CHAPTER 5

Costing health services: health economics

Chapter contents

Introduction

With the current emphasis on the purchase of health and social services that are effective and also cost-effective, there is an increasing need for policy-makers, health professionals and managers, and researchers to be aware of the basic concepts of health economics. There is a related need to be aware of the type of data that should

be collected for economic analyses. This chapter describes the main concepts and techniques used by economists and the types of data that are required in relation to each. Cost data are complex to collect, and collaboration with a professional health economist is required in research projects which aim to evaluate the costs, as well as the health and social outcomes, of services. Useful introductions to costing health and also social services include those by Mooney (1992), Netten and Beecham (1993), Locket (1996), Drummond *et al.* (2005) and Brown (2005).

Health economics

Economic evaluation has its foundations in welfare economics (see Sculpher 2004, for a brief overview). The underlying assumption of economics is that the resources available to society as a whole are scarce, and thus decisions have to be made about their best use. For economists, resources are best employed when they maximise the benefit to society. This is as true for health care as any other resource area. Health economics, therefore, is about how health care resources are used in order to produce the greatest benefit to the population. This inevitably involves choosing between competing calls on scarce resources (e.g. should resources be spent on building another community clinic or on employing more nurses in existing clinics?). Decisions have to take account of what services have to be given up, or which planned services deferred, in order to pay for the alternative. In other words, the *opportunity cost* has to be assessed.

The basic assumption of economic analysis consists of 'rational' individuals or organisations operating in an 'ideal' market where goods and services are exchanged for resources. The 'ideal' market is where many buyers and sellers have free entry and exit, all organisations seek to maximise profits, maximum profit is made when the *marginal cost* of production is equal to the market price and there is a situation of *perfect knowledge*. Knowledge is necessary because individuals must be able to exercise informed choices which achieve a desirable outcome (their choice is, of course, limited to the opportunities presented to them, which are determined by price and income, which are related to the amount sold). They are said to have a preference for a good or service that gives satisfaction (utility), and they work towards maximising that utility, in a world in which financial resources are scarce. The 'ideal' market is not always achieved, and is threatened by monopolies, monopsonies and oligopolies. A monopoly is a situation in which there is only one producer, who has the power to influence price and can price goods and services discriminately, selling to different buyers at different prices. A monopsony is a situation in which there is a single purchaser. An oligopoly is a situation in which a few producers compete and output and prices are subject to the interrelationships between the producers.

In economics, idealised markets (the collection and interaction of buyers and sellers) operate according to the laws of supply and demand (see later). The aim of organisations (e.g. hospitals) is assumed to be the maximisation of profit (or its equivalent), and their constraints relate to the production process. Health economics, however, deals with an 'imperfect' market situation. The health care market is frequently referred to as the internal market or quasi-market. This is the application of rules to ensure an increase in efficiency and improved allocation of resources within the framework of the organisation (see Locket 1996, for examples of different methods of financing and organising health care).

The 'social good' is also relevant to economics, as expressed in the concepts of the *efficiency* of the distribution of resources and *equity* (which may conflict). Efficiency can be defined in relation to allocative efficiency (the allocation of resources to maximise the benefits to the population) and technical efficiency (the achievement of maximum benefits at minimum costs). Equity can be interpreted in a number of ways: for example, the fairness of the distribution of resources; entitlement to resources in relation to need or contribution; the production of the greatest good for the greatest number. Whitehead (1994) argued that equity related to everyone having a fair opportunity to attain their full health potential, and that no one should be disadvantaged from achieving this potential if it could be avoided.

Patient choice and equity

Most economic analyses involve the study of individuals making choices under constrained conditions (Le Grand 2003). This is because the availability of choice may lead to some patients choosing more expensive (though not necessarily more effective) health technologies and systems (Oliver 2003). Patient choice is arguably a feature of *equity of health care*. But the theoretical free will of individuals to make choices is constrained by the social system. It may have equitable or inequitable consequences if some patients (e.g. the more educated and affluent) are more aware of the choices being offered, and more adept at making more beneficial, though not necessarily more cost-effective, choices than others. In addition, the range of choices offered may vary depending on the geographic area and its characteristics.

Macro- and micro-level analyses

Health economists, then, are concerned with economic evaluations of health care in terms of costs and benefits. The costs and benefits of health care are analysed at the *macro* (the larger scale of organisations, communities and entire societies) and *micro* (the individual organisation, community and society) levels. Farrar and Donaldson (1996) have provided examples of the macro and micro levels in relation to care for elderly people. At the *macro level* one question is whether the ageing population is a growing burden that can be afforded by western societies. They break this question down into two issues. First, is an ageing population going to constitute an increasing economic burden? Second, what does 'affordability' in relation to health care mean? In relation to the first question, economists work with demographers and social scientists in order to assess trends in the age structure of the population, including information on morbidity rates and, in particular, on morbidity compression (the concentration of morbidity into the last years of life rather than spread out across older age groups). They then relate this to information on the costs of addressing these patterns of morbidity to provide estimates of trends in health care costs and expenditure for older populations. Regarding the second question, on affordability, the concept of opportunity cost is relevant: what has to be given up in order to provide the care in question, and is that what society wants? In other words, what proportion of society's scarce resources should be devoted, not just to health care, but to the health care of elderly people? At the *micro level*, economists are concerned with the costs and benefits of different ways of caring for elderly people *within* societies and ensuring that health care resources are spent in the best possible way. At this level it is essential to include the costs and benefits incurred by all relevant sectors, regardless of budget demarcations (e.g. primary and secondary health services, social

services, voluntary sector), as well as the public (i.e. patients) and wider society (Farrar and Donaldson 1996).

Demand, utility and supply

Although health care markets operate differently from other markets, economists still use the concepts of demand, utility and supply in their analyses. *Demand* refers to consumers' willingness to pay for desired goods and services, in the context of limited resources. It assumes that the consumer is in the best and most knowledgeable position to decide what values should be attached to various goods and services, though this is less likely to be the case in relation to health services (Mooney 1992).

The demand curve for a good or service illustrates the relationship between its price and the quantity desired (holding other variables constant, such as income and the price of other goods). The curve usually indicates that the lower the price, the greater the quantity desired (sloping down from left to right). *Elasticity* refers to the degree to which demand responds to price.

The concept of *utility* underlies the concept of demand. It simply refers to consumers' satisfaction. Economists assume that the greater the utility (satisfaction) obtained from the good or service, the greater will be the price the consumer is willing to pay for it. Related to this are the concepts of marginal utility (the additional utility obtained from consuming one extra unit of the good or service) and diminishing marginal utility (as more units of the good or service are consumed, the utility obtained from each additional unit of consumption tends to fall).

Supply refers to how the costs of producing goods and services and the prices of the final product affect the quantity supplied. The supply curve illustrates the relationship between the price of the good or service and the quantity supplied, holding other variables constant (e.g. the price of other goods). The prices result in revenue, and the additional revenue for each extra unit produced is the marginal revenue. The curve reflects the incentives for the producer, in that the higher the price of the commodity, the more the producer will be prepared to devote resources to producing it because, if nothing else changes, profits can be increased (thus the curve slopes upwards from left to right). Maximum profit is earned when the output is set at the point where the marginal cost is equal to the marginal revenue.

The concept of *costs* is related to that of supply. In theory, the producer will supply goods only if costs can at least be covered. Producers have fixed and variable costs. Fixed costs are constant regardless of the volume of output; variable costs vary with the volume of output. The higher the price, the greater is the likelihood that costs will be met, and thus greater profits obtained; hence the supply curve is usually positive. In contrast, the consumer aims to maximise utility. This is where the notion of competition is relevant, because it refers to the negotiation between producers and consumers over prices.

The limits of economic analysis

Economists agree that the crude application of these concepts to health care is obviously inappropriate. Microeconomic methods of analysis are of limited value in situations where there is an agency relationship between the consumer and the provider, and where consumer choice is constrained by several factors – from lack of technical information

for consumers to exercise informed choice to limitations in provision. A further danger of economic analysis is that the values given to individual and aggregate utilities become 'more real' than those of the individual or the groups they are said to represent (Ashmore *et al.* 1989). Locket (1996) pointed out that economic analysis should only be performed where there is information that a health care intervention works, is acceptable to patients (i.e. they will use it) and is accessible to those who need it.

Economic appraisal

Economic appraisal is the comparative analysis of alternatives in terms of their costs and consequences, and can take a variety of forms. It covers a range of techniques but the main approaches (which include cost minimisation, cost-effectiveness analysis, cost–benefit analysis and cost–utility analysis) are described in the following sections. Each involves systematic approaches to the identification and measurement of the costs and consequences of a particular service or intervention. These concepts have also been clearly described by Drummond (1994), who also points out that many economic evaluations do not fall neatly into one of these categories. For example, some investigators report a range of costs and consequences without attempting to aggregate the costs or the health benefits or calculate a cost-effectiveness ratio. This approach is labelled a cost–consequences analysis. The decision-maker (e.g. health care purchaser) then has to make the trade-offs.

With all costings it is important to collect up-to-date information, and each piece of cost information should relate to the same time period. The cost information must also be comparable between sampling units. The process is far from straightforward in relation to data collection, interpretation and analysis. For example, Kelly and Bebbington (1993) have described the considerable problems of reliability (the consistency of measures across location and time) of organisations' measures of unit costs and the caution required in interpreting and analysing these. Because of such inconsistencies, and because of the unavailability of data, economists are often forced to make assumptions about costs and organisational characteristics in their costing formulae. These assumptions are not always admitted to or made explicit but they should be, so that the reader can critically assess the validity of the exercises. Economics, while quantitative, is not an exact science, and many value judgements underpin costing analyses. Barendregt and Bonneux (1999) criticised the high level of arbitrariness and lack of transparency in economic evaluations of health care. They pointed out that standardisation would increase transparency, but not if the standard required researchers to make additional assumptions and use controversial methods of imputation. They argued that a modest standard would be a 'boon for transparency'.

Box 5.1 A review of health economics evaluations in RCTs

Barber and Thompson (1998) carried out a review of statistical methods used for health economics evaluations in RCTs published in 1995, identified from Medline. They found that information about the completeness of the cost data was given for only 24 (53 per cent) of the studies, there were major deficiencies in the way cost data in RCTs was analysed, and misleading conclusions about the costs of alternative therapies were often reported in the absence of supporting statistical evidence. They called for guidelines to improve practice.

The challenges involved in obtaining the prices of resources, the necessity of adopting a hybrid approach in many cases, as well as choosing estimates from the research data (with assessments of how typical they are) and nationally available statistics have been outlined by Drummond (1994).

Cost minimisation

Cost minimisation compares the costs of achieving a given outcome. This approach is used when the outcomes of the procedures being considered are known to be the same (e.g. two drugs whose efficiency and side-effects are the same). This makes it possible to focus on identifying the least cost option without having to worry about measuring and comparing outcomes. Cost minimisation should be undertaken *only* where there is a very high confidence that the outcomes are the same, because if they are, in reality, different, then the analysis will give misleading results. There are few cases in which health care interventions are identical in this way.

Cost-effectiveness

Because it is rare to find health interventions which are similar in effects to permit cost-minimalisation analyses, it is more usual to compare the difference in costs between interventions with the difference in their consequences. Where there is just one main parameter in this respect (e.g. cost per life year), cost-effectiveness analysis is used.

Cost-effectiveness analysis is an approach to the assessment of efficiency which is concerned with the measurement of outcomes in 'natural' units (e.g. improvement in health status), which are compared with the monetary costs of the health care. The cost-effectiveness of a health care intervention is defined as the ratio of the net change in health care costs to the net change in health outcomes. For example, if the total costs of the care have been calculated, and if a health status or health-related quality of life scale has been administered to a sample of the patient group of interest before and after exposure to the care under study, then the cost per change in health status/health-related quality of life can be calculated. An incremental analysis can examine the incremental change in effectiveness and costs of moving from one type of care to another (e.g. outpatient care to GP care). A decision will have to be made when the results are interpreted as to whether any observed increase or reduction in costs is enough to compensate for any increase or decrease in resulting health status/health-related quality of life.

With cost-effectiveness analysis, the costs are more narrowly defined than with cost–benefit analysis. They are generally confined to monetary measures of output (effectiveness) and are limited, as they have difficulties coping with more than one output.

Cost–benefit analysis

Cost–benefit analysis refers to approaches which assign a monetary value to the benefits of a project and compare this with the monetary costs of the project. This enables comparisons between alternative projects to be made in terms of efficiency.

Cost–benefit analysis values *all* costs and benefits in monetary units and enables the total service cost to be calculated (see Allen and Beecham 1993, for details). This is the broadest method. Once calculated, costs should be disaggregated to a unit of measurement that is as close as possible to client-level data in order to obtain a relevant unit cost for each service (e.g. hospital use is counted by the number of inpatient days or outpatient attendances) or to even more detailed levels (e.g. ward costs).

Cost–benefit analysis is used in decision-making about whether to introduce (or maintain) a particular programme (i.e. service). The principles underlying cost–benefit analysis are that programmes should be implemented only where benefits exceed costs, and they should not be implemented where costs exceed benefits (Mooney 1992). The point about cost–benefit analysis is that it allows *different* services to be compared (e.g. renal dialysis with rheumatology clinics). Because of the methodological complexities of measuring and including all health benefits in the analysis, some economists use 'willingness to pay' techniques instead.

Marginal cost

The marginal cost can be defined as the additional cost of producing one extra unit of output (e.g. of treating an extra patient), and includes staffing and treatment costs, but not buildings and large-scale capital equipment costs.

In *marginal analysis* the basic rules of cost–benefit analysis are applied at the margin (it is not to be confused with the use of the same term to refer to a method of asking groups of professionals to reach a consensus on where to spend or cut a given monetary amount of resources). The assumption is that a programme can be expanded or contracted to the point where marginal benefit equals marginal cost, except if there is budgetary constraint, when all programmes should operate at the level at which the ratio of marginal benefit to marginal cost is the same for all.

In relation to marginal costs, Allen and Beecham (1993) pointed out that short-run marginal costs are inappropriate for most costing exercises, as they do not include the full costs of, for example, creating new services. Long-run marginal costs enable analysis of the differences which the alternative service being studied will make to available resources. However, as knowledge of future events and costs is uncertain, the convention is to use short-run averages, which include both revenue and capital elements as an approximation for long-run marginal costs (on the assumption that *relative* price levels remain stable).

Complete costs

With cost–benefit analysis, all costs and benefits, from all sources (e.g. health and social services, voluntary sector and individuals as well as wider society) that arise from implementing the objective are relevant because the welfare of the whole society is regarded as important, and not just the health service. They are not confined to monetary measures of costs, but also encompass benefit valuations. Because costs are usually measured in monetary terms, economists want to make benefits commensurate with these and to measure them in monetary terms.

In addition, complete costings should include the costs to the individual patients and to any carers, as well as their opportunity costs (i.e. what they would have been doing instead and the costs of this). The economic costing of patients' and carers' time

has not been resolved and is still fairly crude. The problem with costing people using a labour market cost, for example, is that not everyone works and this does not take leisure time into account. Some economists in the UK use the Department of Transport's (1994) estimate for the cost of leisure time, but this produces an embarrassingly low value of the cost of people's time (i.e. in relation to a few pence). It is good practice to carry out a *sensitivity analysis* using guestimates of the value of leisure time.

Where prices are charged (without a subsidy) for a health treatment, it is easier to set a monetary value on the services received. In socialist health care systems, however, there are, in the main, no charges for services. In such situations, economists sometimes consider the possibilities of public 'shadow prices'; that is, prices fixed by the state with the aim of reflecting the amount of resources that the community is willing to give up in return for a unit improvement in health. The attraction of shadow prices is that they can provide a practical approach to the problem of assigning monetary values, but they are a crude answer to a complex question.

In summary, time can be costed in relation to market activity (e.g. wages and salaries), leisure activities, meeting physiological needs (e.g. sleep) and productive, non-market activity (e.g. housework, caring for dependent people). Ideally, the impact of each type of activity that was forgone as a consequence of the service (or illness itself) would be costed separately. This is complex because of the lack of valid information on the cost of leisure time (based on the impact it has on market and non-market productivity). These issues have been discussed by Allen and Beecham (1993).

Intangible costs

When one is undertaking a cost–benefit analysis, an important issue is deciding which 'intangible costs' should be included. Intangible costs include things like work time and leisure time forgone (see above), the value of reassurance that accompanies a negative diagnostic test result and the reduction in stress gained by carers from respite care. In deciding which intangibles to include, it is useful to consider whether the gathering of more data on intangibles will change the results of the study significantly and whether the costs of gathering the data are prohibitive (Drummond *et al.* 1997).

Event pathways

The *Cochrane Collaboration Handbook* (1994) states that the following information is required for economic evaluations:

- identification of all main event pathways that have distinct resource implications or outcome values associated with them;
- estimation of the probabilities associated with the main event pathways;
- descriptive data to enable the resource consequences associated with each pathway to be measured;
- descriptive data to enable the outcomes associated with each pathway to be valued.

Event pathways are defined as a clinical event, details of its management and resources used for it, associated subsequent events and the cost of these resources.

Opportunity cost

The cost of spending resources in a particular way is not necessarily the monetary cost of the resource, but it is the opportunity lost (benefit forgone) by loss of its best alternative use. As described earlier, scarcity of resources implies choice, and this choice gives rise to the concept of opportunity cost. Given the scarcity of health care resources, it follows that the allocation and use of resources for one type of health care involves sacrifice of another. While the financial concept of cost simply relates to monetary outlays, the economist's concept of cost takes other considerations into account. Economists are interested in the health benefits that could be obtained by using the resources in a different way. Therefore, they measure costs in terms of the benefit that would be derived from the alternative use of the resource (*Cochrane Collaboration Handbook* 1994). In practice, money is a convenient yardstick against which to measure benefits and is generally used (Knapp 1993).

Problems with the calculation of opportunity costs

Opportunity costs are not straightforward to calculate. In particular, there is the issue of non-marketed items, on which economists attempt to put monetary values. These have been described by Knapp (1993), who points to three approaches to their valuation:

- the *human capital approach*;
- *implicit valuation methods*;
- clients' *willingness to pay*.

Human capital approach

With the human capital approach, earnings are used to value the effects. For example, the treatment may enable patients to return to work, or take less time off work, and this could be valued in societal terms of the extent of growth in national productivity. However, as some people are unemployed or retired or do not work for other reasons, there is little scope for using this approach. For the same reasons, loss of earnings is also problematic in relation to valuing the individual patient's opportunity costs. In addition, some people are salaried and do not necessarily lose earnings through time off work (e.g. to attend for treatment). In relation to predicting demand for health care, Torgerson *et al.* (1994) pointed to the importance of the private opportunity costs of time itself (i.e. the time taken to utilise health services) as a preferable measure to wages forgone.

Implicit valuation methods

Implicit valuation methods are based on the preferences for services that patients, clients and professionals reveal by their explicit behaviour. People are asked to put a price on the alternatives available in terms of how much they would be prepared to pay for them. This enables their expression of preference, and indirectly of satisfaction, to be calculated in financial units, which can then be directly compared with the actual financial cost. In theory, this facilitates policy decision-making about which alternative to purchase. It assumes an unproblematic relationship between price and cost. It is essential to be explicit about the assumptions and methods used.

Willingness to pay

Despite economic evaluation of health care being dominated by cost-effectiveness and cost per QALY comparisons between treatments, it has been argued that more use should be made of willingness to pay methods, especially in relation to cost–benefit analyses (Hanley *et al.* 2003). There are two main methods for setting a monetary value on a specific package of health or other benefits: '*contingent valuation method*' and '*choice experiments*' (previously called conjoint analysis, now referred to as discrete choice experiments).These methods have generally been used to set a monetary value on a package of health and/or non-health benefits in the context of a specific intervention. These methods have also been criticised as subject to bias and lack of sensitivity (Cookson 2003).

Gerard *et al.* (2008) described willingness to pay questioning, in which people are asked in surveys how much they are willing to pay for a good or service, as a contingent valuation method. This is because they are asked about their willingness to pay contingent on a hypothetical scenario and description of the good or service being valued. Willingness to pay is based on observed trade-offs between resources or states of health/ill health (Donaldson 1993; Drummond *et al.* 1997, 2005). In relation to health and social care, particularly in societies with government-controlled services, such exercises are often too hypothetical and difficult for many people to conceptualise: health care does not have an *explicit* monetary value. Some would also object, on ideological grounds, to asking people to consider the costs of health care when it is provided free at the point of consumption (see Ryan 1996).

Discrete choice experiments

Some economists have used *discrete choice experiments* to elicit people's values, in which preferences for scenarios (levels of attributes of the good or service) are obtained through surveys asking people to rank, rate or choose between scenarios. This provides a more realistic estimation of the relative importance of different attributes in the provision of a good or service, the trade-offs between the attributes, and the total satisfaction or utility the individual derives from the good or service with specific attributes (Ryan 1996) (see later section on discrete choice experiments).

Price stability

Knapp (1993) also outlined the problem of price stability. Even with valid information on market costs, does the economist take the initial or final price as the measure of opportunity cost? One example is that if a health authority, or other large health care purchaser, decided to increase greatly the number of elderly people discharged from hospital to the care of a hospital at home scheme (as an alternative to a longer hospital inpatient stay), then this would affect the supply price of the hospital at home scheme. So should the previous or the subsequent supply price of services be used? Knapp suggested using a formula that takes account of both.

Other problems with opportunity costs

Other complications include the issue of apportioning joint costs (e.g. where costs are met by social and health services, or social and health services and the individual), the issue of private costs (e.g. services provided within the organisation by external public

or independent agencies), and costs to society, and price distortions. For example, in relation to price distortions, if the resources are supplied by a monopoly organisation, the price and cost will differ; indirect taxation distorts prices; if staff would otherwise be unemployed, they will have a zero shadow price (Knapp suggested setting their price as equal to forgone leisure time and the costs of travelling to and from work, although other complications, such as government policy, need to be taken into account). Knapp listed the following implications of an opportunity costing approach to social care, which can be applied to health care: the opportunity cost of using a resource in a particular way cannot be measured without knowing what alternative uses are available; costs are forgone benefits; opportunity costs are context-specific; some apparently costly items are costless; some apparently free items have non-zero costs.

Discounting

Discounting is designed to standardise different cost–time profiles, though the concept is untested. It is important to take into account the time period of the incurred costs and benefits. Future benefits are valued less than current benefits, regardless of inflationary effects (e.g. because desires may change). Discounting of the future is also based on the assumption that most people's real income increases over time. The UK Treasury sets a percentage discount rate for public sector projects. If discounting is employed, it is prudent to consider a range of discount rates as part of a sensitivity analysis. Mooney (1992) pointed out that this is particularly problematic in relation to health promotion services, where the benefits will not be obtained until the future.

Cost–utility analysis

Cost–utility analysis provides a fuller analysis of health care benefits than cost–benefit analysis, because it takes patients' quality of life into account. However, societal costs and benefits are generally still ignored. Different interventions often have several different health outcomes (e.g. efficiency versus side-effects of treatment; length versus quality of life). In such cases cost–utility analysis is used, where the different changes in health states are valued relative to each other. Cost–utility analysis is a technique that relates the cost of the project to a measure of its usefulness of outcome (utility). This produces an overall index of health gain, or health status in relation to output. The quality-adjusted life year (QALY) (see later section) is one index used, which attempts to combine quantity and quality of life into a single index, which gives it an advantage over single-dimensional measures of output (as in cost-effectiveness analysis).

This form of analysis requires the different impacts of the treatments on length and quality of life to be measured. It also requires a set of values for health states ('utilities'), in order that the different health states can be compared. These values can be derived from existing data, where available, and where relevant to the population of interest, or they may need to be derived as part of the specific study. Whether QALYs really are utilities is open to debate (Drummond *et al.* 1997).

Cost–utility analysis also provides one approach to addressing issues of *efficiency* of resource allocation in relation to the determination of *health* priorities. The advantage

is that the approach is not solely monetary. However, it has several disadvantages in that it does not adequately address issues of equity in health care, or take account of objectives of health services other than the maximisation of health. It also follows the questionable assumption that it is based on an adequate measure of health.

Cost–utility analysis and summary health indices

Cost–utility analyses require outcome measures which combine quality of life and mortality outcomes. A *value of health states* is also necessary for cost–utility analysis. Economists have developed questionnaires that aim to measure quality of life with the advantage that the data derived can be applied to a pre-scaled matrix of health state preference values for use in cost–utility analysis (Drummond 1994). The values used are the expressed preferences for states of function on a scale anchored at the extremes by death and optimum functioning. The preference values can be derived from existing research data, where appropriate, or by undertaking direct utility measures within the study concerned. The measures can be used, in theory, as either health-related quality of life measures (if all domains are tapped) or as instruments in cost–utility analyses.

QALYs

Cost–utility analysis uses the QALY which claims to take account of quality of life and length of life. This is used for health care decision-making as it enables treatments for different conditions to be basically compared. The QALY is a form of health status measurement which places mortality and morbidity on the same measurement scale. The QALY figure reflects the change in survival (known as 'life years') with a weighting factor for quality of life. QALYs are used in making comparative assessments about the effectiveness of various treatments. Costs of the treatment per QALY are calculated and generally presented in QALY league tables (e.g. showing QALYs for hip replacements, bypass surgery, etc.). Caution is required in interpreting QALY league tables in view of the relatively crude methods underlying the calculation of QALYs, and the assumptions made.

The QALY takes one year of perfect health-life expectancy as worth a value of 1, and one year of less than perfect health-life expectancy as less than 1. Phillips (1996) explained the formula clearly as follows. An intervention which increases a patient's life expectancy by four extra years, rather than the patient dying within one year, but where quality of life falls from 1 to 0.6 on the continuum, generates the following value:

4 extra years of life at 0.6 quality of life values 2.4

Minus one year at reduced quality (1 – 0.6) 0.4

QALYs generated by the intervention 2.0

The assumptions underlying QALYs are open to criticism, as is their construction. Measures that include time as a dimension register fewer benefits for elderly people because of their shorter life expectancy, in comparison with younger people (Farrar and Donaldson 1996). Defenders of the QALY counter that it simply reflects the public's ageist attitudes, as the QALY values were developed from public surveys (Edgar *et al.* 1998).

QALYs can also be criticised on the grounds that they focus on cures rather than care, and are thus less appropriate for use in the priority setting of chronic care, in comparison with acute services. The fact that every QALY calculation places hip replacements over home or hospital dialysis in relation to value for money raises deep moral concerns (Lockwood 1988; Butler 1999). These limitations of the QALY are likely to result in treatment decisions which are inequitable and less than optimal. While various health economists have questioned the restrictive assumptions inherent in the QALY, and attempted to build in utility functions to address its shortcomings, their attempts remain inconclusive. There is also evidence that society expects broader benefits from health care than simply health and utility, including empowerment, social participation, feelings of safety, self-respect and dignity (Coast *et al.* 2008a, 2008b; Byrne *et al.* 2010).

QALYs have been reported to be less sensitive than other measures of physical functioning and emotional well-being when used to assess the health status of elderly people (Donaldson *et al.* 1988), suggesting that their use in priority setting might place elderly people lower down on the priority list than they ought to be. Oddly, however, as Farrar and Donaldson (1996) pointed out, the QALY league tables have ranked hip replacements and chiropody highly. There is undoubtedly a need for caution, particularly given the relative lack of robust evidence on costs and effectiveness of many treatments and procedures.

Decisions about priorities for health care interventions, owing to limited resources, entail making trade-offs between their estimated benefits and their estimated harms and costs. QALYs can be used in decision-making about health priorities, though this use is controversial. Different health care programmes can be compared in league tables in relation to their marginal costs per QALY obtained. The practice is that the programmes with the cheapest QALY are given the highest priority. This is based on the assumption that, with limited health care resources, the aim is to maximise the number of QALYs purchased within the budget (Mooney 1992). For example, in England and Wales, the National Institute for Health and Clinical Excellence (NICE), which was established in 1999, was charged with making decisions about whether treatments should be made available by the NHS, taking cost-effectiveness (or 'value for money') into account. NICE aims to purchase the greatest number of QALYs possible – i.e. to maximise health gain in relation to available funds. The QALY indicator is also used in other countries, including by the Pharmaceutical Benefits Advisory Committee in Australia, and the Canadian Agency for Drugs and Technologies in Health.

Eliciting values

Methods of eliciting values for specific health states include the *rating scale* (also called the *visual analogue scale*), *time trade-off* and the *standard gamble* (Torrance 1976, 1986). On the basis of their review of the literature of these techniques, Brazier *et al.* (1999) concluded that each was a practical method for use with most populations, though evidence on reliability and validity was generally lacking for each method. If anything, the rating scale appeared to be the most practical and acceptable technique. Other methods include the *magnitude estimation* and the *person trade-off*, though these two methods have little supportive evidence in favour of their psychometric properties or their practicality (Brazier *et al.* 1999).

UK economists (Kind *et al.* 1982; Williams and Kind 1992) developed an alternative method initially based on the Rosser Disability Index (Rosser and Watts 1972). However, this early method has long been criticised as limited, and of questionable reliability and validity. Brazier *et al.* (1999) argued that there is no place for this method of preferences in economic theory. The QALY formula now used in the UK is based on the utility value of a health state derived from the EuroQoL-5D (EQ-5D). In the USA, Kaplan and Bush (1982) and Kaplan *et al.* (1984) developed a slightly different approach using their more detailed Index of Well-being Scale. The Health Utilities Index, which is broader than the EuroQoL, is also increasingly popular (Feeny 2005), particularly in North America.

Alternatives scales to QALYs have been developed, such as the Time Without Symptoms and Toxicity scale (TWIST) (Gelber and Goldhirsh 1986; Gelber *et al.* 1989). It has been argued that a measure based on 'healthy years equivalent (HYE)' is more representative of people's preferences (Mehrez and Gafni 1989). The HYE also combines quality and quantity of life, and is regarded as an improvement on the QALY because it obtains the utility for the whole health profile (rather than each state separately), and therefore more fully represents the patient's preferences. It also allows attitudes towards risk to be incorporated. HYEs are calculated using two standard gamble questions and respondent burden is relatively high. Other alternatives to the QALY are the Saved Young Life Equivalent (Nord 1992), which compares treatment outcomes in terms of units of saving a young life and restoring the young life to full health; Quality-Adjusted Lives (Stevenson *et al.* 1991) (treatments are assessed in terms of number of lives saved rather than length of life); and Healthy Life Expectancy (Bone 1992) (an indicator of the health status of a population, combining morbidity and mortality into a single index used in epidemiological and demographic studies; there are three different methods for calculating this; Barendregt *et al.* 1994). Methods of eliciting utility values are time-consuming and complex, as well as imposing considerable respondent burden.

The rating scale (VAS)

The rating scale involves a horizontal line (a visual analogue scale or VAS) anchored at one end with 0 which is equal to death, and at the other with 1 or 100, which is equal to the best/most desirable state of health. It is used with a given health state. The scale is given to study members, in conjunction with a description of the given health state, who are asked to make judgements about where on the line various intermediate states of health lie. For example, if a particular state of (ill) health (e.g. diabetes) is judged to be 0.75 or 75, then the respondents perceived this state to reduce their health status by a quarter.

Torrance *et al.* (1982) specified attributes, which are graded, that should be included in a health state: physical function, emotional function, cognitive function, self-care and pain. The characteristics of the given health state include a description of these attributes either in written vignettes or shown on video. A technique known as multiple attribute theory is used to determine the value for each level of the attributes and the utility value of the associated health state (Torrance *et al.* 1982).

Time trade-off

This method involves asking respondents to establish equivalents. They are asked to consider an ill-health state that is to last for a fixed period of time. They are informed that a new health care procedure will give the individual normal health for a shorter period of

time, but with the likelihood of death or severe disablement at the end of that time. The respondent is asked to 'trade off' the time with the ill-health state with normal health for a shorter period of time. The time spent in normal health is varied until the point of indifference is found. Variations include trading off the number of people helped by different treatments (e.g. how many people in state B must be helped to provide a benefit that is equivalent to helping one person in state A?). This method has been reported to be more reliable than the standard gamble technique (Dolan *et al.* 1993). Others have reported that the standard gamble has better construct validity than the time trade-off (Puhan *et al.* 2007).

Standard gamble

This asks the respondent to make a choice between remaining in a current state of ill health and the probability of being immediately restored to perfect health, with some chance of immediate death (e.g. in relation to a specific health care intervention). The respondent is asked to vary the level of probability until the point of indifference between choices is reached. As indicated above, the reliability and validity of these methods are debatable. Time trade-off and standard gamble techniques both suffer from a disjuncture between stated preferences and actual choices (Ryan *et al.* 2006).

The Rosser Index of Disability

The Rosser Index of Disability is an early measure, based on the concept of a health index, with people (or descriptions of health and ill-health states) being graded by respondents, recruited to make the assessments, into one of eight areas of disability, from no disability, to slight social disability, through to confined to bed, and unconscious. Each state is graded on a four-point distress scale: none, mild, moderate or severe. States are scored on a scale ranging from 0 at death to 1 for healthy (with negative values for states judged to be worse than death). Once these rankings have been completed, respondents are asked to undertake a series of complex priority ranking exercises in relation to the conditions assessed. For example, they are asked to place the conditions (or 'health states') on a scale in relation to 'how many times more ill is a person described as being in state 2 than state 1'; they are also asked to place a state of death on a scale of permanent states (e.g. vegetative state), and to assign a value to it (see Kind *et al.* 1982). There is no justification for this method as a measure of preferences in economic theory (Brazier *et al.* 1999). Results for inter-rater reliability and construct validity, including sensitivity to clinical outcomes, have been inconsistent (see Brazier *et al.* 1999).

Kaplan's Index of Well-being

The Quality of Well-being Scale provides an index value as well as quality of life descriptors. It is the best-known measure in this field. It was developed in order to operationalise 'wellness' for a general health policy model, in an attempt to develop an alternative to cost–benefit analysis for resource allocation (Kaplan *et al.* 1976, 1978, 1984; Kaplan and Anderson 2004). The instrument defines levels of wellness on a continuum between death and optimum function and integrates morbidity and mortality into the same number. It classifies respondents according to their level of functioning on three scales – mobility, physical activity

and social activity – combined with their (most undesirable) problem/symptom. The level of function and the reported complaint (symptom) are weighted by preference on the scale of 0 (dead) to 1.0 (optimal functioning). The aim was to produce a point-in-time expression of well-being, so it has a fairly short-time reference ('preceding six days'). The scale is interviewer-administered, is lengthy and difficult to administer. It has been used extensively in clinical trials and population studies in the USA, and has good levels of construct validity when tested against other physical health status scales, but correlates poorly with measures of emotional well-being and psychological health (see review by Brazier *et al.* 1999). Its retest and inter-rater reliability is also unknown (Brazier *et al.* 1999). It has relatively few floor or ceiling effects. However, its widespread use has been hindered because it is complex to administer, although a self-completion version has also been developed (it is still 80 items and the time reference for reporting symptoms and difficulties functioning is either 'current' or scaled as 'no days/yesterday/2 days ago/3 days ago') (Andresen *et al.* 1998).

To derive a single utility score ('Kaplan's Index of Well-being'), Kaplan and his colleagues (Kaplan and Bush 1982; Kaplan *et al.* 1984) placed people with given health states into categories of mobility, physical activity and social activity, and then classified their symptoms and health problems on a given day. Case histories were compiled in order to illustrate the combinations of functional levels, symptoms or problems. The scale also includes death. Random samples of the public were asked to rate preferences to the descriptions, and weights were derived for each level of mobility, physical activity, social activity and symptom or problem. A utility value was assigned to each functional level, and questionnaire responses were used to assign the health states to one of a number of discrete function states. Kaplan's Index of Well-being, which provides a single score, developed out of this methodology (Kaplan *et al.* 1976, 1978; Bush 1984; Kaplan and Anderson 2004). The scale quantifies the health outcome of a treatment in terms of years of life, adjusted for changes in quality.

EuroQoL

The aim of the EuroQoL was to provide a self-administered, standardised, generic instrument for describing health-related quality of life and to generate a single index value for each health state. The EQ-5D, after revision, now contains five questions which can be used to generate a single summary index of health status, while still permitting analytical breakdown into its five domains, and a self-rating of health on a vertical visual analogue 'thermometer' scale from 0 (worst imaginable health state) to 100 (best imaginable health state) (EuroQoL Group 1990; Kind 1996; Dolan 1997; Kind *et al.* 1998). It measures current health on five domains: mobility, self-care, usual activities, pain/discomfort, anxiety/depression, and the items use three-point response categories. The EuroQoL was also later revised (middle values were added to increase sensitivity) and shortened. It does not cover broader health-related quality of life. A tariff, using time trade-off methods, derived from population samples in the UK, is then used to value the person's health state.

UK preference values for the EQ-5D were initially derived from time trade-off techniques, which is an accepted method for deriving preference values, with just over 3000 people (Dolan 1997; Gudex *et al.* 1997). The earlier version achieved adequate construct and convergent validity (Brazier *et al.* 1993a), there is evidence of its test–retest reliability (see review by Brazier *et al.* 1999), and there have been some design improvements. While research indicates it is less sensitive to changes in specific disease states (Wolf and Hawley 1997), it has been reported to be sensitive to variations in

response to selected self-perceived health questions in a general population survey in Canada (Houle and Berthelot 2000) but with reduced sensitivity at the ceiling (i.e. at the lower levels of perceived health) (Brazier *et al.* 1993b). The EuroQoL has been reported not to be responsive to some conditions (e.g. vision – see Datta *et al.* 2008), and some items are so extreme that few people endorse them (see Wolfe and Hawley 1997).

Brazier *et al.* (1999), on the basis of their review of the literature, judged the EQ-5D, along with the Health Utilities Index (Mark III) to be superior to the other preference-based measures.

However, the instrument still contains several fundamental design flaws. It has also been criticised as being insensitive to changes in health status that are important to patients (Jenkinson *et al.* 1997; Jenkinson and McGee 1998). The item wording is inconsistent in parts which can increase response error (e.g. the self-care domain scale wording varies inconsistently from asking about problems with 'self-care' to more specifically 'wash or dress'). The mobility domain scale leaps from 'no' problems and 'some' problems walking about to 'confined to bed' (omitting the group of older people who are largely confined to chairs during the day). Also, neither the pain/discomfort domain or the anxiety/depression domain scaling make any provision for those who suffer 'a little' (both domain response scales leap from 'no' symptoms to 'moderate' and then to 'extreme' symptoms). The scoring is linear and additive. It is still relatively crude and produces skewed results (Brazier *et al.* 1992, 1993a, 1993b; Carr-Hill 1992), with variable response rates from moderate to good (Brazier *et al.* 1993b; Gudex *et al.* 1997; Bowling 1998; Kind *et al.* 1998). The aim of such instruments is to produce a point-in-time expression of health-related quality of life, and thus the time reference period is 'today'. The respondent is asked to rate statements which best describe their own health state 'today' by ticking 'at least one box in each group'. Short time frames, while most likely to produce the most accurate data (e.g. less prone to recall bias) do, however, increase the potential for regression to the mean in follow-up studies, as ratings for just one day are less stable than for longer periods of time.

Health Utilities Index (HUI-3)

Another single summary health utility index is the Health Utilities Index, version 3 (HUI-3), which contains 31 questions in eight dimensions (vision, hearing, speech, mobility, dexterity, emotion, cognition and pain) (Furlong *et al.* 1998). These attributes were selected for inclusion because members of the general population rated them as the most important dimensions of their health (Feeny *et al.* 1996). The HUI employs five- and six-point response choices and incorporates the full range of no, mild, moderate and severe problems. Preference values for the HUI-3 were derived from standard gamble techniques with 500 Canadians, and scoring is multiplicative (utilities).

The HUI-3 has been reported to have ceiling effects (Sung *et al.* 2003) and it carries modest study and respondent burden (Feeny 2005). There is only fragmentary evidence to support the scale's sensitivity to change (Feeny *et al.* 1995), although tests with the third version indicate that its responsiveness to change is similar to the EuroQoL (Houle and Berthelot 2000). As stated earlier, Brazier *et al.* (1999) judged the EQ-5D and the HUI-3 to be superior to the other preference-based measures. However, neither instrument is apparently as good at predicting changes in health status as a simple VAS of self-rating of health (0 'least desirable state' and 100 'perfect health') (Houle and Berthelot 2000). Results for the reliability and validity of earlier versions of the

scale were poor (see Brazier *et al.* 1999). More recent results have been better and an improvement on the EQ-5D (Grootendorst *et al.* 2000; Fisk *et al.* 2005).

Discrete choice experiments

There is increasing interest in using discrete choice experiments, as well as other methods (e.g. ranking and scaling exercises), to estimate QALY values for cost–utility analysis. None are without problems (see Flynn 2010), as existing methods for generating preference weights are limited to changes in health states. Some health economists have adopted valuation methods used in other fields, in order to value processes of health care. Conjoint analysis is used in marketing, transport and environmental economics. In health care, the technique has been adapted (called 'stated preference discrete choice experiments') to predict service acceptance and utilisation, or option values. They are known as attribute-based approaches (Gerard et *al.* 2008).

This technique presents patients with hypothetical choices to make between services/treatments, and varies them with different attributes (components) that might underlie preferences for the choices. If cost is included as an attribute, then the aim is also to measure 'willingness to pay'. One example is the trade-offs people may be willing to make between location of treatment, or choice of hospital, and waiting times for treatment (Ryan and Farrar 2000; Burge *et al.* 2004). Statistical modelling of the results is used to provide estimates of the extent to which components of the service or treatment contribute to the preferences elicited.

It is essential to minimise measurement error and ensure that all attributes are included, and to check that respondents interpret the task accurately. As Lloyd (2003) has pointed out in his critical review, however, preference elicitation methods assume that people's preferences are stable and complete, and assume that they are consistent (Ryan *et al.* 2006). He suggested that little attempt has been made to apply psychological theories of judgement and decision-making which challenge the assumptions of the method. It is unknown to what extent preference elicitation is influenced by cognitive processing strategies and the employment of heuristics (cognitive strategies which result in shortcuts to simplify the cognitive functioning required for the task). Indeed, the measurement of human judgements, by both experts and lay people, is a highly specialised psychological subject (see Harries and Kostopoulou 2005; Harries and Stiggelbout 2005). As Harries and Stiggelbout concluded, preferences are prone to the inconsistencies inherent in making judgements, and are influenced by many external factors, including how information is presented, and time.

Disadvantages of methods

Each method has its limitations, and no gold standard exists. A major disadvantage of all these methods is their cost, owing to their time-consuming nature, the requirement for highly skilled interviewers and their complexity, leading to high respondent burden. The last point leads to a general reliance on convenience sampling, rather than random sampling, leading to results based on unrepresentative samples of the population.

One of the main debates surrounding the use of these techniques is whose values should be sought to provide utility values: those of the public, those of the health professionals, the patients and/or their families? Economists argue that patients' values would not be constant over the course of the illness, and thus the utility values would

not be stable. The issue remains one for ethical and methodological debate. There is some evidence that the methods used to elicit values in economic analyses do not tap underlying true preferences (Kahneman and Tversky 1983).

Costing health services

This section provides some examples of the types of costs that are collected by economists in health service evaluations, which include cost–benefit studies. The economic costs of health care technically come under the umbrella of the structure of health services. However, health economists aim to incorporate costs into the assessment of outcomes of health care because clinical effectiveness needs to be interpreted in relation to economic, or cost, effectiveness. Decisions about priorities for health care interventions, owing to limited resources, entail making trade-offs between their estimated benefits and their estimated harms and costs.

Costings are rarely straightforward: there are many methodological obstacles when one is making cost comparisons, and costings often require assumptions to be made that are seldom applicable across settings (Wright 1993) and would probably be unacceptable in many other scientific disciplines. The implication is that costings and comparisons of costs must be interpreted with caution. When any cost comparisons are made, it is important to ensure that the same service is being costed, given the sometimes enormous variations in the organisation and quality of care within any one type of service in different places. This is often extremely difficult to achieve. The valuation of cost and benefit in economic terms inevitably involves elements of subjective judgement. When cost and benefit are presented in quantified form, this point is often, unfortunately, forgotten. While health professionals' time can be costed using their salaries and overhead costs, the costs of lay carers, for example, are difficult to value. Mooney (1992) pointed out that even when these intangible costs cannot be valued, it is important to note them to prevent them being ignored in decision-making processes.

Thus it is important that the data collected for economic evaluations are accurate and comprehensive, that assumptions underlying any categorisations are made explicit and that the time periods for follow-up in the data collection are carefully planned in order that they incorporate the 'subsequent events'.

Capital costs

Capital costs are building costs, equipment and land and other capital-intensive items (e.g. expenditure on structural alterations). There are two components of capital cost: opportunity cost of resources tied up in the asset, and depreciation over time of the asset itself. Building costs require information on the valuation of capital, and can be based on annual payments of capital, plus any charges for depreciation and interest, and then apportioned to the unit of interest. At a simple level, if the total is divided by the number of patients booked per clinic, then a building cost per consultation can be derived.

The costs of the buildings (annuitised over the lifespan of the buildings) used for the services need to be included in the total costs. This enables an estimate of the long-run *marginal (opportunity) costs* of services to be calculated. The capital costs are counted alongside revenue costs to enable the total costs of a service to be presented in one figure.

The opportunity costs of the capital (buildings and stock) also need calculation. Allen and Beecham (1993) explain that it is convention to calculate opportunity costs of capital by assuming that the best alternative use of the resources is investment. The value of the resources thus includes interest which could have been earned had the money not been tied up in buildings and equipment.

Allen and Beecham have described how, in the case of private sector care where information on the valuation of buildings and other capital-intensive items might not be accessible or easily available, an acceptable compromise is to take the fee charged, on the assumption that this (market price) approximates the real cost and includes the cost of the original capital investment.

Overhead costs

Overheads relate to those resources that service different programmes: for example, expenses related to the building (e.g. power, rates), staffing costs and other costs of providing the service (e.g. associated with administration, transport, catering, laundry, maintenance, cleaning, stationery). This information is obtained from accounts of expenditure and salaries. Overhead costs include direct and indirect overhead costs. Where individual programmes are being costed, these overheads should be shared out.

There are costs associated with the building and stock, such as power, water and sewage charges and building rates, repair and maintenance, cleaning and other operating costs. They also include day-to-day expenses for supplies and services, immediate line management, telephones, and so on. These can be difficult to calculate, and where information on total overhead costs is obtained from the organisations themselves, additional information on how costs were apportioned is required, and should be adjusted if necessary in order to ensure that like is being compared with like.

In order to calculate overhead costs, there are two options: to accept the organisation's figures on these costs and the costs of, for example, a clinic attendance, with information on how they apportioned costs in order to ensure the comparison of like with like across the study; or to measure the square footage of the space occupied by the clinic, ward or other unit under study and the square footage of the total building, collect all cost data and reapportion costs independently. Most investigators opt for the former, given the time and resource implications of the latter alternative.

Salaries and costs

The total salaries of staff members need to be obtained. Staff costs are calculated by multiplying the hourly rate of salaries at the appropriate grades. There are several other factors that will need to be included in staff costings, such as weighting factors for labour market variations, merit awards of consultants, employers' costs and contributions. In the British NHS, for example, employees who work in London are given a London weighting allowance. These need to be taken into account if the cost of services (e.g. clinics) in London is to be compared with that of clinics elsewhere in the country. Some costings take average salaries, or mid-points on the relevant scales (if the mid-point of the salary scale is used, then it needs to be adjusted for merit awards, e.g. the total value of distinction awards given to consultants in a specialty is

divided by the number of consultants in that specialty; the average is then added to the consultant's salary).

The total costs for staff need to be calculated in relation to the unit under investigation (e.g. hospital outpatients' clinic) and will need to be allocated to that unit by dividing the total staff costs by the number of patients (e.g. booked to attend the clinic). This will give the cost per patient booked.

As before, the various staff costs should be spread over all the units of interest (e.g. patients booked into a clinic; appointment times) to give a cost per consultation.

Allen and Beecham (1993) describe the complexity of costing the time of staff employed in community and primary care. For example, in Britain the income of a GP is partly dependent on the type and amount of work done (e.g. certain minor surgical procedures for which additional payments are made) and the type of patients registered with their practices (there are higher capitation payments for older people).

Other costs may need to be taken into account. In evaluations of outreach clinics held by specialists in GP surgeries, for example, the travelling costs (e.g. a marginal cost mileage rate) of the specialist between sites also had to be included in the overall costings (see Gosden *et al.* 1997; Bond *et al.* 2000).

Apportioning to unit of study

As before, all costs need to be extracted and apportioned to the unit of study (e.g. clinics). They can be averaged, for example, in a costing of outpatients' clinics, by the number of patients booked per clinic. Annual overhead costs can be converted into an hourly rate by dividing by the average number of working weeks in a year and the average number of hours worked per week. The hourly rate is then equally apportioned between the clinics operating on the day on which the clinic is evaluated. Alternatively, overhead costs can be apportioned per hour to each type of clinic in a building by dividing total overhead cost by the total number of hours for which the practice or hospital departments were open.

Resource costs: patients' treatment costs to the health service

The allocation of treatment costs to individual patients involves tracking patients' use of investigations, including biochemistry (checking site of analysis in case costs vary), procedures, prescriptions, surgery, and so on. For this exercise the patients' notes are used, supplemented with reports from health professionals and patients themselves. The costs of each item have to be obtained. At the crudest and simplest level, the costs for diagnostic tests, procedures and operations can be obtained by reference to price lists compiled by the hospitals or community service units in the areas of the study, or to national cost sources, where held and accessible, by government departments. (See Box 5.2.) An example of the collection of cost data from individual health care sites and national sources, as well as patient-level data is found in Roderick *et al.* (2005). This is not without problems, as these costs may not always reflect true costs, and the higher prices of some procedures may be subsidising the lower prices of others. With prescriptions, the unit cost of items can be obtained from formularies (e.g. in Britain from the *British National Formulary*, which is published annually by the British Medical Association and the Royal Pharmaceutical Society of Great Britain). The information required for this is the name of

Box 5.2 Example of use of records

An example of medical and billing records being used to cost health care is the study of the costs and outcome of standardised psychiatric consultations in the USA by Smith *et al.* (1995). Their study was based on an RCT comparing an immediate with a delayed (for one year) standardised psychiatric consultation. The study was carried out with 56 somatising patients from 51 study doctors. The measures included the patient-completed Rand SF-36 to measure health status and analysis of medical and billing records. Smith *et al.* standardised the costs by costing all items according to Arkansas Blue Cross–Blue Shield charges, inflated at an annual compound rate of 7.3 per cent. There was a two-year follow-up period. This study reported that, using these methods, the intervention reduced annual medical charges by 33 per cent (particularly through a reduction in the number of hospital stays) and physical function was found to have improved slightly. The weakness of the study, however, is that it only focused on direct organisational costs, and did not take the intangible costs into account, nor those incurred by the patients and their families. Where intangible costs have not been included, this should be made clear.

the prescribed item, dose, form and duration. An alternative is to calculate defined daily doses (Maxwell *et al.* 1993), though this can be complicated and time-consuming.

Patients' costs

Patients' costs include their travel costs and other expenses (e.g. direct financial expenditure on goods and services, such as diet, prescriptions, equipment, aids; waged and non-waged time; care costs for dependants; future costs; and, in some cases, accommodation) in relation to their health care.

Patients also incur *opportunity costs*, which include forgone leisure time or time off work to attend hospital (e.g. clinics, day or inpatient stays). Having identified what patients have given up, one must then put a monetary value on it. Economists take society's valuation of the cost of time, rather than the individual's. However, the issues of estimating how much time a lay carer spends providing care and costing it, and costing the opportunity cost of carers' and patients' time, are complex and unresolved (see earlier).

Study methods used for costings

In relation to studies of costs and effectiveness, health economists use the full range of research methods and techniques to obtain cost data in relation to the unit of study. Gosden *et al.* (1997), in their study of the cost-effectiveness of specialists' outreach clinics in general practice, in comparison with specialists' hospital outpatient clinics, used a before–after study method, with cases (specialist outreach patients) and controls (hospital outpatients), and designed self-completion questionnaires as instruments for the collection of data from the patients, the doctors, the practices and hospital managers. In some cases, economists obtain their data by undertaking document research (e.g. they access and analyse billing records in private health care systems); but there is still the problem of how to standardise costs across providers to facilitate comparisons.

Summary of main points

- The underlying assumption of economics is that the resources available to society are scarce, and decisions have to be made about their best use.
- Health economists use the basic economic concepts of demand, supply and utility in their analyses.
- Cost-effectiveness is an approach to the assessment of efficiency that compares the monetary costs of different projects that produce the same kinds of non-monetary benefits.
- A cost–benefit analysis assigns a monetary value to the benefits of a project and compares this with the monetary costs of the project.
- The marginal cost is the additional cost of producing one extra unit of output.
- The opportunity cost refers to the opportunity lost (benefit forgone), when resources are spent, for spending them in their best alternative way.
- Discounting standardises different cost–time profiles (future benefits are valued less than current benefits).
- Cost–utility analysis relates the cost of the project to a measure of its usefulness of outcome (utility).
- Cost–utility analysis is based on an index of health status in relation to output (e.g. the QALY).
- The QALY attempts to combine quantity and quality of life into a single index, for use in making comparative assessments about the effectiveness of different treatments. Costs of the treatment per QALY are calculated and generally presented in QALY league tables.

Key questions

1 What are the underlying assumptions of economic analysis?
2 Distinguish between demand, supply and utility.
3 What are cost-effectiveness and cost–benefit studies?
4 Explain opportunity cost.
5 What is discounting?
6 Describe cost–utility analysis.
7 What is a QALY?
8 What are the main techniques used to develop QALYs?

Key terms

cost–benefit
cost-effectiveness
cost–utility
demand
discounting
economic appraisal
Index of Well-being Scale
marginal cost
opportunity cost
QALYs
rating scale
standard gamble
supply
time trade-off
utility

Recommended reading

Brown, J. (2005) Economic evaluation, in A. Bowling and S. Ebrahim (eds) *Handbook of Health Research Methods: Investigation, Measurement and Analysis*. Maidenhead: Open University Press.

Drummond, M. (1994) *Economic Analyses Alongside Controlled Trials*. Leeds: NHS Executive, Department of Health.

Fox-Rushby, J. and Cairns, J. (2005) *Economic Evaluation*. Maidenhead: Open University Press.

Locket, T. (1996) *Health Economics for the Uninitiated*. Oxford: Radcliffe Medical Press.

Mooney, G. (1992) *Economics, Medicine and Health Care*. London: Harvester Wheatsheaf.

Netten, A. and Beecham, J. (eds) (1993) *Costing Community Care: Theory and Practice*. Aldershot: Ashgate Publishing.