helpful, it can be seen within the decision-aiding tradition of economic evaluation (Sugden and Williams 1978). The disadvantages are that the basis for a decision can often be unclear and will not be based on explicit values.

2.3 Measures for describing health

2.3.1 What is health?

Health has been variously defined, but one of the most enduring and influential definitions is the statement in the Constitution of the World Health Organization (1948) that health is: 'A state of complete physical, mental and social well-being, and not merely the absence of disease and infirmity'. While this definition has been highly influential in the development of measures in this field, it is very broad and not easy to operationalize. It could be argued, for example, that social well-being is not health *per se*, but an aspect of quality of life that is affected by health. Some developers have chosen to take a much narrower definition of health in the construction of their measure. As a consequence, measures of health or health-related quality of life often differ considerably in their content. To add to the potential for confusion, some measures claim to measure quality of life, and yet their content is narrowly focused around symptoms.

In the health literature terms such as health, health status, health related quality of life and quality of life are used to mean different things by different instrument developers. In a sense terminology in itself does not matter, but underlying this issue of semantics is an important implication for policy in terms of what should be counted as a benefit and what might be ignored for the purpose of informing health policy. Whether or not social activites are counted as health, health related quality of life or quality of life does not matter, the important question is whether impact on social activities should form the part of the description of the benefits of health care. benefit Chapter 4 addresses this question in some detail.

An important feature of health is that it includes elements (such as pain, affect or various symptoms) that are experienced by an individual but difficult to measure with any external instrument or test that is valid, reliable and comparable. There is evidence of significant differences between what professionals report on a patient and what the patients say themselves (Jachuck *et al.* 1982). It is increasingly recognized in clinical and health services research that descriptions of the experience of a health state should be elicited from the patients themselves or, where this is not possible, from others on their behalf, in order best to reflect the actual experience of the disease and its treatment (Fitzpatrick *et al.* 1998).

2.3.2 What are self-reported measures of health?

A common approach to describing an individual's health is to elicit self-reports of levels on various health dimensions using standardized numerical scoring systems. This approach aims to provide quantitative descriptions of health states in terms of the components of health regarded as most relevant to patients with health problems, caused either by disease, the treatment of disease or other processes such as natural ageing, trauma and pregnancy. The domains included in most self-reported health measures do not include biomedical measures (such as blood pressure, forced expiratory volume, cholesterol levels,etc.) or relate to specific diagnostic instruments used in clinical practice. While both of these are important in the clinical management of a patient and potentially important in predicting a person's future health, they are less pertinent to understanding the health experience of a patient.

Self-reported measures of health have been available since the 1940s (Karnofsky and Burchenal 1949), but did not become widely used until the 1960s and 1970s. However, by 1987, there were over 200 measures of health identified by Spilker and colleagues (1990). They can assess symptoms, function or well-being depending on the specific definition of health being used, and can be 'generic' and hence designed for use across all conditions, or specifically designed for a particular disease. See Chapter 4 for a further discussion on this topic.

Patient-perceived measures of health are increasingly used to assess the efficacy and effectiveness of health care interventions. Some researchers have attempted to use them in conducting economic evaluations alongside clinical trials (e.g. Buxton *et al.* 1985; Nichol *et al.* 1992; Brazier *et al.* 1999) but they were not designed to be used in this way. The use of these measures in economic evaluation has been criticized by health economists, largely because they do not have the necessary interval properties to undertake CEA. However, such measures will continue to be widely used in clinical trials and other studies due to their popularity with clinicians and health services researchers who are seeking to gather evidence for other purposes. It is therefore important to examine the potential use of health status measures in economic evaluation, if only from an opportunistic viewpoint. More importantly, this provides an important background to the reasons why measures have been specially developed for use in economic evaluation.

2.3.3 The content of instruments

Measures of health vary widely in terms of content, format and scaling. The principal features of a sample of five measures are presented in Table 2.1.

Table 2.1 Characteristics of five health status measures (from Brazier *et al.* 1999)

Questionnaire	No. of dimensions	Description of dimensions/items	No. of items	Source of responses	Method of administration	Source of values	Results
Condition specific							
St George's Respiratory Questionnaire	4	Symptoms (e.g. shortness of breath and wheezing), activity (e g. walking and playing games), impacts (e.g.embarrassment)	50	Patient	Interview or self-completion	Patients (using VAS)	Profile and index
Chronic Respiratory Questionnaire	4	Dyspnoea, fatigue, emotional function, mastery	20	Patient	Interview	Assumed	Profile
Barthel	1	Mobility, grooming, dressing, continence	10	Professional	Professional assessment	Assumed	Index
Generic							
SF-36 health survey	8	Physical functioning, role limitations (physical and emotional problems), social functioning, pain, mental health, general health perception	36	Patient or proxy	Self-completion, interviewer administration	Assumed	Profile
Nottingham Health Profile	6	Mobility, social isolation, pain, emotional reactions, energy	38	Patient	Self-completion	Thurstone's method	Profile

The instruments have been selected to demonstrate the diversity of measures in terms of their size, coverage of health domains, method of administration and sources of values, not for being typical or even representative.

The content varies considerably between the measures. They cover generic concepts of functioning (such as physical functioning), through to specific symptoms (e.g. dyspnoea for respiratory disease, dexterity for arthritis, and so forth). The methods of completing the questionnaires include clinical interview, professional assessment, researcher interview and self-completion, either in the clinic or at home. While most of these measures are self-reported, we note that the Barthel Index is an exeption, being completed by the health professional. The Chronic Respiratory Questionnaire incorporates a further development, where patients are asked to identify the important activities which make them breathless, as well as providing the assessment. The developers have argued that this approach has the advantage of generating a score more responsive to health change (Guyatt *et al.* 1987), though it is of doubtful use in interpersonal comparisons. Responses to items are combined into either a single index or a profile of several subindices of scores using a scoring algorithm.

Most measures of health have a simple 'summative' scoring system. The SF-36 health survey, for example, is a standardized questionnaire used to assess patient health across eight dimensions (Ware *et al.* 1993). It consists of items or questions which present respondents with choices about their perception of their health (see Table 2.2). The physical functioning dimension, for example, has 10 items to which the patient can make one of three responses: 'limited a lot', 'limited a little' or 'not limited at all'. These responses are coded 1, 2 and 3, respectively, and the 10 coded responses summed to produce a score from 10 to 30. The same procedure is used for all eight dimensions. These raw dimension scores are transformed linearly onto a 0–100 scale. The eight dimension scores of the SF-36 are not comparable across dimensions. This procedure has been mistakenly described in the psychometric literature as being 'unweighted' (Jenkinson 1991), yet implies an *equal* weighting.

Scoring can be more sophisticated, such as the use of statistical techniques like factor analysis for the SF-36 (Ware *et al.* 1993). More recently, Rasch models have been suggested to take into account degree of difficulty, or severity, of an item in relation to the underlying, unobserved (latent) scale that the item is presumed to measure. Some instruments, such as the St George's Respiratory questionnaire and the Sickness Impact Profile, weight items using explicit valuation procedures. This involves asking people to rate the importance of each item using a visual analogue scale (VAS). Other instruments ask the patients to record how bothersome they find the attribute described by the item, and this is then used to weight the item.

Table 2.2 An example of a measure of health status: the SF-36 health survey (from Brazier et al. 1999)

Dimension	No. of items	Summary of content	No. of response choices	Range of response choice
Physical functioning	10	Extent to which health limits physical activities such as self-care, walking, climbing stairs, personal hygiene and moderate and vigorous exercises	3	'Yes, limited a lot' to 'no, not limited at all'
Role limitations-physical	4	Extent to which physical health interferes with work or other daily activities, including accomplishing less than wanted, limitations in the kind of activities, or difficulty in performing activities	2	Yes/No
Bodily pain	2	Intensity of pain and effect of pain on normal work, both inside and outside the home	5 and 6	'None' to 'very severe' and 'not at all' to 'extremely'
General health	5	Personal evaluation of health, including current health, health outlook and resistance to illness	5	'All of the time' to 'none of the time'
Vitality	4	Feeling energetic and full of life versus feeling tired and worn out	6	'All of the time' to 'none of the time'
Social functioning	2	Extent to which physical health or emotional problems interfere with normal social activities	5	'Not at all' to 'extremely' and 'All of the time' to 'none of the time'
Role limitations-emotional	3	Extent to which emotional problems interfere with work or other daily activities, including decreased time spent on activities, accomplishing less and not working as carefully as usual	2	Yes/No
Mental health	5	General mental health, including depression, anxiety, behavioural-emotional control, general positive affect	6	'All of the time' to 'none of the time'

Health Technology Assessment 1998; 2(3) 'A review of the use of Health Status Measures in Economic Evaluation' by Brazier, Deverill, Green, Harper and Booth © Crown copyright material is reproduced with the permission of the controller of HMSO and Queen's Printer for Scotland.

2.4 The use of non-preference-based measures of health in economic evaluation

2.4.1 Criticisms of non-preference-based health status measures

Scoring by dimension

The simple summative scoring algorithms described above have been used by most measures of perceived health. This assumes equal intervals between the response choices and that the items are of equal importance. However, there is no reason to suppose, for example, that patients perceive the intervals of the responses to items of the physical functioning dimension of the SF-36 of 'not limited at all' and 'limited a little' to be equivalent to the interval between 'limited a little' and 'limited a lot'. To take another example from the SF-36, the intervals for an item on how much bodily pain a person has had in the last 4 weeks are 'none' to 'very mild', 'very mild' to 'mild', 'mild' to 'moderate', 'moderate' to 'severe', and 'severe' to 'very severe'. This would imply that in a trial, a reduction in pain from 'mild' to 'very mild' would be equivalent to a reduction from 'severe' to 'moderate'. Yet evidence using visual analogue and standard gamble valuation techniques suggests that people are often unable to perceive a significant difference between 'very mild' and 'mild', but that there is a very large and significant difference between 'moderate' and 'severe' (Brazier *et al.* 1998, 2002).

The summing of scores across items makes equally untenable assumptions about the value people would place on different items. In the physical functioning scale of the SF-36, the item 'limitations in climbing one flight of stairs' is assumed to be of equal importance to 'limitations in walking more than one mile'. For someone living in a bungalow, limitations in walking would probably be regarded as a far worse problem. Given the lack of any empirical basis for these assumptions, there must be doubts about even the ordinal properties of these scales as indicators of people's preferences, particularly over small changes in the dimension scores.

The equal interval assumptions underlying most measures of health have been defended by some researchers. It has been claimed that the relative importance of the different health concepts (as perceived by the instrument developers) is in part taken account of by the number of items used to represent them. It has also been claimed that it makes little difference in practice whether or not equal interval weighting is used (Jenkinson 1991). However, the numerous valuation studies with the EQ-5D, SF-6D and HUI3 have all shown that intervals between response choices are not equal and that items do not have the same weight (see Chapter 8 for a review of these measures). Furthermore, studies have found only a low to moderate correlation between

health status measures and various preference- or value-weighted measures (Brazier *et al.* 1999).

Some instruments have adopted more sophisticated scoring methods using factor analysis or Rasch modelling. Factor analysis weights items according to the extent to which they contribute to some underlying latent variable. The stronger the correlation, the larger the weight of an item or dimension (depending on the unit of analysis). This has been used to re-score the SF-36 dimensions into two summary scores, one for physical health and the other for mental health (Ware *et al.* 1993). The scores have also been transformed so that a score of 50 represents the mean level in the general population and each movement of 10 points from this score represents a standard deviation of the score in the general population. Whilst this scoring system offers a statistical basis for understanding score differences between populations, there is no reason why weights based on correlation between items should reflect their relative importance to people in their daily lives.

More recently there has been interest in using Rasch models to re-score instruments based on item response theory. This was originally developed in education to provide a way of estimating how difficult different questions are against a unidimensional construct, for example numeracy. In health, the analysis will estimate the degree of severity represented by different items, where the underlying construct can be physical functioning, or pain. It is claimed that Rasch models result in a linear interval scale against which any item, regardless of the instrument from which it came, can be calibrated. One example of its application was a re-scoring of the 10 item physical functioning of the SF-36 (Raczek *et al.* 1997).

Whilst Rasch models provide a useful technique for understanding the position of items within a construct, they do not provide an appropriate method for valuing health for economic evaluation (such as in a cost per QALY analysis). The fact that one item is found to be more difficult to do, say against the construct mobility, does not mean it is more or less important in people's lives. While it may represent an improvement on summative scoring and has an important role in constructing measures (this is discussed in the context of constructing preference-based measured in Chapter 4), it does not provide preference-based (or experienced-based) weights needed for cost per QALY analyses.

Over the years, there have been a number of interesting debates between psychometricians and economists. The main source of confusion seems to arise from the distinction between measuring the construct or constructs of health and the value of health. This difference is summarized in Box 2.2. Whilst the dichotomy may seem rather artificial, since measures of health do contain valuations (some items contain a degree of valuation–more on this in

Chapter 4–and all instruments are weighted in some way), it does provide a useful way to understand the difference between the two approaches.

Score profile

Most health status measures present a profile of scores. The generation of a single index score for health has been opposed by many developers of measures of health. The developers of the Nottingham Health Profile, for example, have argued: 'The simple addition of affirmative responses gives misleading results because of the features of pain, social life, emotion, and so on are qualitatively distinct and made up of different facets which cannot have common denominators' (Hunt *et al.* 1986). This view is understandable when the purpose is to derive a *measure* of different aspects of health, but this is not sufficient for use in economic evaluation. To undertake economic evaluation, it will often be necessary to be able to combine the dimensions into an overall indicator of health. In a comparison of surgical and medical management of a condition, for example, one might perform better against one dimension, but worse against another. At the end of a clinical trial it would not be possible to determine which treatment was most effective, let alone whether it was cost-effective. A trade-off needs to be made between dimensions in order to determine effectiveness, and for assessing cost-effectiveness some means of valuing the difference between interventions needs to be found.

Some measures of health combine dimensions to form a single index (e.g. St George's Respiratory questionnaire, the Sickness Impact Profile and the

Barthel Index). As for the aggregation of items, many assume an equal weighting between dimensions (e.g. Barthel), while others combine the items using item weights estimated using valuation techniques such as the VAS (a critique of VAS as a measure of preference is provided in Chapter 5). In addition to the criticisms of these methods of valuing, the scoring systems make an assumption of simple additivity between dimensions, where the value of one dimension is assumed to be unaltered by the level of another dimension. This rules out the prospect of any interaction between dimensions.

Health status and survival

An important limitation of conventional measures of health is that they do not include mortality. One criticism of separating mortality from health status is the same as that made of profiles in general (i.e. how to combine them in order to assess the overall cost-effectiveness of an intervention). Another is that it creates a statistical artefact known as the 'survivor' effect. In a clinical study, a lower survival rate in one arm of the trial can increase its mean health status score(s) compared with the other arms of the trial. This arises because the patients who have died probably have a lower than average health status. Assuming increased survival is regarded as a good thing, then the analyst is without any means for deciding which is the better treatment. There are some statistical solutions to this problem, but they fail to address the central problem that these outcomes need to be combined in some way (Billingham *et al.* 1999). This is a serious limitation of conventional measures of health status, since many health care interventions have consequences for survival and health status.

Time profile of outcomes

The outcome of a treatment is often estimated as the mean difference between health scores before and after the treatment of patients in the trial. A more sophisticated approach to analysing repeated measures is to estimate the health change as the difference between the mean pre-treatment scores and a weighted average of mean scores across the post-treatment assessments, with the weights proportional to the time between each assessment. In other words, it is the 'area under the curve' where levels of health are plotted against time along the horizontal axis (Matthews *et al.* 1990). However, this method of analysis ignores the impact of time on people's preferences over different outcomes.

Uncertainty in outcomes

Outcomes in health care are rarely certain. Even common interventions such as cholecystectomy are associated with a wide dispersion of outcomes, such as in the relief of pain (Nicholl *et al.* 1992) and mortality. Most interventions

come with risks, including mortality in many cases, and numerous complications and side effects. When treatment begins, neither the doctor nor the patient knows the outcome for certain. Conventional analyses of measures of health assume people are risk-neutral (i.e. their decision is unaffected by the degree of uncertainty around the mean value). Yet, in health care, there is evidence that many people are averse to risk (Loomes and McKenzie 1989). Patients may choose a treatment that achieves a lower expected or mean improvement in the health than another, but is associated with less variance (below we consider how this concern also pertains to QALYs)

2.4.2 Limitations to using health status measures in economic evaluation

The usefulness of non-preference-based health status measures in assessing the relative efficiency of interventions depends on the results of the study. In Table 2.3 we present seven scenarios of costs and outcomes in a comparison of a new intervention with the existing one, and consider whether it is possible to assess cost-effectiveness using health status measures.

Table 2.3 Assessing the relative efficiency of two interventions given different cost and health outcome scenarios

Scenario	Cost	Health status measure	Can relative efficiency be evaluated?
1	Lower	Better in at least one dimension and no worse on any other	Yes, by dominance[1]
2	Same	Better in at least one dimension and no worse on any other	Yes[1]
3	Lower	Same across all dimensions	Yes, by cost minimization[1,2]
4	Lower	Better on some dimensions and worse on others	No
5	Same	Better on some dimensions and worse on others	No
6	Higher	Better in at least one dimension and no worse on any other	No
7	Higher	Better on some dimensions and worse on others	No

[1] Assuming the scale at least indicates ordinal preferences over the range being considered (see discussion in text).

[2] It has been argued that cost minimization is rarely achievable, given the uncertainty that exists around most estimates (see discussion in text).

From: © The EuroQol Group. www.euroqol.org.

The first scenario is a case of dominance where new treatment is cheaper *and* better on at least one of the dimensions of the health status measures, while being no worse on any other. In the second scenario, it is also straightforward to assess relative efficiency, since it is simply a question of choosing the treatment with the better health status measure scores since the two have been found to cost the same. In the third scenario, the dimension scores are the same across all dimensions of the health status measure, and so the decision as to whether to adapt the new intervention could be seen as a *cost-minimization analysis*. Even for these three scenarios, however, it is necessary to demonstrate the ordinality of the scale of the health status measure scores in relation to preferences and that there are no differences on other outcomes, such as survival. Furthermore, it has been argued that in practice there will always be uncertainty around outcomes and so it is not possible to be sure that outcomes are equal (Briggs and O'Brien 2001) or that one intervention is always better than another. The handling of uncertainty is considered in Chapter 9, but essentially the best way to handle it would be through probabilistic sensitivity analysis (Claxton *et al.* 2006) and this requires a single measure of outcome. Health status measures would not be appropriate for this type of analysis.

The result is even less straightforward for scenarios 4–7. In scenarios 4 and 5 the new treatment performs better on some dimensions but worse on others, and hence it could have a lower cost per unit of health gain on some dimensions but higher on others. Even where one treatment is apparently more cost-effective across all the health dimensions, care must be taken in the interpretation. A review of the evidence found that health status measures do not possess the interval properties required to undertake such comparisons (Brazier *et al.* 1999). Furthermore, it is the incremental cost-effectiveness ratio that is important for resource allocation purposes. Therefore, where the least cost-effective intervention costs more and yields a higher benefit, then the greater benefit could be worth the extra cost.

For multiple outcomes, one approach is to present the costs and benefits of the alternatives in a CCA. This type of presentation uses an economic framework, but is unlikely to be helpful since scores on health status measures have no obvious intuitive meaning. Score differences cannot be compared between dimensions, nor can health status measures scores be compared with other outcomes, such as survival, or cost. Non-preference-based health status measures cannot be used to assess the efficiency of interventions in such circumstances.

Overall, health status measures have a very limited role in economic evaluation in their current form, if any at all. For this reason, a different class of health status measures has been developed, known variously as preference-based

measures of health or multiattribute utility scales, for calculating QALYs. However, this is not to say that the descriptive data collected by these non-preference-based health measures could not be used in other ways. In Chapter 8 we review a range of methods for using these data in an economic evaluation, including mapping onto preference-based measures and constructing new preference-based measures from them (as has been done with the SF-36). The next section provides an introduction to QALYs and to preference-based measures of health.

2.5 An introduction to quality-adjusted life years

One of the great innovations in the subdiscipline of health economics has been the development of a new method for valuing benefits for use in economic evaluation, namely the QALY. The QALY attempts to value the benefits of health care in terms of a measure that combines the impact on longevity with quality of life into the common numeraire of a year in full health. Health economics has really broken with mainstream economics by using the QALY rather than the traditional WTP approach (see Chapter 3 for a review of theoretical foundation of the QALY and why it has gained prominence over the more conventional monetary approaches used by mainstream economists). This section introduces the reader to the QALY and preference-based health status measures that provide a key ingredient, the quality adjustment.

2.5.1 Basic description

Since health is a function of both length of life and quality of life, the QALY has been developed in an attempt to combine the value of these attributes into a single index number. It has been defined as 'a measure of health outcome which assigns to each period of time a weight, ranging from 0 to 1, corresponding to the health related quality of life during that period, where a weight of 1 corresponds to optimal health, and a weight of 0 corresponds to health state judged to be equivalent to death' (Gold *et al.* 1996). The basic idea is that, for any individual, the prospect of living Y years in less than full health, or 'optimal health', may be equated to a prospect of living X years in full health where X<Y. If different 'Ys' can be converted into equivalent 'Xs' (i.e. QALYs), and if more QALYs are preferred to fewer, then QALYs can be used to inform resource allocation decisions.

The QALY can be represented graphically. Figure 2.1 shows the expected quality and length of life profiles of patients with severe angina and left main vessel disease (Williams 1985). The graph shows the length of life along the horizontal axis and the quality of life, measured on the zero to one scale, along