states preferred to death [Torrance et al. (1973), Torrance (1976)]. It is an implicit technique like the standard gamble – that is, unlike the rating scale in which the subject provides his preference
values explicitly, the time trade-off and the standard gamble methods derive the preference values implicitly based on the subject’s responses to decision situations. However, compared to the standard gamble, the time trade-off method has the advantage of being simpler to use.

The application of the time trade-off technique to a chronic state considered better than death is shown in fig. 6. The subject is offered two alternatives – alternative 1: state $i$ for time $t$ (life expectancy of an individual with the chronic condition) followed by death; and alternative 2: healthy for time $x < t$ followed by death. Time $x$ is varied until the respondent is indifferent between the two alternatives, at which point the required preference value for state $i$ is given by $h_i = x/t$.

![Fig. 6. Time trade-off for a chronic health state preferred to death.](image)

If the chronic state is considered worse than death, the time trade-off technique can be applied as shown in fig. 7. Here again the subject is offered two alternatives – alternative 1: healthy for time $x < t$ followed by state $i$ until time $t$, followed by death; and alternative 2: immediate death. Time $x$ is varied until the respondent is indifferent between the two alternatives at which point the required preference value for state $i$ is given by $h_i = x/(x-t)$. This equation is derived by equating the two alternatives, $1.0x + h_i(t-x) = 0$, and solving for $h_i$. One difficulty which has arisen in practice is that although the method imposes an upper limit of 1.0 on states preferred to death, it imposes no comparable lower limit on states dispreferred to death. A solution which has been used tentatively to overcome this difficulty is to scale the preference values of states worse than death such that the worst possible state is assigned a preference value of $-1.0$ [Torrance et al. (1982)]. Further work on the measurement of utilities for states worse than death is underway [Torrance (1984)].

Preferences for temporary health states can be measured relative to each other using the time trade-off method as shown in fig. 8. As with the rating scale and the standard gamble, intermediate states $i$ are measured relative to the best state (healthy) and the worst state (temporary state $j$). The subject is offered two alternatives – alternative 1: temporary state $i$ for time $t$ (the time...
duration specified for the temporary states), followed by healthy: and alternative 2: temporary state j for time \( x < t \), followed by healthy. Time \( x \) is varied until the respondent is indifferent between the two alternatives, at which point the required preference value for state \( i \) is \( h_i = 1 - (1 - h_j)x/t \). If we set \( h_j = 0 \), this reduces to \( h_i = 1 - x/t \). Fig. 8 shows the basic format, but other variations are possible. State \( j \) need not be the worst state as long as it is any state considered worse than \( i \). In using variations, however, care must be taken to ensure that all preference values can be calculated. In one systematic variation which has been used [Torrance et al. (1972), Torrance (1976), Sackett and Torrance (1978)], state \( j \) is always the next worst state to state \( i \). Finally, as with the rating scale and the standard gamble, if the preference values for the temporary states are to be transformed to the 0–1 dead–healthy scale, the worst of the temporary states must be re-defined as a short duration chronic state and measured by the method for chronic states described above.
3.5.4. Other techniques

An equivalence technique has been used occasionally to measure cardinal preferences for health states [Berg (1973), Bush et al. (1973)]. In this technique the subject is asked the following kind of question: 'If there are x people in adverse health situation A and y people in adverse health situation B, and if you can only help (cure) one group (for example, due to limited time or limited resources), which group would you choose to help?' One of the numbers x or y can then be varied until the subject finds the two groups equivalent in terms of needing or deserving help. If x and y are the equivalent numbers as judged by the subject, the undesirability (disutility) of condition B is \( \frac{x}{y} \) times as great as that of condition A. By asking a series of such questions all conditions can be related to each other on the undesirability scale.

Ratio scaling has been used in one study [Rosser and Rind (1978)]. Here the subjects were asked to provide the ratio of undesirability of pairs of health states – for example, is one state two time worse, three time worse, etc. compared to the other state? Then, if state B is judged to be x times worse than state A, the undesirability (disutility) of state B is x times as great as that of state A. By asking a series of questions all states can be related to each other on the undesirability scale. One can see that the ratio method of Rosser and Rind and the equivalence method are very similar. Indeed, to help their subjects understand the question, Rosser and Rind described two implications of the subject’s ratio assignments, and the second implication is precisely the equivalence method.

3.5.5. Multi-attribute utility theory

The measurement techniques covered above have all been described in terms of how they are used to measure single health states. If the health states are described by a classification system of the type appropriate for multi-attribute utility theory methods, the approach is more complicated. An appropriate multi-attribute utility model (additive, multiplicative, etc.) must be determined, and this in turn specifies the utility measurements which need to be made [Torrance et al. (1982)]. However, the fundamental utility measurements are then made using the basic measurement techniques described above. Thus, the multi-attribute utility theory approach does not introduce any new methods of measuring preferences. Rather, it introduces a way of selecting specific preferences to be measured and combining them into a mathematical model of the subject’s utility structure. The model can then be used to determine the utility values for all possible health states in the classification system.

3.6. Accuracy

The utility values are valid if (i) the subjects are appropriate, (ii) the health
state descriptions are adequate to properly describe the states and neutral in their influence on the measurement, (iii) the measurement questions are framed in a balanced or neutral way, and (iv) the measurement technique itself is reliable and valid. The first three of these points were discussed earlier. The fourth issue of reliability and validity of the measurement techniques, is discussed here.

A measurement technique is reliable if it is consistent – if the same phenomenon can be measured a second time with identical results. Internal reliability refers to a second measurement taken as part of the original interview. The coefficient of internal reliability (as measured by the product moment correlation coefficient $r$) for the rating scale ranges from 0.86 to 0.94 [Torrance et al. (1982)], for the standard gamble 0.77 to 0.92 [Torrance (1976)], and for the time trade-off 0.77 to 0.88 [Torrance (1976), Torrance et al. (1982)]. These would all be judged acceptable. Test–retest reliability refers to a second measurement taken sometime later. Test–retest reliabilities for the time trade-off method have been reported to range from 0.63 to 0.80 for a six-week retest [Churchill et al. (1984b)], and to be 0.81 for a four-week retest [Churchill et al. (1985)]. The recent experience with 0.81 is certainly quite acceptable. On the other hand, one-year test–retest r's have been poor for all techniques, ranging from 0.49 to 0.62 [Torrance (1976)]. However, this may simply indicate, at least in part, that people's preferences shift over time. On balance, then, the work to date suggests that over short intervals of time the reliability of the measurement instruments is satisfactory.

Reliability can also be expressed in terms of the precision $\sigma_e$ of an individual measurement, where $\sigma_e$ is the standard deviation of the measurement error. For the rating scale $\sigma_e$ values have been reported from 0.09 to 0.15 [Torrance et al. (1982)], while values for the time trade-off and standard gamble are 0.13 [Torrance (1976)]. This means that if an individual responds to a standard gamble or a time trade-off question with a utility of 0.60 for a particular health state, the 95% confidence interval would be 0.34–0.86, a rather wide range. Thus, single individual measurements are not particularly precise.

Moreover, individuals differ greatly in their health state preferences and the differences cannot be explained by the usual demographic characteristics such as age, sex, socio-economic status, religion, illness, occupation, etc. For example, individual differences among the general public for the same health state on the 0–1 utility scale result in a standard deviation of scores of approximately 0.30 [Sackett and Torrance (1978)]. On the other hand, subjects who are more homogeneous and more knowledgeable about the health state seem to differ less. For example, in the study reported by Torrance (1976) 29 home dialysis patients rating the home dialysis scenario with the time trade-off method resulted in a standard deviation of 0.18 compared to 0.28 for the general public. [Torrance (unpublished)].
Fortunately, the imprecision of individual measurements and the considerable differences among individuals can be ameliorated by taking the mean of a large group of subjects. Since the standard error of the mean is $\sigma/\sqrt{N}$, the mean utility value for a health state can be made as precise as desired by increasing the group size $N$. Moreover, group mean values have been found to be remarkably stable regardless of the make-up of the group [Sackett and Torrance (1978), Boyd et al. (1982), Wolfson et al. (1982)]. These are fortunate findings since group mean utilities are the values normally required in programme evaluation.

A measure is valid if it accurately reflects the concept or phenomenon it claims to measure. Two approaches have been taken here. In one approach, health state utilities are claimed to be utilities obeying the axioms of von Neumann–Morgenstern utility theory for decisions under uncertainty [von Neumann and Morgenstern (1953)]. In this case, the standard gamble measurement technique is valid by definition and the validity of the other techniques can be determined by comparison. In one study using this approach, the time trade-off technique was found to be relatively valid while the rating scale method was not [Torrance (1976)]. In the other approach to validity, health state utilities are claimed to measure the overall quality of life associated with the health state and thus should be strongly correlated with other trusted measures of health-related quality of life. Using this approach, Churchill et al. (1984b, 1985) determined that patients' measured utilities correlated significantly with nephrologists' ratings of the patients' quality of life. Moreover, groups of patients with different clinical status differed in their measured health state utilities in predictable ways.

Given the uncertainties still surrounding the reliability, validity and precision of utility values, it is important to perform sensitivity analysis on them. If wide changes in the utility values have no impact on the study's conclusions, as was the case in one recent study [Boyle et al. (1983)], nothing further need be done. If, on the other hand, the conclusions of the study are found to be sensitive to the utility values for specific health states, these values should be remeasured on additional subjects, perhaps with other techniques, to improve the confidence in the overall result.

4. Conclusions

Health state utilities can be measured, albeit with some imprecision. The measures can be used in the economic appraisal of health care programmes through the application of cost–utility analysis. Imprecision in the measures can be, to some extent, ameliorated through larger sample sizes and by the careful use of sensitivity analysis.

The approach represents an important tool in the kit bag of economic evaluation. In health care programmes where the quality of life associated
with the outcomes is a significant consideration, health state utilities provide a method to incorporate formally these considerations into the analysis. To omit consideration of quality of life in such circumstances because of measurement difficulties would be quite inappropriate.

The methods of measuring health state utilities are being improved continually as we gain more experience and more understanding of them. We commend the approach to health care analysts everywhere, not only for the direct benefits in their particular study, but also for the increased knowledge and improved techniques that come from further shared experiences.

References


Arrow, K.J., 1963, Social choice and individual values, 2nd ed. (Yale University Press, New Haven, CT).

Berg, Robert L., 1973, Establishing the values of various conditions of life for a health status index, in: Robert L. Berg, ed., Health status indexes (Hospital Research and Education Trust, Chicago, IL), 120–134.


Boyle, M.H. and G.W. Torrance, 1984, Developing multiattribute health indexes, Medical Care 22, no. 11, 1045–1057.


Bush, J.W., S. Fanshel and M.M. Chen, 1972, Analysis of a tuberculin testing program using a health status index, Socio-Economic Planning Sciences 6, 49–68.


Drèze, J., 1962, L'utilité sociale d'une vie humaine, Revue Français de Recherche Opérationelle 6, 93–118.


Harris, A. et al., 1971, Handicapped and impaired in Great Britain (HMSO, London).


Torrance, George W., 1970, A generalized cost-effectiveness model for the evaluation of health programs, Research report no. 101 (Faculty of Business, McMaster University, Hamilton).


