THEORY AND METHODS OF ECONOMIC EVALUATION OF HEALTH CARE

Developments in Health Economics and Public Policy

VOLUME 4

The titles published in this series are listed at the end of this volume.

THEORY AND METHODS OF ECONOMIC EVALUATION OF HEALTH CARE

by

Magnus Johannesson
Stockholm School of Economics

SPRINGER-SCIENCE+BUSINESS MEDIA, B.V.

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN 978-1-4419-4757-4 ISBN 978-1-4757-6822-0 (eBook)
DOI 10.1007/978-1-4757-6822-0

Printed on acid-free paper

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 10

LIST OF TABLES

PREFACE

Most economic evaluations of health care programmes at the moment are cost-effectiveness and cost-utility analyses. The problem with these methods is that their theoretical foundations are unclear. This has led to confusion about how to define the costs and health effects and how to interpret the results of these studies. In the environmental and traffic safety fields it is instead common to carry out traditional cost-benefit analyses of health improving programmes. This striking difference in how health programmes are assessed in different fields was the original motivation for writing this book. The aim of the book is to try and provide a coherent framework within cost-benefit analysis and welfare economics for the different methods of economic evaluation in the health care field. The book is written in an easily accessible manner and several examples of applications of the different methods are provided. It is my hope that it will be useful both for teaching purposes and as a guide for practitioners in the field.

Glenn C. Blomquist, John D. Graham, Rich O'Conor and four anonymous referees provided helpful comments on previous versions of the manuscript. I would also like to express my gratitude to the following persons for helping me to prepare the manuscript: Carl-Magnus Berglund, Carin Blanksvärd, Ann Brown, and Ziad Obeid. Financial support was received from the Harvard Center for Risk Analysis, the Swedish Council for Social Research and the National Corporation of Swedish Pharmacies. Needless to say, all remaining errors are my sole responsibility.

1. INTRODUCTION

Cost-benefit analysis is an analytic tool designed to promote economic efficiency in the allocation of scarce resources to public projects or technologies. The criterion of economic efficiency states that if those citizens who benefit from a project had to bear its entire cost, they would consider it worth paying for. Where projects are inefficient, those who benefit would reject the project if they were required to pay for it in its entirety.

The intellectual roots of cost-benefit analysis have been traced to the writings of Jules Dupuit (1844), a 19th century French economist. The ethical underpinnings of economic efficiency were further refined by the Italian social scientist Vilfredo Pareto and the British economists Nicholas Kaldor (1939) and Sir John Hicks (1939,1941). By the end of World War II, an entire field called welfare economics was being developed in economics and political science to define the optimal allocation of a society's scarce resources. It was not until 1971, however, that the British economist E.J. Mishan authored the first comprehensive textbook on the subject entitled Cost-Benefit Analysis.

The logic of economic efficiency has been used for decades. In one of the earliest laws to promote efficiency, the USA's Flood Control Act of 1936 states that projects shall be considered for congressional action only if "the benefits to whomever they accrue exceed their costs." Of all sectors of economic activity, the use of formal cost-benefit analysis is probably most widespread in the transportation sector.

The phrase "economic evaluation" includes cost-benefit analysis and related tools such as cost-effectiveness analysis, cost-utility analysis, and cost-minimization studies (Drummond et al 1987). The distinctive feature of cost-benefit analysis is the explicit effort to express all the benefits and costs of a project in a common unit (usually a monetary unit). As we shall see, the related tools of economic evaluation are not as ambitious precisely because they do not attempt to collapse all programme consequences into a single measure of value.

Interest in economic evaluation of health-related programmes began in the 1960s. Economist Burton Weisbrod (1961) of the University of Wisconsin conducted some of the pioneering economic evaluations of public health programmes such as vaccination of children against measles. Statistician Dorothy Rice (1967) of the USA's Social Security Administration published some of the first estimates of the economic costs of illness, which were defined as the sum of resources consumed in treatment and foregone economic production. In accordance with the cost of illness studies, early economic evaluations of medicine and public health defined health-related benefits as: the estimated reductions in subsequent treatment costs plus increases in production due to improved health. The renewed interest in this approach called the "human capital" approach, which can be traced back to the writings of Sir William Petty in the

seventeenth century (see Dublin & Lotka (1945) for a survey), can be explained by the development of the theory of human capital during the 1960s (Becker 1964).

The limitations of the human capital approach were, however, soon realized. In 1968 economist Thomas Schelling argued that the theoretical foundation of the approach is not rooted in the proper concept of willingness to pay, as defined by Kaldor and Hicks. Schelling argued, moreover, that the intrinsic value of good health and quality of life is ignored in human capital calculations. A disturbing consequence of using the human-capital approach was that the method assigns very small values to the health of poor people and those who are not in the labour force. While some economists argued that human-capital estimates might serve as valid lower bounds on the correct numbers, even this suggestion did not withstand careful technical scrutiny (Berger et al 1987; Rosen 1988).

Interestingly, the weaknesses of the human capital approach led to two quite different responses: (1) the development of methods to measure willingness to pay for improved health (Acton 1973), and (2) the development of cost-effectiveness analysis (Klarman et al 1968). The latter response has proved to be more influential in medicine, while the former response has been influential in some aspects of public health such as environmental protection.

In cost-effectiveness analysis, costs are measured in monetary terms and health effects are measured in non-monetary units. The most common effectiveness measure is life-years gained. One of the main challenges for cost-effectiveness analysis has been to develop a single outcome measure that incorporates information about the quality as well as the length of life. The notion of quality-adjusted life-years (QALYs), where life-years are multiplied by a weight reflecting quality considerations, was developed for this purpose (Bush et al 1973; Weinstein & Stason 1976). When QALYs or related utility concepts are used as the outcome measure in cost-effectiveness analysis (CEA), the method is frequently referred to as cost-utility analysis (CUA).

Although the development of CEA and CUA were important steps forward, these methods are not free of problems. It has proved difficult to construct an outcome measure that is consistent with individual preferences and practical to measure. Perhaps a more severe limitation is that CEA and CUA cannot determine whether a project is efficient or worthwhile, i.e., whether benefits exceed costs. It is possible only to compare the cost-effectiveness ratios of various options.

While cost-benefit analysis was proposed from a societal perspective, CEA and CUA are often referred to as the "decision-maker approach" to economic evaluation, indicating that the aim is to maximise whatever the decision-maker wants to maximise (Sugden & Williams 1978; Williams 1985). It is, however, rare that a specific decision-maker is actually identified. This creates confusion about what costs should be specified in the numerator of the ratio and what effects should be taken into account in the denominator.

In response to the limitations of the human capital method and the problems with CEA and CUA, some analysts have developed methods for explicitly quantifying the monetary value of health gains and losses. In these applications, it is accepted that the appropriate benefit measure is the amount of money individuals are willing to forego in order to obtain a specific health improvement. Methods have been developed to infer willingness to pay from either actual choices involving health risks or hypothetical choices in surveys.

The most common market where revealed preference has been studied is the labour market, where wage premiums are offered to induce workers to accept more risky jobs. The classic work has been performed here by the U.S. economists Rosen, Thaler, Smith and Viscusi (Smith 1974; Thaler & Rosen 1976; Viscusi 1978). Other actual safety choices that have been analysed include whether to use automobile safety belts (Blomquist 1979), whether to purchase smoke-detectors (Dardis 1980), whether to purchase homes in safer and cleaner neighborhoods (Portney 1981), whether to purchase new cars with superior safety features (Atkinson & Halvorsen 1990), and whether to undertake steps to reduce indoor radon exposures (Åkerman et al 1991).

The expressed preference approach refers to the use of survey methods to obtain willingness-to-pay estimates from respondents. It is also called the contingent valuation method (Cummings et al 1986; Mitchell & Carson 1989). Jan Acton of the Rand Corporation (1973) was the first to use this method in health care when he assessed the willingness of individuals to pay for mobile coronary care units that would reduce the risk of dying after a heart attack. The British economist M.W. Jones-Lee (1976) later applied this technique to estimate how much airplane passengers might be willing to pay for measures to achieve enhanced airline safety. More recently, the method has been used to assess willingness to pay for traffic safety (Jones-Lee et al 1985), worker safety (Gerking et al 1988) and the health benefits of improved air quality (Viscusi et al 1991). The contingent valuation method has also been used extensively in environmental economics to assess the ecological benefits of environmental protection such as the preservation of a recreational area (Daubert & Young 1981) and the non-use value of pristine water quality in Alaska (Carson et al 1992).

The development of willingness-to-pay methods in environmental economics has also led to renewed interest in the application of cost-benefit analysis to health care (Tolley et al, 1994). A small number of studies have used the contingent valuation method to assess the value of health care programmes such as the treatment of hypertension (Johannesson et al 1991,1993).

In this book the theory and methods of economic evaluation of health care programmes are described. The methods are also illustrated with applications from the literature. The book provides a coherent framework within cost-benefit analysis, for the different methods of economic evaluation. We begin in Chapter 2 by discussing

the foundations for cost-benefit analysis in the branch of microeconomics called welfare economics. We then define the costs and benefits of health care programmes and it is shown how the revealed preference approach and the expressed preference approach can be used to estimate the value of health changes in monetary terms. The monetary benefits also have to be supplemented by estimations of the societal costs of the programmes in order to carry out cost-benefit analysis, and it is shown how these costs can be estimated. Thereafter we outline the principles of cost-effectiveness analysis. The decision rules of cost-effectiveness analysis are demonstrated and cost-effectiveness analysis and cost-benefit analysis are compared. The next chapter is about cost-utility analysis. The relationship between different outcome measures such as QALYs and HYEs and individual preferences are analysed and cost-utility analysis and cost-benefit analysis are compared. Finally, the relationship between economic evaluation and health policy is briefly considered before the book ends with some concluding remarks.

REFERENCES

Acton JP. Evaluating public programmes to save lives: the case of heart attacks. Santa Monica: The Rand Corporation, RAND Report R- 950-RC, 1973.

Åkerman J, Johnson FR, Bergman L. Paying for safety: voluntary reduction of residential radon risks. Land Economics 1991;67:435-446.

Atkinson SE, Halvorsen R. The valuation of risks to life: evidence from the market for automobiles. Review of Economics and Statistics 1990;72:133-136.

Becker GS. Human Capital. Chicago: University of Chicago Press, 1964.

Berger MC, Blomquist GC, Kenkel D, Tolley GS. Valuing changes in health risks: a comparison of alternative measures. Southern Economic Journal 1987;53:967-84.

Blomquist G. Value of life saving: implications of consumption activity. Journal of Political Economy 1979;87:540-558.

Bush JW, Chen M, Patrick DL. Cost-effectiveness using a health status index: analysis of the New York State PKU screening programme. In Berg R (Ed.). Health status indexes. Chicago: Hospital Research and Educational Trust, 1973.

Carson RT, Mitchell RC, Hanemann MW, Kopp RJ, Presser S, Ruud PA. A contingent valuation study of lost passive use values resulting from the Exxon Valdez Oil Spill. Report to the General Attorney of the State of Alaska, 1992.

Cummings RG, Brookshire DS, Schulze WD. Valuing environmental goods. New Jersey: Rowman and Allanheld, 1986.

Dardis R. The value of a life: new evidence from the marketplace. American Economic Review 1980;70:1077-1082.

Daubert JT, Young RA. Recreational demands for maintaining instream flows: a contingent valuation approach. American Journal of Agricultural Economics 1981;63:666-676.

Dublin LI, Lotka AJ. The Money Value of a Man. New York: Ronald Press, 1946.

Dupuit JH. De la mesure de l'utilité des travaux publics. Annales des Ponts et Chaussees 1844. Translated by Barback RH in International Economic Papers 1952;17:83-110.

Drummond MF, Stoddard GL, Torrance GW. Methods for the economic evaluation of health care programmes. Oxford: Oxford Medical Publications, 1987.

Gerking S, de Haan MH, Schulze W. The marginal value of job safety: a contingent valuation study. Journal of Risk and Uncertainty 1988;1:185-189.

Hicks JR. The foundation of welfare economics. Economic Journal 1939;49;696-712.

Hicks JR. The four consumer's surpluses. Review of Economic Studies 1941;11:31-41.

Johannesson M, Jönsson B, Borgquist L. Willingness to pay for antihypertensive therapy: results of a Swedish pilot study. Journal of Health Economics 1991;10:461-474.

Johannesson M, Johansson P-O, Kriström B, Gerdtham U-G. Willingness to pay for antihypertensive therapy: further results. Journal of Health Economics 1993;12:95-108.

Jones-Lee MW. The value of life: an economic analysis. London: Martin Robertson, 1976.

Jones-Lee MW, Hammerton M, Philips PR. The value of safety: results of a national sample survey. Economic Journal 1985;95:49-72.

Kaldor N. Welfare propositions of economics and interpersonal comparisons of utility. Economic Journal 1939;49:549-552.

Klarman HE, Francis JOS, Rosenthal G. Cost-effectiveness analysis applied to the treatment of chronic renal disease. Medical Care 1968;6:48-54.

Mishan EJ. Cost-benefit analysis. New York: Praeger, 1971.

Mitchell RC, Carson RT. Using surveys to value public goods: the contingent valuation method. Washington DC: Resources for the Future, 1989.

Portney PR. Housing prices, health effects, and valuing reductions in risk of death. Journal of Environmental Economics and Management 1981;8:72-78.

Rice D. Estimating the cost of illness. American Journal of Public Health 1967;57:424-440.

Rosen S. The value of changes in life-expectancy. Journal of Risk and Uncertainty 1988;1:285-304.

Schelling TC. The life you save may be your own. In Chase SB (Ed.). Problems in public expenditure analysis. Washington D.C.: Brookings Institution, 1968.

Smith RS, The feasibility of an 'injury tax' approach to occupational safety. Law and Contemporary Problems 1974;38:730-744.

Sugden R, Williams A. The principles of practical cost-benefit analysis. Oxford: Oxford University Press, 1978.

Thaler R, Rosen S. The value of saving a life: evidence from the market. In Terleckyj NE (Ed.). Household production and consumption. Cambridge MA.: NBER, 1976.

Tolley GS, Kenkel D, Fabian R. Valuing health for policy: an economic approach. Chicago: University of Chicago Press, 1994.

Viscusi WK. Labor market valuations of life and limb: empirical estimates and policy implications. Public Policy 1978;26:359-386.

Viscusi WK, Magat WA, Huber J. Pricing environmental health risks: survey assessments of risk-risk and risk-dollar trade-offs for chronic bronchitis. Journal of Environmental Economics and Management 1991;21:32-51.

Weisbrod B. Economics of Public Health: Measuring the Impact of Diseases. Philadelphia: University of Pennsylvania Press, 1961.

Weinstein MC, Stason WB. Hypertension: a policy perspective. Cambridge, MA: Harvard University Press, 1976.

Williams A. Economics of coronary artery bypass grafting. British Medical Journal 1985;291:326-329.

2. WELFARE ECONOMICS

Economics is concerned with the allocation of scarce resources. The amount of resources (labour, materials, natural resources, etc) available to a society can be considered as fixed at a given moment of time. This means that choices have to be made concerning how to use these resources, e.g. the amount of education that is provided can only be increased if the production of some other goods is decreased.

In economics it is usual to make a distinction between positive and normative economics. Positive economics is only concerned with analysing the consequences of different changes or policies, without making judgements about the desirability of alternative allocation of resources. In positive economics it is for instance possible to analyse the consequences of a proposed health care reform on the supply of physicians, prices of health care etc, without making any judgement about whether the health care reform should be carried out.

Normative economics on the other hand is concerned with analysing the desirability of different changes or policies, e.g. to judge whether the proposed health care reform should be carried out. Normative economics is often referred to as welfare economics, and ultimately welfare economics is concerned with providing criteria to rank different alternative changes or policies. To be able to rank different policies it is necessary to impose some value judgements.

2.1 The Pareto Principle

The first fundamental value judgement that is made in welfare economics is known as the Pareto principle. The Pareto principle states that a change is desirable if it makes some individual(s) better off without making some other individual(s) worse off. The Pareto principle is usually coupled with the consumer sovereignty principle. The consumer sovereignty principle states that individuals are assumed to be the best judges of their own welfare, i.e. it is the individuals themselves who decide whether they are better off or worse off with a change.

The Pareto principle is related to the working of the market economy. All trade fulfills the Pareto principle, since it is based on voluntary exchange. If there is perfect competition (i.e. no market failure, see below) it can be shown that the economy may attain Pareto optimality (Boadway & Bruce 1984). Pareto optimality is a situation where it is impossible to improve the situation of some individual(s) without making at least one other individual worse off. According to the so called first theorem of welfare economics (Boadway & Bruce 1984), a competitive general equilibrium is under certain assumptions Pareto optimal. A competitive general equilibrium is a situation where we have a set of market prices so that all markets clear, i.e. there is no excess demand or excess supply.

Demand, supply and equilibrium for one market are demonstrated in Figure 1. Figure 1a shows the aggregate demand curve for good X. Each point on the demand curve shows the total quantity that all potential buyers of the good would like to purchase at that price, holding all other factors constant. The slope of the demand curve shows the marginal benefit of additional units of the good measured as the marginal willingness to pay for additional units. The demand curve is usually assumed to slope downwards since the marginal benefit in terms of willingness to pay for additional units is assumed to decline with further units of the good. Since the slope of the demand curve shows the marginal willingness to pay for an additional unit of the good, the area below the demand curve provides a measure of how much the consumers in total are willing to pay for the good (see also chapter 3).

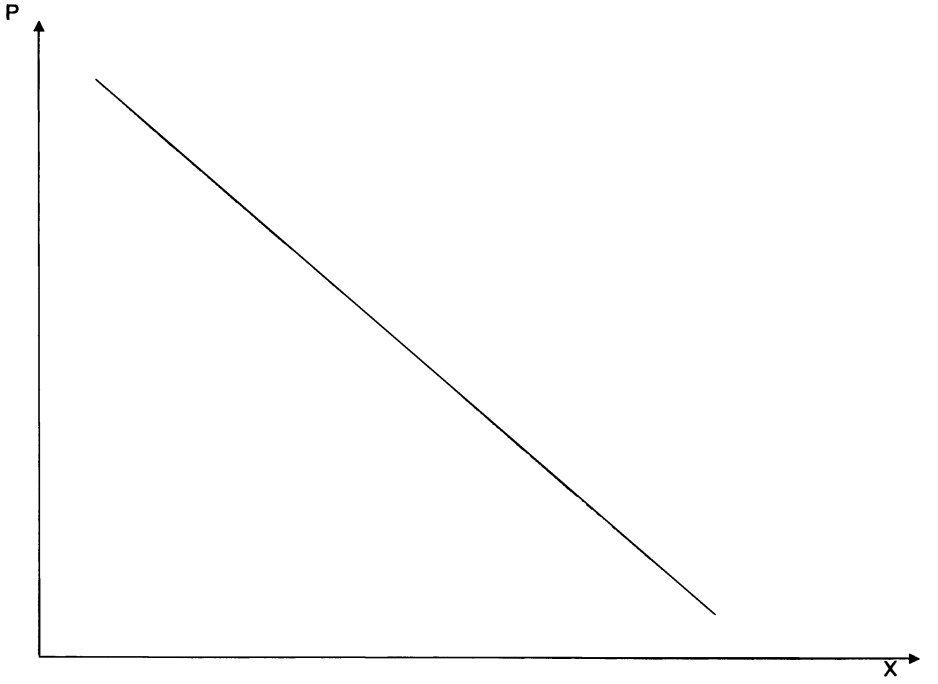

Figure 1a. The market demand curve for a good

Figure 1b depicts the aggregate supply curve for good X. Each point on the supply curve shows the number of units that all suppliers of the good would like to sell at that price, holding all other factors constant. The slope of the supply curve shows the marginal cost of producing additional units of the good, measured as the compensation needed to release the production factors used in producing the good from their best alternative uses. The supply curve is usually assumed to slope upwards since the marginal cost is assumed to increase for the production of additional units of

the good, due to an assumption of diminishing returns to scale in the production of the good.

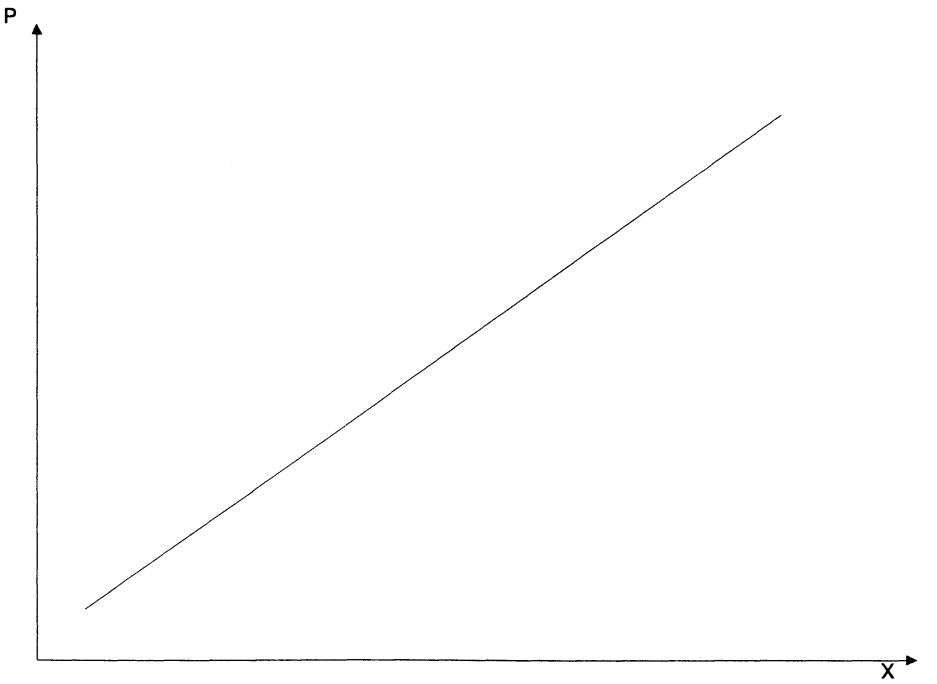

Figure 1b. The market supply curve for a good

In Figure 1c the market demand and the market supply curve are combined. The price P* gives the equilibrium price of the market and the quantity X* gives the equilibrium quantity. At this point the marginal benefit and the marginal cost are equal to the price of the good. If the price was above P* in the figure then the quantity demanded would fall short of the quantity supplied (i.e. excess supply), and market forces would cause the price to fall until the market cleared. Similarly if the price was below P* the quantity supplied would fall short of the quantity demanded (i.e. excess demand) and market forces would cause the price to rise until the market cleared. For more on the concept of a competitive general equilibrium, see Boadway & Bruce (1984).

There are a number of different possible Pareto optimal situations in an economy, and which point is reached in a perfect market economy will depend on the initial distribution of income. The second theorem of welfare economics states that under certain assumptions it is possible to attain any Pareto optimal situation as a result of a competitive general equilibrium given the distribution of income (Boadway & Bruce

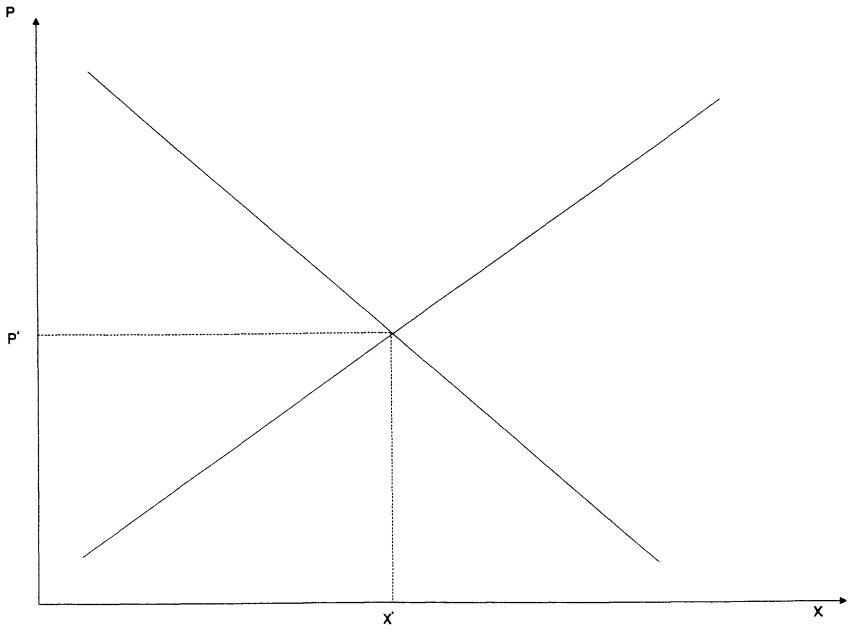

Figure 1C, Equilibrium in a market

1984. If redistribution of income is costless then it is therefore in principle possible to attain any Pareto optimal situation with a perfect market system.

For a situation to be Pareto optimal three sets of conditions must hold (Boadway & Bruce 1984). The first condition is usually referred to as efficient exchange. According to this condition the marginal rate of substitution between two consumer goods, i.e. the rate at which a household is willing to substitute one good for another good, must be the same for all households that consume the goods. If this condition does not hold then households could benefit from further trade. Assume for instance that society only consists of two households, 1 and 2, which consume only two goods, apples and oranges. If household 1 is willing to trade one apple for three oranges and household 2 is willing to trade one apple for one orange, both households would benefit from further trade and trade would continue until the marginal rate of substitution was equal for both households (e.g. until both household 1 and 2 are willing to trade one apple for two oranges).

The second condition is that of efficient allocation of factors. According to this condition the marginal rate of transformation between two goods, i.e. the rate at which one good can be transformed into another by reallocating the supply of a production factor, has to be the same for all production factors. Assume that there are only two production factors, namely labour and materials. According to this efficiency condition

the marginal rate of transformation between apples and oranges has to be the same for labour and materials. If for instance three oranges can be produced at a cost of one apple by transfering labour from apple to orange production, and one orange can be produced at a cost of one apple by transfering materials from apple to orange production, then the production of both oranges and apples can be increased by reallocating the production factors until the marginal rate of transformation is the same for both inputs (e.g. until two oranges can be produced at the cost of one apple for both labour and materials).

The third condition for a Pareto optimum is efficient output choice. According to this condition the marginal rate of substitution between two goods has to equal the marginal rate of transformation between the goods. If for instance households are willing to trade one apple for three oranges and it is possible to produce one apple at the cost of one orange, the production of apples would increase until the marginal rate of substitution equalled the marginal rate of transformation (e.g. until the households were willing to trade one apple for two oranges and it was possible to produce one apple at the cost of two oranges). In a competitive market equilibrium the ratio of the prices of two goods will equal the marginal rate of substitution and the marginal rate of transformation between the goods; if this was not the case there would be excess supply or excess demand and the market would not be in equilibrium.

It is important to note that the Pareto principle says nothing about the distribution of goods, e.g. a Pareto optimal situation may be one where the goods are highly unequally distributed in the economy. Figure 2 depicts the so called utility possibilities frontier in an economy with only two households, 1 and 2. The utility possibilities frontier shows all the Pareto optimal situations that can be reached in the economy given the initial endowment of goods and production factors. If utility is assumed to be only ordinal (i.e. it only ranks alternatives, but says nothing about the strength of preference between different alternatives) it is impossible to assign any specific shape to the utility possibilities frontier, and it has therefore been drawn as neither convex nor concave in Figure 2. The utility possibilities frontier has to slope downward, however, otherwise the utility could be increased for one individual without its being decreased for another individual, and the point would not be Pareto optimal.

The Pareto principle cannot be used to rank different Pareto optimal situations such as points A, B and C in Figure 2, since in order to move from for instance point A to point B one household would gain and another household would lose.

The Pareto principle as such is a rather weak value judgement that is probably acceptable to most people. Note also that a project which increases the income very much for the rich and just a little bit for the poor may be rejected by the Pareto principle if individuals are concerned about the distribution of income in society (i.e. some people feel worse off because of the increased income inequality). If a perfect market system existed and redistribution of income was costless, the role of the public

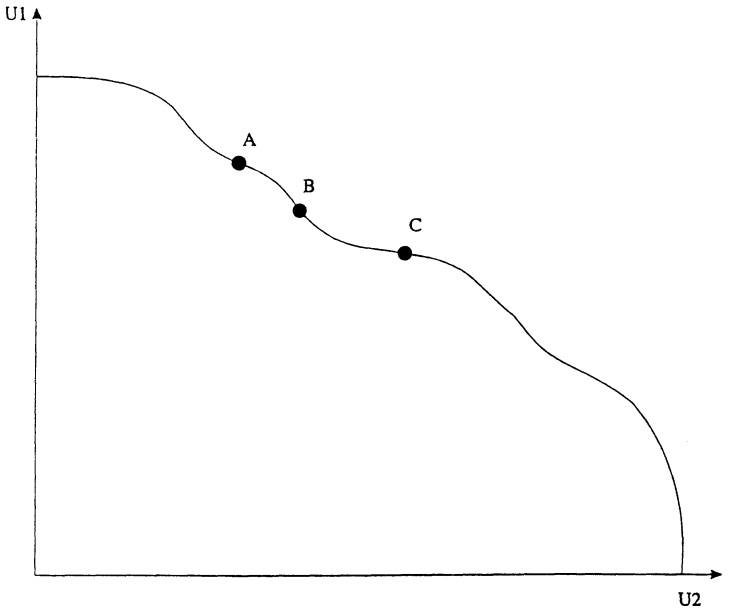

Figure 2. The utility possibilities frontier

sector could be restricted to chosing a point on the utility possibilities frontier (i.e. to redistributing income so that the competitive market equilibrium would lead to that point). In the real world redistribution of income is, however, not costless since it will affect the incentives in the economy in a negative way, e.g. a tax on labour will affect the incentives to work and distort the trade-off between working time and leisure time. Even with a perfect market system it would therefore not be possible for the government to choose any point on the utility possibilities frontier, unless the ownership of the initial endowment of goods and factors in the economy (i.e. the property rights) could somehow be redistributed once and for all.

Unfortunately there exist a number of cases where the market fails to achieve an efficient outcome, i.e. situations where some gains from trade are not exploited. Some of the most common causes of market failure are reviewed below. Market failure provides an argument for public interventions in the market. Such interventions lead to changes where some individuals gain and some other individuals lose, and in those cases the Pareto principle cannot be used to determine whether a change should be carried out or not. Since the majority of policies produce both gainers and losers the Pareto principle is thus of little practical use.

2.2 Market Failures

For a market economy to operate efficiently it has to be assumed that there are a large number of utility maximizing consumers and profit maximizing firms, each unable to affect the going market price, i.e. both firms and consumers act as if the price is given. It also has to be assumed that both consumers and firms have perfect information and that no externalities or public goods exist. Finally, it has to be assumed that firms operate under diminishing returns to scale, that prices are flexible and that markets are complete (i.e. that markets exist for all goods and services). Some of the most common causes of market failure are reviewed below. For a more complete and detailed description of market failure and its welfare consequences, see Boadway & Bruce (1984) and Johansson (1991).

In the case of a monopoly there is only one producer of a specific good. In this case the assumption that firms are price-takers and cannot affect the going market price is violated. In such a case the firm can decide the market price. The monopoly case is depicted in Figure 3. Since the demand curve slopes downward the monopolist must reduce the price to increase the sales. This means that the price on all the sold units will decrease. Because of this the marginal revenue of the last unit is lower than the price, since the monopolist has to deduct the loss in revenue from all other units sold in order to get the marginal revenue. Since the monopolist is assumed to maximise profits, the monopolist will choose the output where marginal cost equals marginal revenue. In the Figure X1 units will be sold at the price P1. The efficient output is X* since at that point the marginal benefits equal the marginal costs. The welfare loss due to the monopoly is equal to the area below the demand curve reflecting the willingness to pay of the consumers minus the area below the marginal cost curve reflecting the cost of producing the good for the move from X* to X1. It should be noted that a monopolist may act as if it was a competitive market if there is free entry to the market, since if profits are made this would induce other firms to enter the market. Such a situation is usually referred to as a contestable market (Baumol 1982).

Increasing returns to scale are also a source of market failure. It is normally assumed that firms operate under diminishing returns to scale, leading to increasing marginal costs and an upward sloping supply curve. If, however, the marginal costs of production decrease then a firm will lose money if the price is set equal to the marginal cost and it will cease to operate on the market. If the area below the demand curve measuring the total willingness to pay of the consumers exceeds the total production costs it would, however, be profitable for society if the good was provided at a price equal to the marginal cost. An example of this could for instance be the construction of a bridge, where the marginal cost of using the bridge once it is constructed is close to zero and a firm which sets the price equal to the marginal cost will make a loss and cease to operate, even though the construction of the bridge may be profitable.

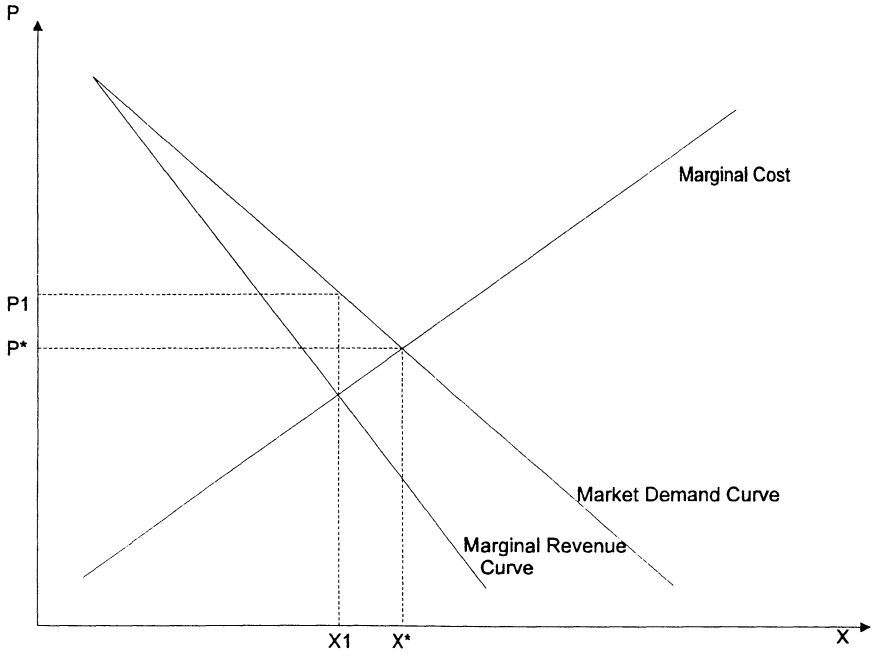

Figure 3. A monopoly

Another source of market failure is the presence of externalities, i.e. costs and/or benefits that are not taken into account by the actors on the market. In Figure 4 the case of a negative externality in production is shown, e.g. a firm that is polluting the air. MC Private is the marginal cost curve for the firm and MC Social is the marginal cost curve for society including the cost of the pollution. The market will result in the quantity X1 and the price P1. From a societal viewpoint, however, the optimum should be set at X* and P* where the marginal benefit equals the marginal social cost. The welfare loss due to the externality is equal to the area below the MC social curve minus the area below the demand curve for a move from X* to X1. The example in Figure 4 is an example of a negative externality, but there can also be positive externalities. An example of a positive externality is the presence of a caring externality for health care (i.e. that individuals for altruistic reasons derive benefits from the consumption of health care by other individuals), which means that the market demand curve will underestimate the benefits of health care.

Public goods also lead to market failure. Pure public goods are goods where no one can be excluded from the consumption of the good and consumers can use the good at zero extra marginal cost. Examples of public goods are cleaner air and defence. In a market for private goods the marginal benefit of each consumer is set equal to the price at the efficient outcome. For public goods, efficiency is reached at the point

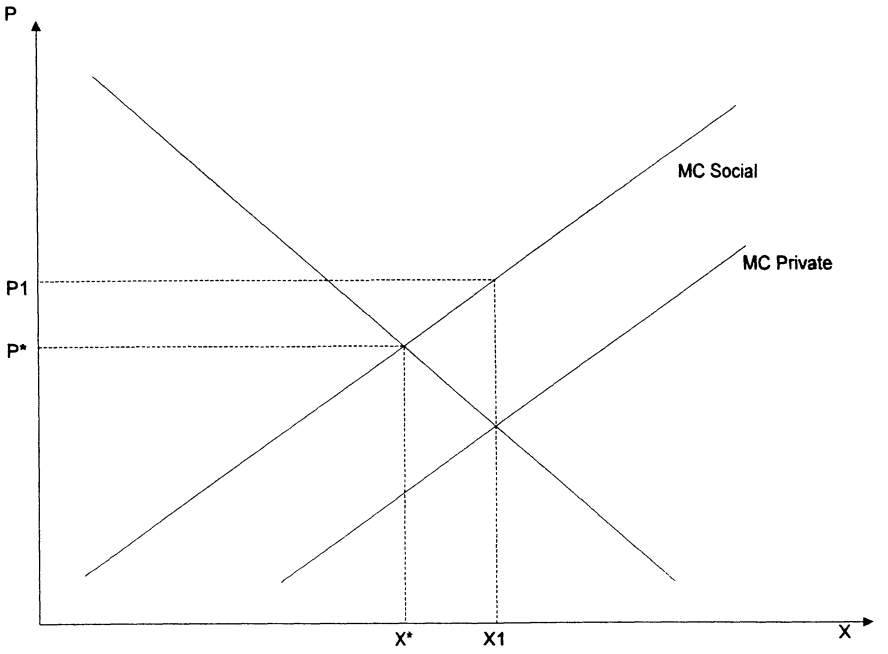

Figure 4. A negative externality in production

where the sum of the marginal valuations of each consumer equals the marginal cost of providing the public good. Markets will therefore undersupply public goods and will not be efficient.

A final source of market failure to be considered here is when the assumption of perfect information does not hold. In for example the health care field, imperfect information is an important source of market failure. An example of imperfect information is shown in Figure 5. In Figure 5 it is assumed that the consumers of health care exaggerate the benefits of health care due to for instance false advertising; the curve D Imperfect shows the demand curve based on the imperfect information and D Perfect shows the demand curve based on correct information about the benefits of health care. The market outcome in Figure 5 is quantity X1 and price P1. This should be compared with the outcome X*, P* based on correct information. The welfare loss due to the imperfect information is equal to the area below the marginal cost curve minus the area below the demand curve for full information for the move from X* to X1. In the health care case imperfect information may also lead to private insurance markets not operating efficiently, or not existing at all. For more on the specific nature of the health care market see Phelps (1992).

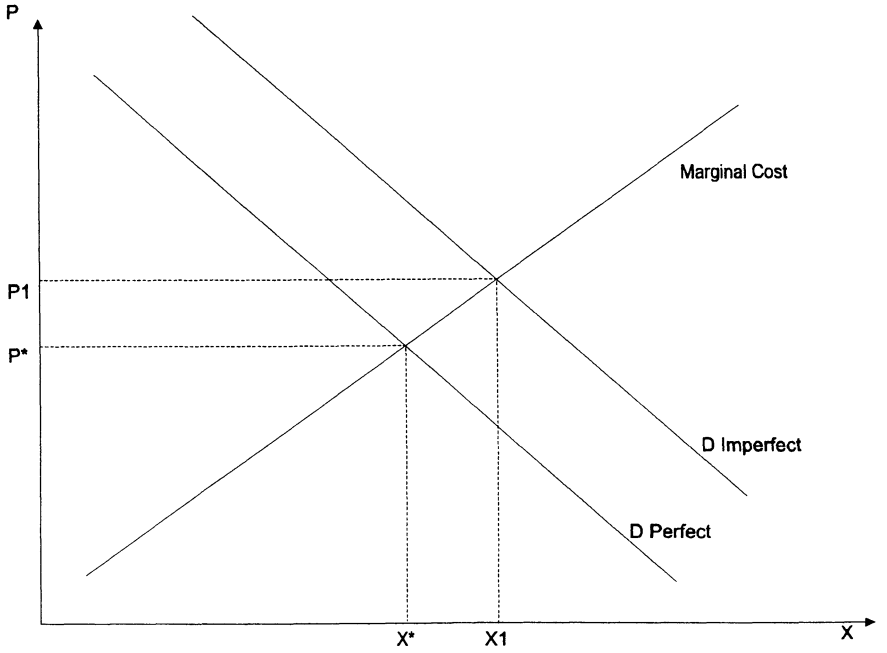

Figure 5. Imperfect information

In this chapter we have shown a number of reasons why markets may fail to achieve Pareto efficient outcomes. This means that public interventions could potentially lead to efficiency gains in these cases. It is important to note, however, that the comparisons above are between imperfect markets and the outcome of a perfect market. Since public interventions are by no means perfect it is not obvious that public intervention will lead to improved efficiency even if market failure exists. In the health care field, arguments for public intervention in the health care market are often based on information assymetries and the existence of altruistic externalities.

2.3 The Compensation Principle

Since markets will not always lead to Pareto efficient outcomes, due to market failure, there may be a role for public intervention apart from pure redistribution of income. If redistribution of income was costless it would also be possible to pursue only projects that satisfy the Pareto principle, i.e. losers could always be compensated for their losses through redistribution of income so that nobody loses from a policy change. Since redistribution is not costless, however, applying the Pareto principle would lead to a bias towards status quo.

To overcome the limitations of the Pareto principle, the notions of potential Pareto improvements or compensation tests were introduced. There are two different compensation tests. According to the Kaldor (1939) version a change is desirable if gainers can hypothetically compensate losers and still be better off than without the change. The Hicks (1939) version on the other hand states that a change is desirable if losers cannot hypothetically "bribe" gainers and still be better off than with the change. Unlike the Pareto principle, the compensation principle does not require the actual payment of compensation. If compensation was actually paid then a project would be an actual Pareto improvement. The rationale for the compensation test is that if benefits exceed costs, then if compensation was costless it would be possible to redistribute income so as to achieve an actual Pareto improvement.

In a sense, by focusing on hypothetical redistributions only the efficiency aspects of the project are considered, with efficiency defined as a change where the monetary benefits exceed the monetary costs. Even if benefits exceed costs, however, it is not obvious that gainers can hypothetically compensate losers and still be better off. This interpretation can only be made if compensation does not affect the prices in the economy, since if prices change due to the compensation then gainers may be unable to compensate losers even if benefits exceed the costs of the project; see Boadway & Bruce (1984) for a further discussion of this result.

Since according to the compensation principle compensation is hypothetical, projects that are carried out on the basis of these critera will yield both gainers and losers. Since the gains are measured in monetary terms one objection to the compensation principle is that it discriminates against the poor, since it is based on the existing income distribution. One defence of the compensation principle has been that even though compensation is not actually paid for each policy change, everybody will gain in the long run by applying the compensation principle (Leonard & Zeckhauser 1986). The validity of this argument is, however, unclear since there is no empirical evidence to support it.

2.4 Cost-Benefit Analysis

The compensation principle is the basis of cost-benefit analysis, where a project is considered to be beneficial if the benefits of the project exceed the costs. The benefits are defined as the amount of money the gainers are willing to pay to make sure that a project is carried out (willingness to pay, WTP). The costs are defined as the amount of compensation the losers require to accept that the project is carried out (willingness to sell, WTS). This is therefore consistent with the Kaldor version of the compensation test, although it would also be possible to define cost-benefit analysis based on the Hicks version of the compensation test (i.e. the WTS of the gainers to accept that the project is not carried out versus the WTP of the losers to make sure that the project is not carried out). Using the Kaldor version seems most intuitively appealing since this

is similar to the working of the market, where individuals have property rights and are compensated for giving something up.

The net benefit of a programme can then be defined as the sum of the WTP of the gainers minus the sum of the WTS of the losers according to the Kaldor version of the compensation principle. The decision rule in cost-benefit analysis is that a project should be carried out if the net benefits are positive (i.e. benefits exceed costs). The size of the net benefits also shows the net gain to society, which means that programmes can be ranked according to their net benefits. The ranking of programmes is important if there are mutually exclusive programmes, i.e. it is only possible to carry out one of a number of programmes, e.g. different alternative drugs for the same patient group (for more on mutually exclusive programmes see the chapter about cost-effectiveness analysis). If programmes are independent (i.e. if the costs and benefits of programme A (B) are not affected by whether programme B (A) is carried out, programmes A and B are independent), the ranking of programmes is not important since all programmes with positive net benefits should be carried out. The net benefits, however, show the size of the efficiency gain of different programmes.

It is a well known problem in cost-benefit analysis that benefit-cost ratios cannot be used to rank programmes on the basis of net benefits. The ratio is affected by whether for instance reduced costs are entered as a cost saving or as a benefit. Consider for instance a project that yields $30,000 in benefits, costs $20,000 and leads to reduced morbidity costs of $10,000. Entering the reduced costs as a benefit gives the benefit-cost ratio of 2 ([30,000+10,000]/20,000) and entering the cost savings as reduced costs gives a benefit-cost ratio of 3 (30,000/[20,000-10,000]); thus the ratios differ.

This means that benefit-cost ratios should not be interpreted as a ranking of projects, but they should only be used to determine whether benefits exceed costs or not, i.e. whether the ratio is above or below 1. This is also all that is needed for independent programmes if we do not face a budget constraint, since all independent programmes with positive net benefits should be carried out according to the decision rule.

If, however, a budget constraint exists then it may be the case that not all programmes with positive net benefits can be carried out. In such a case the net benefits should be maximised for the available budget. If a real budget constraint exists it is not obvious that a ranking of programmes according to net benefits will maximise the net benefits. This can be easily shown by an example. Assume that a budget of $10,000 is available and that we can choose between three independent programmes. Programme A uses all the $10,000 of the budget and has other costs and benefits of +$15,000, thus resulting in $5,000 in net benefits. Programme B uses $5,000 of the budget and has other costs and benefits of +$9,000, thus resulting in $4,000 in net benefits. Programme C also uses $5,000 of the budget and has other costs and benefits of +$8,000, thus resulting in net benefits of $3,000. According to net benefits programme

A should be carried out first, and since this programme uses all the budget a total net benefit of $5,000 results. However, if the budget had been spent on programmes B and C instead the net benefits would have been $8,000 (3,000+5,000). Thus a ranking according to net benefits does not result in the maximisation of net benefits with the budget. The problem is that the cost of a dollar from the budget has a higher opportunity cost than one dollar, due to the budget constraint.

If a real budget constraint exists then cost-benefit ratios may be useful (we use cost-benefit ratios rather than benefit-cost ratios here to make it equivalent to the estimation of cost-effectiveness ratios in the chapter on cost-effectiveness analysis). The cost-benefit ratios of different programmes can then be estimated, with the costs paid by the budget defined as costs and all other costs and benefits defined as benefits. By ordering programmes according to their cost-benefit ratios (with the incremental cost-benefit ratios calculated for mutually exclusive programmes, i.e. the increased costs divided by the increased benefits, see chapter 9 about cost-effectiveness analysis), programmes can be implemented in order of these ratios until the budget is exhausted. Such an approach would lead to the maximisation of the net benefits for the budget constraint, if the cost-benefit ratios of the different programmes were independent of the scale of the programmes (i.e. constant returns to scale) (Baumol 1972).

Alternatively if the assumption of constant returns to scale does not hold, non-linear programming techniques can be used to maximise net benefits (Winston 1991). In practice it will probably be difficult to use this kind of budget optimisation since it implies a knowledge of all the costs and benefits of the programmes of interest for a specific budget (and all their scale effects if this assumption is relaxed and non-linear programming is employed).

Using cost-benefit analysis based on the compensation principle implies that one dollar is given the same weight for everyone in society. In principle it is, however, possible to weight the costs and benefits for different people using the concept of a social welfare function to obtain a measure of the overall change in social welfare. The concept of a social welfare function is reviewed in the next section.

2.5 Social Welfare Functions

A complete ranking of all social states is sometimes referred to as a social welfare function. Such a function is simply a function of the utility of all households in society. If we assume that there are only two households in society the function can simply be written as W=f(U1,U2), i.e. social welfare is a function of the utility levels of the two households. Such a welfare function is often referred to as a Bergson/Samuelson social welfare function (Bergson 1938; Samuelson 1947). Social welfare functions are usually assumed to satisfy the property of welfarism, which means that they are only a function of the utility of the households. Social welfare is also assumed to be increasing in each household's utility, so that the Pareto principle is satisfied (e.g. if

the utility of one household increases and the other household's utility does not change, social welfare increases). Furthermore, it is often assumed that it does not matter who gets a high or a low utility (known as anonymity).

In order to be able to define a social welfare function that consistently ranks all social states, strong assumptions have to be made concerning the measurability and comparability of the utility of households. In standard demand theory it is usually assumed that the utility functions of households are ordinal and that the utility levels cannot be compared between households. If that is the case it has been shown by Arrow (1951) that no social welfare function exists that consistently ranks social states (if dictatorship is ruled out, as is done by Arrow). This result is often referred to as Arrow's impossibility theorem.

If, however, full measurability and comparability of utility functions is assumed then a consistent ranking of social states will result (Boadway & Bruce 1984). The functional form of the social welfare function f(U1,U2) still has to be determined, however. If the social welfare function is assumed to be the sum of the utilities of the households (W=U1+U2) we get the utilitarian social welfare function depicted in Figure 6a. All points along each social welfare indifference curve in the figure give the same social welfare.

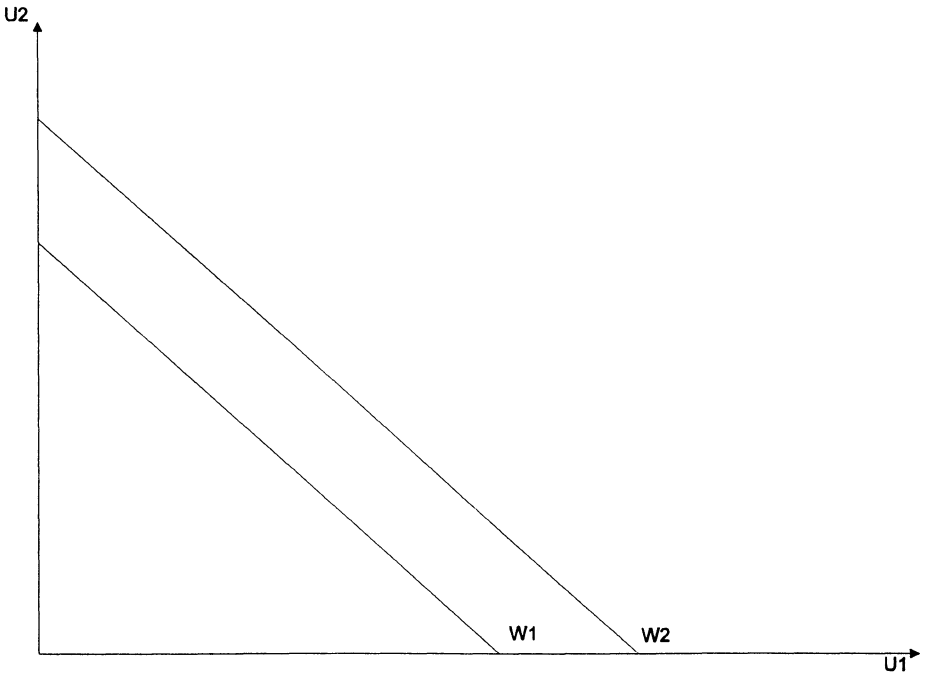

Figure 6a. A utilitarian social welfare function

In figure 6b it is assumed that society is concerned not only about the sum of the utilities, but also about the distribution of utilities as seen by the convex social welfare indifference curves. Such a welfare function is achieved by multiplying the utilities of the households with different welfare weights that depend on the utility level attained, with greater weights for poor households than rich households.

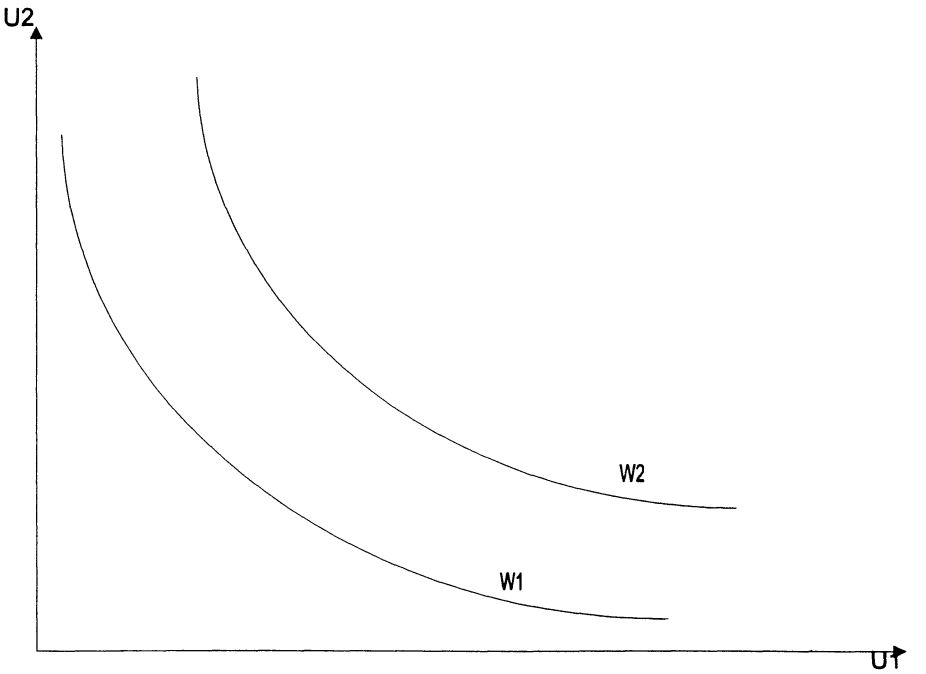

Figure 6b. A social welfare function with aversion towards inequality in utilities

Finally in Figure 6c a Rawlsian social welfare function is depicted, where it is assumed that the social welfare depends only on the utility of the poorest or worst-off household (Rawls 1972).

If the social welfare function is known, social welfare can be maximised in society by reaching the highest social welfare indifference curve given the endowment of goods and production factors. This case is depicted in Figure 7 where a utilitarian social welfare function is assumed and the maximum is reached when a social welfare indifference curve is tangential to the utility possibilities frontier, leading to utility level U1* for household 1 and utility level U2* for household 2. In Figure 7, the utility possibilities frontier has been drawn as convex in contrast to Figure 2, since full measurability and comparability of utility functions is now assumed.

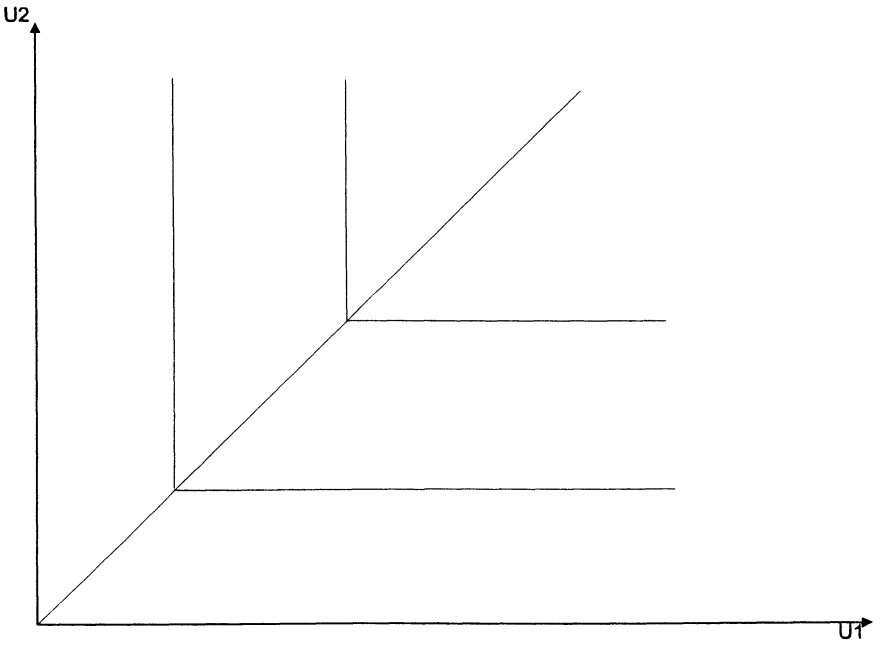

Figure 6c. A Rawlsian social welfare function

If cost-benefit analysis is interpreted in terms of a utilitarian social welfare function it has to be assumed that the social marginal utility of income is equal for all households in society, which would be the case if we are at a social welfare optimum with an optimal distribution of income (Johansson 1995). If the social marginal utility of income is known, the monetary costs and benefits can also be converted to utility to get a measure of the change in social welfare due to a project. However, at the moment there is no generally agreed way of estimating the social marginal utility of income of different individuals or determining the shape of the social welfare function.

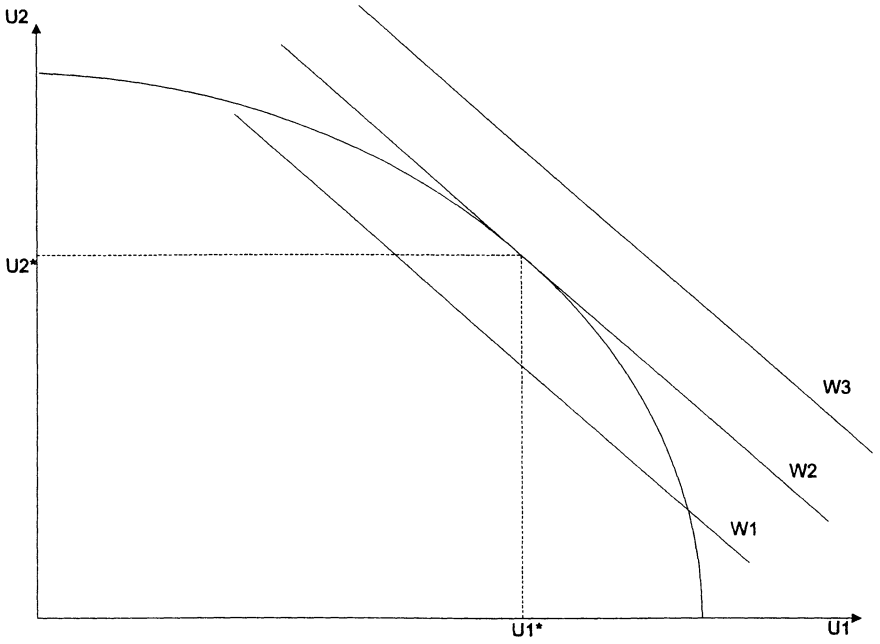

Figure 7. Social welfare optimum

REFERENCES

Arrow KJ. Social choice and individual values. New Haven Connecticut: Yale University Press, 1951.

Baumol WJ. Economic theory and operations analysis. 3rd edition. London: Prentice Hall, 1972.

Baumol WJ. Contestable markets: an uprising in the theory of industry structure. American Economic Review 1982;72:1-15.

Bergson A. A reformulation of certain aspects of welfare economics. Quarterly Journal of Economics 1938;52:310-334.

Boadway RW, Bruce N. Welfare economics. Oxford: Blackwell, 1984.

Hicks JR. The foundation of welfare economics. Economic Journal 1939;49;696-712.

Johansson P-O. An introduction to modern welfare economics. Cambridge: Cambridge University Press, 1991.

Johansson P-O. Evaluating health risks: an economic approach. Cambridge: Cambridge University Press, 1995.

Kaldor N. Welfare propositions of economics and interpersonal comparisons of utility. Economic Journal 1939;49:549-552.

Leonard HB, Zeckhauser RJ. Cost-benefit analysis applied to risks: its philosophy and legitimacy. In MacLean D (Ed.). Values at risk. Totowa N.J.: Rowman & Allanheld, 1986.

Phelps CE. Health Economics. New York: HarperCollins, 1992

Rawls J. A theory of justice. Cambridge MA.: Harvard University Press, 1971.

Samuelson PA. Foundations of economic analysis. Cambridge MA.: Harvard University Press, 1947.

Winston WL. Operations research: applications and algorithms. Second edition. Boston: PWS-Kent Publishing Company, 1991.

3. THE MONETARY VALUE OF HEALTH CHANGES

Cost-benefit analysis is based on the summation of the monetary valuations of the changes due to a programme among all the affected individuals (Boadway & Bruce 1984; Johansson 1991). This implies that the costs and benefits to all individuals affected by a health care programme should be included.

The costs and benefits of a health care programme can be divided into the costs and benefits to the recipients of the health care programme, and the costs and benefits arising for other people due to the programme. The costs and benefits for other people can further be divided into the external costs and a caring externality (altruism). The external costs are the effect on the consumption of people other than the individual who receives the health care programme, and the caring externality is the altruistic value of changes in other people's health status.

In principle, in cost-benefit analysis the benefits of a programme refer to the monetary valuation of the change among the gainers of a programme while the costs refer to the monetary valuation of the change among the losers of a programme. In practice, however, both the terms costs and benefits are used to reflect both positive changes (i.e. gains) and negative changes (i.e. losses). For instance, the cost term can be used to reflect both the fact that increased resources are used in a treatment programme (e.g. the cost of drugs for the treatment of high blood pressure) and the fact that the programme leads to less resources being used in treating a disease (e.g. the decreased costs of treating heart attacks due to the treatment of high blood pressure). Similarly, the benefits term can be used to reflect for instance both health gains (i.e. increased life-expectancy due to the treatment of high blood pressure) and health losses (i.e. reduced health gains from blood pressure treatment due to side-effects).

In this book we will make a distinction between the different types of change that result from a health care programme and between different recipients of the consequences of a health care programme. The changes due to a programme will be divided into the effects on health status per se and the resource consequences of the programme. The resource consequences are defined as the effect on the consumption of goods and services and the effect on the consumption of leisure. For both the health consequences and the resource consequences a distinction can be made between the recipients of these consequences. The health consequences can be divided into the value of the health change for the recipient of the health care programme (e.g. the patients who receive a blood pressure treatment) and an altruistic externality (e.g. the fact that the utility of individual A depends on the health status of individual B).

The resource consequences can similarly be divided into the resource consequences of the recipient of a health care programme and the external costs (i.e. the effect on the consumption of people other than the recipients of the health care programme). The size of the external costs will depend on the institutional arrangements in society.

In this chapter we will analyse the value of health changes; the resource consequences of a health care programme are analysed in the next chapter. Below we first analyse the value of the health change for the individual who receives the health care programme and then consider the value of the health change for other people in society, i.e. the altruistic externality.

3.1 Private WTP

We will start with the simplest possible model to define the private WTP for a health change. Assume that the utility of an individual depends on the consumption of private (non-health) goods, and the health of the individual. This leads to the following utility function:

$$U=U(C,H) \quad (1)$$

In equation 1, C is the consumption of private non-health goods and H is the health of the individual. The health of the individual is here assumed to be exogenous to the individual, which means that we do not include the possibility that the individual can produce health. The individual is assumed to maximise utility subject to the budget constraint Y-PC=0, where P is the price of non-health goods and Y is assumed to be income after tax (all the costs of health programmes are assumed to be covered by a tax). Indirectly the utility of the individual can therefore be written as a function of income and the price of non-health goods. This gives the following indirect utility function (Johansson 1995):

$$V=V(Y,P,H) \quad (2)$$

The indirect utility function can be used to define monetary measures of health changes. It is possible to define the willingness to pay (WTP) and the willingness to sell (WTS) for a health change by holding either the utility level before the change in health or the utility level after the change in health constant. If the utility level before the change in health is held constant we investigate the WTP for an improvement in health and the WTS to accept a deterioration in health. These WTP and WTS measures are usually referred to as compensating variation (Hicks 1941). If the utility level after the change in health is held constant we investigate the WTP to avoid a deterioration in health and the WTS to accept that a health improvement is not carried out. These WTP and WTS measures are usually referred to as equivalent variation (Hicks 1941).

Using the sum of compensating variations in a cost-benefit analysis corresponds to the Kaldor (1939) version of the compensation principle and using the sum of equivalent variations in a cost-benefit analysis corresponds to the Hicks (1939) version of the compensation principle. The choice between the measures depends on which reference utility level is held constant when defining the measures, the utility level before or

after the change in health. Here we will generally use the compensating variation definitions of WTP and WTS which follow from the market analogy: in a market, individuals are compensated for giving something up and individuals pay to receive something. It should be noted, however, that it would be possible to define WTP and WTS measures in the same way as below by using equivalent variation instead of compensating variation.

To illustrate the definition of the WTP for a health improvement, we can assume that a drug is introduced which improves the health status of an individual from arthritis (HA) to full health (H*). The WTP of this drug for the individual is then defined by the following equality:

$$V(Y-WTP,P,H^*)=V(Y,P,HA) \qquad (3)$$

WTP is the amount of money that, if paid, keeps the individual at the initial utility level (the utility level with arthritis). The WTP of the drug is also illustrated in Figure 1. In Figure 1 the utility function with respect to income is shown for the health states arthritis and full health. The individual is initially in the arthritis state with an income of Y0. In order to define the WTP we determine the income in the healthy state that leads to the same utility as income Y0 with arthritis. This income is Y1 in the Figure and the WTP is equal to the difference between Y0 and Y1.

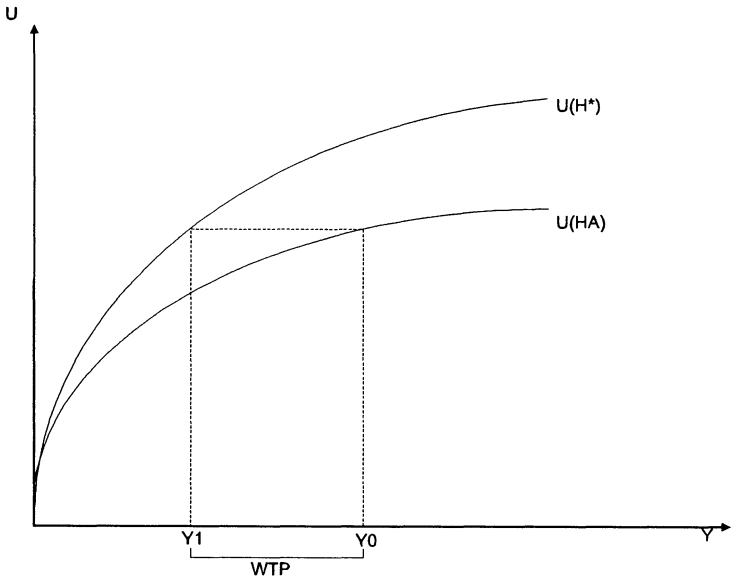

Figure 1. The WTP for a health improvement

To illustrate the definition of the WTS for a health deterioration, we can consider the case where the individual receives the drug and we are analysing the consequences of withdrawing the drug. The WTS of a deterioration in health from full health (H*) to arthritis (HA) is defined by the following equality:

$$V(Y+WTS,P,HA)=V(Y,P,H^*) \qquad (4)$$

WTS is the amount of money that if received keeps the individual at the initial utility level (the utility level with full health). The WTS of giving up the drug is also illustrated in Figure 2. In Figure 2 the utility function with respect to income is shown for the health states arthritis and full health. The individual is initially in the healthy state with an income of Y0. In order to define the WTS we determine the income in the arthritis state that leads to the same utility as income Y0 with full health. This income is Y1 in the figure and the WTS is equal to the difference between Y1 and Y0.

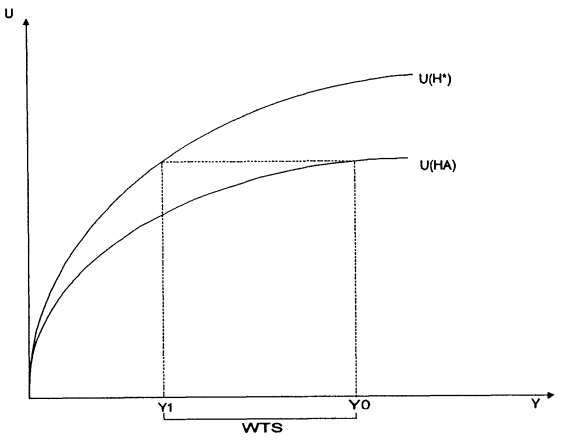

Figure 2. The WTS for a health deterioration

In Figure 3 we have illustrated the WTP of an improvement in health from arthritis to full health by using a marginal WTP curve. The function shows the marginal WTP of additional health improvements with utility held constant at the initial level along the

curve. The WTP for an improvement from arthritis to full health can be estimated as the area below the curve between the arthritis health state (HA) and full health (H*). This curve is sometimes referred to as a compensated demand curve (Hicks 1941, Johansson 1991), since the income is adjusted along the curve so as to hold utility constant.

Sometimes market demand curves (referred to as ordinary demand curves) are used to estimate the WTP of a good. On ordinary demand curves the income rather than the utility is held constant along the curve and because of this the utility of an individual will increase as we move along a market demand curve towards lower prices (called the income effect). Because of the income effect the area below a market demand curve (ordinary consumer surplus) will not give the maximum willingness to pay of an individual in general. It is often used as an approximation of WTP, however, and for small changes this may be a good approximation (Johansson 1993, 1995). It has been shown, however, that in general it is not guaranteed that the change in ordinary consumer surplus will have the same sign as the underlying change in utility (i.e. utility may increase although the ordinary consumer surplus decreases). For more on this issue, see Johansson (1987,1991,1993,1995).

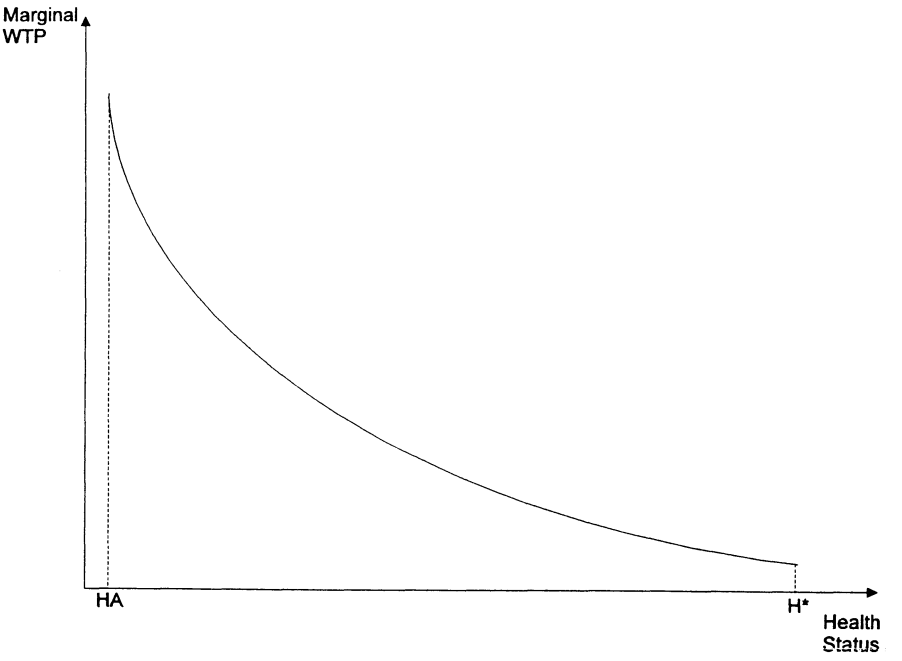

Figure 3. A marginal WTP curve for improved health

As shown by Johansson (1995) the WTP can be converted to the change in utility by multiplying the WTP by the marginal utility of income (all the properties discussed here are valid also for WTS). This means that the WTP will always have the same sign as the change in utility as long as the marginal utility of income is positive, which it will be for a non-satiated individual (i.e. an individual who would like to have a higher income). This also holds in the more general case when all factors of importance for the utility of the individual change (e.g. relative prices and income).

WTP is also what is usually called a path-independent measure, which means that the size of the WTP for a total change is independent of the order in which for instance we change health, prices and income. This also means that the WTP of the different changes can be added together to get the total WTP of a change as long as the WTP of each change is estimated sequentially on the basis of the previous change (i.e. if both income and health change and we estimate the WTP for the change in income first, the additional WTP for the change in health should be estimated on the basis of the individual having already paid the WTP for the change in income). It is important to stress that the WTP of independently assessed changes cannot be added to get the WTP of a combined change.

Sometimes we want to compare more than one final health level with the initial health level (e.g. we want to compare more than one new treatment with the existing treatment or with no treatment). We may for instance want to compare two new drugs (A and B) for arthritis with the current treatment for arthritis. In such a situation it is not guaranteed that the WTP will rank the treatments in the same way as the utility function. The problem with WTP in this case is that it is evaluated at the health, income and price level after the treatment (see the left-hand side of Equation 3 where both WTP and H change and in the more general case income and prices could change as well). Even if the WTP is the same for both drugs A and B in this example, the gain in utility may differ between the drugs if the marginal utility of income differs in the situation with drug A compared to the situation with drug B.

If we are comparing two or more initial health states with one final health state then WTP will always rank the changes in the same way as the utility function. In that case the final health, prices and income levels are the same for all changes but the initial health, prices and income levels differ for the programmes. See Johansson (1995) for more on the ranking properties of WTP.

The treatment so far has been highly unrealistic in the sense that no risk has been assumed. In this section we introduce risk, i.e. probabilities of different events. We use the words risk and uncertainty interchangeably below to refer to the case where the probabilities of different events are known.

In the analysis of situations involving risk we will assume that the individual is an expected utility maximizer. For the axioms underlying the expected utility theory, see von Neumann and Morgenstern (1947). For the certainty case it is enough to assume

that the individual has an ordinal utility function. This means that if the utility function U(C,H) represents the preferences of the individual any other increasing function or monotonic transformation of U(C,H) such as the logarithm of U(C,H) will also represent the preferences in the same way (i.e. both functions will order different combinations of C and H in the same way).

The expected utility theory makes stronger assumptions than ordinality; it is based on a cardinal utility function. If the utility function U(C,H) is cardinal any positive affine transformation of the function, e.g. J(C,H)=a+bU(C,H), where a and b are constants and b>0, will give the same representation of preferences. This makes it possible to define what is usually referred to as risk attitudes, which has to do with the curvature of the utility function.

This is illustrated with respect to income in Figure 4. The utility function with respect to income in Figure 4 (with health and prices held constant) illustrates the concepts risk neutrality, risk aversion, and risk loving (Pratt 1964). If the utility function is linear (Figure 4a) the individual is said to be risk neutral with respect to income (also defined as the case where the second partial derivative of the utility function is 0, i.e. the slope (marginal utility of income) is constant.

If the utility function is concave (Figure 4b) the individual is a risk averter (also defined as the case where the second partial derivative of the utility function is <0, i.e. the slope (marginal utility of income) is decreasing).

Finally, if the utility function is convex (Figure 4c) the individual is a risk lover (also defined as the case where the second partial derivative of the utility function is >0, i.e. the slope (marginal utility of income) is increasing).

According to the expected utility theory the expected utility of situations involving risks can be estimated as the sum of the probabilities of the different possible events (states of the world) multiplied by the utility of the events (von Neumann & Morgenstern 1947). For instance if there is a 50% probability that the individual will have an income of $50,000 leading to the utility level U1 and a 50% probability that the individual will have an income of $100,000 leading to the utility U2, the expected utility of this gamble is equal to 0.5*U1+0.5*U2. A risk neutral individual would be indifferent between this gamble and the expected monetary value of the gamble for sure (i.e. $75,000 for sure), whereas a risk averse individual would prefer the expected value of the gamble from the gamble, and a risk loving individual would prefer the gamble from the expected value. The standard assumption in economics is to assume that individuals are risk averse with respect to income due to decreasing marginal utility of income, i.e. the utility function with respect to income is concave.

We can then proceed to define the monetary value of health changes in a risky world. In principle WTP and WTS can be defined in a similar way as in the certainty case, by defining the amount of money that if paid or received leaves the initial level of

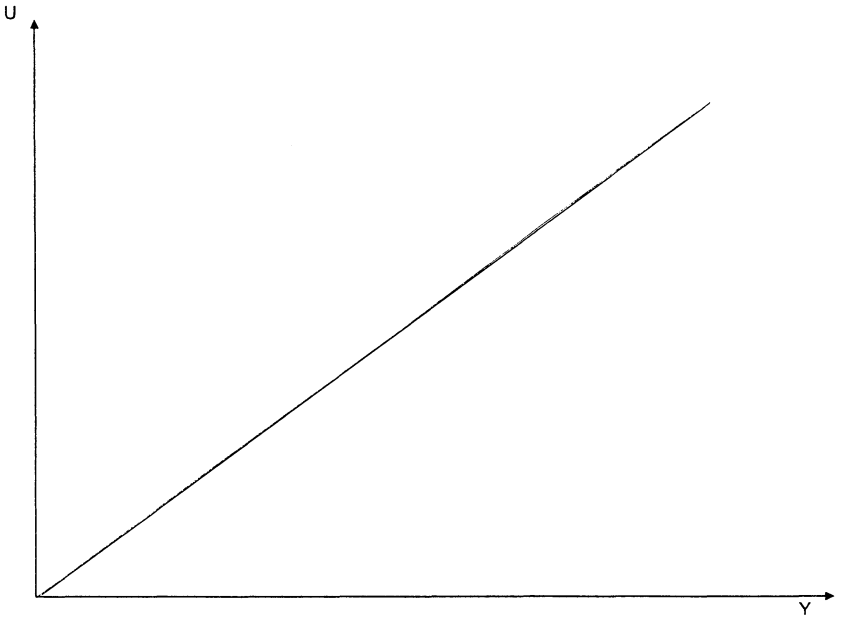

Figure 4a. Risk neutrality with respect to income

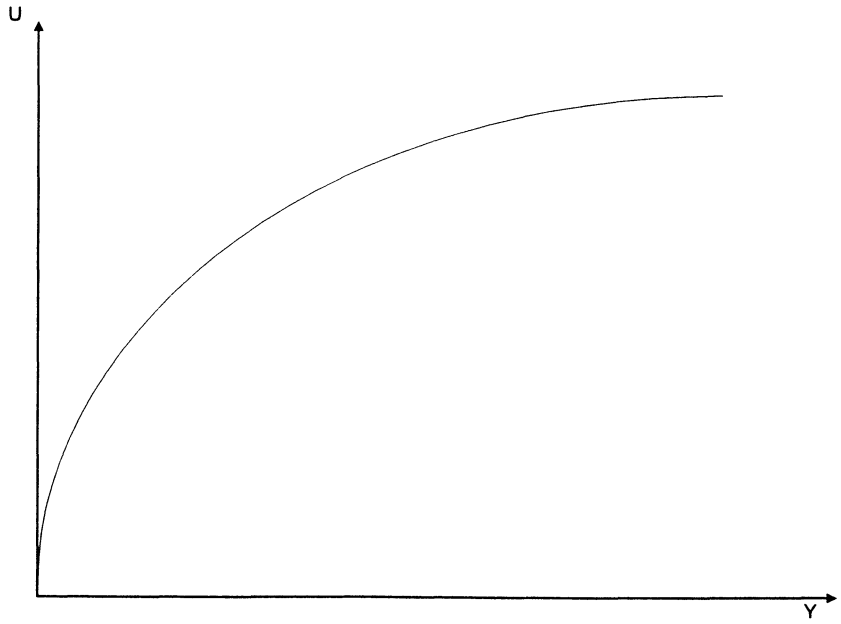

Figure 4b. Risk aversion with respect to income

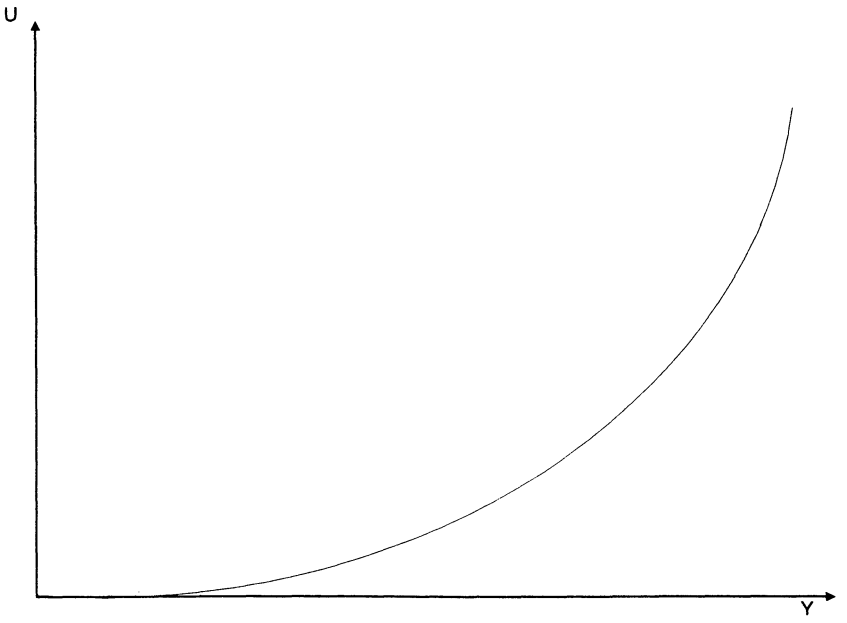

Figure 4c. Risk loving with respect to income

expected utility unchanged. This can be illustrated for two different types of situation involving risks, namely the value of a health care programme when there is uncertainty about what health state the individual will be in and the value of changes in the probability that different health states will occur. Initially we will focus only on morbidity, but after the morbidity cases we will also consider the case when the mortality risk changes.

Assume that there is a 50% chance that the individual will be in the arthritis health state (HA) and a 50% chance that the individual will be in full health (H*), and that a drug is introduced which will cure arthritis. What is the individual willing to pay to make sure that this drug is available? The WTP for the drug is defined by the following equation:

$$0.5*V(Y-WTP,P,H^*)+0.5*V(Y-WTP,P,H^*)=0.5*V(Y,P,HA)+0.5*V(Y,P,H^*) \quad (5)$$

WTP is the amount that if paid to have the drug available keeps the individual on the initial level of expected utility. WTP can here be viewed as an insurance premium that the individual pays to make sure that the drug is available if needed. The other case is when the probabilities of different health states change (note that the above case could also have been defined as a change in the probability of full health from 0.5 to 1.0 and

a corresponding decrease in the probability of arthritis). Assume that a drug decreases the probability of the arthritis state from 0.5 to 0.4 and increases the probability of the healthy state from 0.5 to 0.6. The WTP for this health improvement is then defined by the following equation:

$$0.6*V(Y-WTP,P,H^*)+0.4*V(Y-WTP,P,HA)=0.5*V(Y,P,H^*)+0.5*V(Y,P,HA) \quad (6)$$

WTP is the amount of money that if paid for the risk change keeps the individual on the initial level of expected utility. The WTP for the uncertainty case will have the same properties as the WTP in the certainty case, and will thus always have the same sign as the change in expected utility (see Johansson (1995) for details).

It can be interesting to compare the WTP measure under uncertainty with the WTP measure in the case of certainty. In the uncertainty case it is possible to estimate an expected WTP measure based on the WTP under certainty. With uncertainty about the health state the probability of being in a health state could be multiplied by the WTP for a treatment in that health state (e.g. the probability of the arthritis health state multiplied by the WTP for a cure against arthritis if in the arthritis health state). With a change in the probability of a health state, the change in the probability could be multiplied by the WTP to move from one health state to another with certainty (e.g. the change in the probability of being in full health multiplied by the WTP to move from the arthritis health state to full health with certainty).

It is of interest to investigate the relationship between such expected WTP measures and the measure defined by equations (5) and (6), since it may be easier to collect data empirically about the WTP under conditions of certainty rather than under conditions of uncertainty (especially for small probabilities of health states and small changes in the probabilities). Sometimes we may also have information about the WTP for a treatment among individuals in a specific health state and we would like to estimate what the WTP for the treatment would be in an insurance situation.

Below when we discuss the properties of the expected measures we focus on the WTP measure and not the WTS measure, i.e. the monetary measure under certainty that is used to estimate the expected measure is always the WTP measure. To make a distinction between expected WTP and the measure defined by equations (5) and (6), we refer to the ex ante measure as ex ante WTP. The ex ante WTP is the measure we are interested in as the value of a health change in the case of uncertainty, and we investigate the relationship between this measure and the expected WTP.

In general the relationship between ex ante WTP and expected WTP depends on how the marginal utility of income varies with income when the health status is held constant, and how the marginal utility of income varies with health status when the income is held constant. If individuals are risk neutral with respect to income (i.e. the marginal utility of income does not vary with income) and the marginal utility of income is the same in all health states, ex ante WTP and expected WTP will always

coincide. It is generally assumed in economics, however, that individuals are risk averse with respect to income, i.e. that the marginal utility of income decreases with higher income.

If individuals are risk averse with respect to income, but the marginal utility of income does not vary with health status, expected WTP will be a lower bound for ex ante WTP. If we have information about the shape of the utility function with respect to income (i.e. the degree of risk aversion) it would then be possible to estimate the ex ante WTP based on the expected WTP. This case is illustrated in Figure 5, where the marginal utility of income is the same in all health states (i.e. the slope of the utility function is the same in all health states). In that case only the income level and not the health state matters for the marginal utility of income

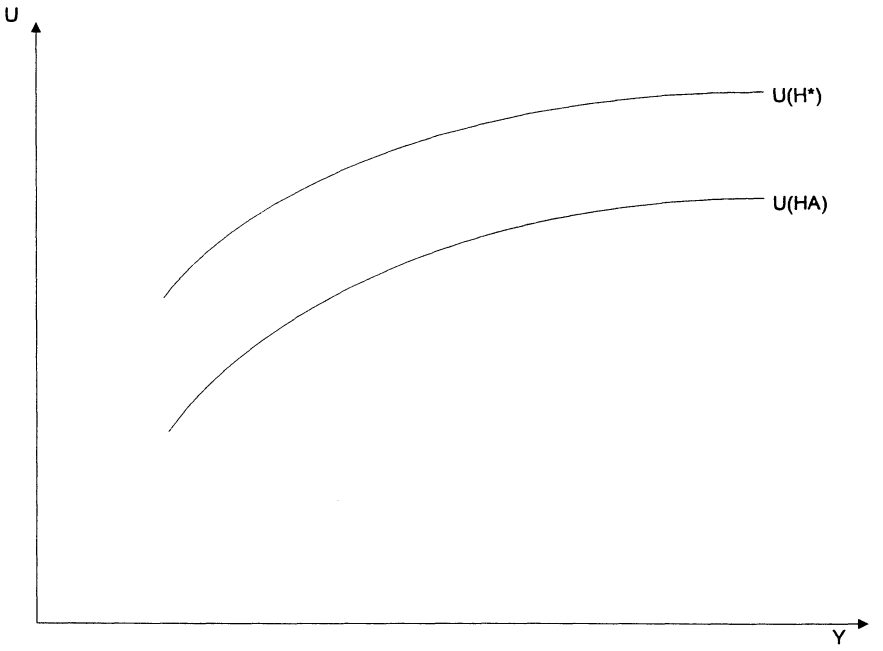

Figure 5. Constant marginal utility of income with health status

Information is limited about how the marginal utility of income varies with health status, but according to a study by Viscusi & Evans (1990) the marginal utility of income increased with better health status. A case with increasing marginal utility of income with health status is shown in Figure 6. In this case the slope of the utility function with respect to income increases with better health status.

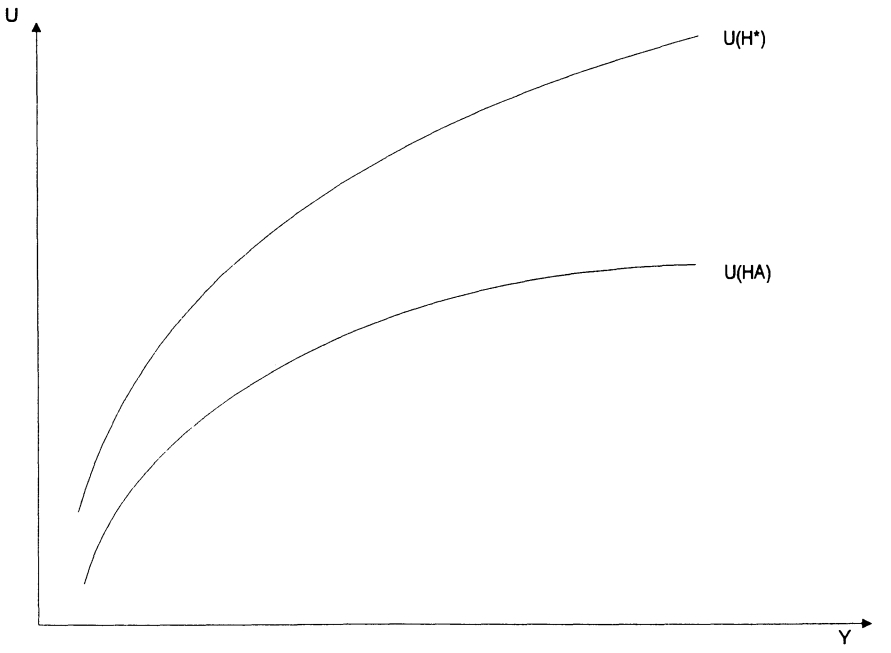

Figure 6. Increasing marginal utility of income with health status

If the marginal utility of income increases with better health status it is possible for the expected WTP to exceed ex ante WTP for a treatment that does not lead to full health even if an individual is risk averse with respect to income. This could be the case, since the marginal utility of income that is used to convert expected WTP to the change in expected utility is assessed in the health state with treatment having a lower marginal utility of income than in full health. It is thus possible that individuals will choose not to fully insure against some health care costs even though the insurance is actuarially fair and the individual is risk averse with respect to income.

Even if the marginal utility of income increases with increasing health, expected WTP is a lower bound for ex ante WTP as long as individuals are risk averse with respect to income, for treatments that restore individuals to full health or for treatments that increase the probability of being in full health (as long as these treatments do not also increase the probability of some health state with less than full health).

To see this, consider first the case where there are only two possible health states: full health and less than full health (here assumed to be arthritis), and a treatment restores the individual to full health if in the less than full health state. Figure 7 illustrates this case using the above example about arthritis. The figure shows the utility function with respect to income in the health states full health (H*) and arthritis (HA). The individual has a 50% chance of being in each health state and the income in both

health states is Y0. A drug is introduced that will cure the individual of arthritis. If the individual is in the arthritis state the WTP would be the maximum income the individual is willing to give up to be in full health, which is equal to Y0-Y" in the figure. The expected WTP is equal to this amount multiplied by the probability of the arthritis state (i.e. half of the distance between Y0 and Y"). Ex ante WTP is the amount that if paid leads to the same expected utility as without the drug. In the figure ex ante WTP is equal to the amount Y0-Y'.

This case is identical to the case where an individual insures against a 50% chance of an income loss of Y0-Y". The reason for this is that both expected WTP and ex ante WTP can be converted to the change in expected utility by multiplying them by the marginal utility of income in the same health state, since the treatment restores the individual to full health. The reason that expected WTP and ex ante WTP do not coincide in this example is that the individual is risk averse, and the marginal utility of income used to convert the monetary measures to the change in expected utility differs for expected WTP and ex ante WTP.

In the figure, expected WTP is converted to the change in expected utility by multiplying it by the mean marginal utility of income for the change in income from Y" to Y0. Ex ante WTP on the other hand is converted to the change in expected utility by multiplying it by the mean marginal utility of income for the change in income from Y' to Y0. Since the marginal utility of income decreases with higher income for a risk averse individual, ex ante WTP will always exceed expected WTP for a treatment that restores an individual to full health if the individual is risk averse.

This means that expected WTP is a lower bound for ex ante WTP in that case. It also means that if we have information about the utility function with respect to income for healthy individuals (i.e. the degree of risk aversion) we could estimate ex ante WTP based on expected WTP in that case.

In a more realistic case with more than two possible health states, expected WTP will still be a lower bound for ex ante WTP for a treatment that restores the individual to full health or increases the probability of the full health state even if the marginal utility of income increases with better health. In such a case the ex ante WTP could be converted to the change in expected utility by multiplying it by a weighted average of the marginal utility of income in all the different possible health states after the intervention. In such a case the relationship between the marginal utility of income and health status would affect ex ante WTP, but if the marginal utility of income increases with better health this would increase ex ante WTP and increase the difference between ex ante WTP and expected WTP further (since the weighted average marginal utility of income will be lower than the marginal utility of income in full health).

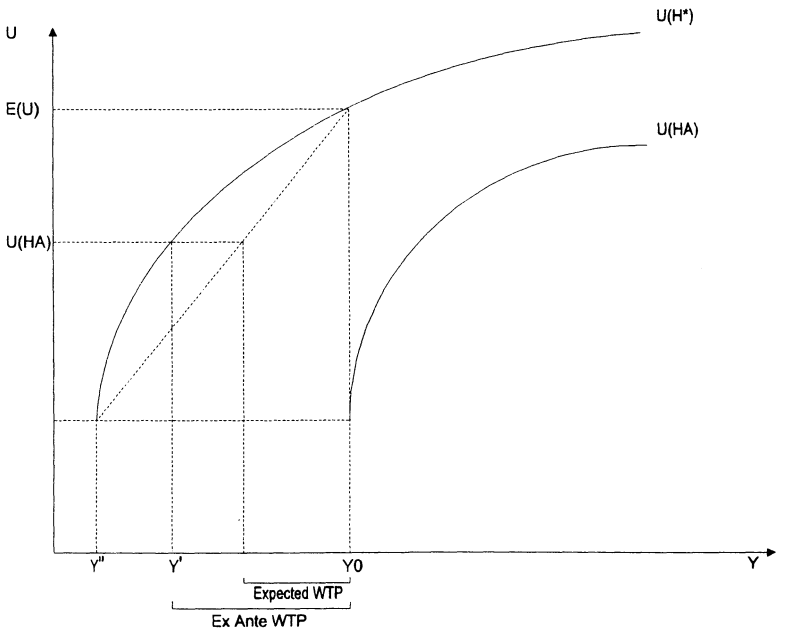

Figure 7. Ex ante WTP VS expected WTP

However, for treatments that do not restore the individual to full health and for treatments that increase the probability of health states with less than full health, expected WTP is not a lower bound for ex ante WTP even if individuals are risk neutral with respect to income and the marginal utility of income increases with better health. This is because the expected WTP should now be converted to the change in expected utility using a marginal utility of income in a health state with less than full health.

For treatments that increase the probability of a better health state, but also increase the probability of a worse health state due to for instance side-effects, it is not necessarily the case that expected WTP even has the same sign as the change in expected utility. Consider for instance a treatment that increases the probability of being healthy rather than having arthritis by 10 percentage units, but also increases the risk of serious side-effects leading to a health state that is worse than arthritis by 5 percentage units. Assume further that the WTP for avoiding the side-effects and being healthy is $10,000 and the WTP for avoiding the arthritis state and being healthy is $5,000. The expected WTP would be zero (0.10*5000+0.05*-10,000), but this does not mean that the change in expected utility is zero. The marginal utility of income that should be used to convert the $10,000 to a change in utility would be higher than the marginal utility of income that should be used to convert the $5,000 to a change in

utility for a risk averse individual. The treatment would thus reduce the expected utility for a risk averse individual.

To summarise the comparison of ex ante WTP and expected WTP, it is important to note that both the variation in the marginal utility of income with income and the variation in the marginal utility of income with health status is important for the relationship between these measures. Even if we assume risk aversion, expected WTP is not necessarily a lower bound for ex ante WTP, since the marginal utility of income may vary with health. If the marginal utility of income is constant with health expected WTP is a lower bound for ex ante WTP, and then it should also be possible to use information about risk aversion to estimate the ex ante WTP based on the expected WTP. For treatments that lead to full health (or close to full health) and for increased probabilities of staying healthy rather than entering a disease state it also seems reasonable to treat the expected WTP as a lower bound to the ex ante WTP and to estimate the ex ante WTP on the basis of information about risk aversion with respect to income. This is true at least so long as the disease or disability probabilities are low, so that ex ante WTP is only marginally affected by variations in the marginal utility of income with health.

It should also be noted that the above discussion has been based on WTP rather than WTS as a basis for the expected measure. For WTS the results would in general be the opposite compared to WTP. If the marginal utility of income does not vary with health status and the individual is risk averse with respect to income the expected WTS would be an upper bound for the ex ante WTS, since the marginal utility of income to convert the expected WTS to the change in expected utility would be lower than the marginal utility of income used to convert the ex ante WTS to the change in expected utility.

Another important issue before we turn to the mortality case is how the ex ante WTP for risk changes is affected by the baseline risk and the size of the risk reduction. Note first that according to the expected utility theory, each additional unit of increased probability of being healthy rather than sick leads to the same increase in expected utility (i.e. the expected utility is simply the sum of the probabilities multiplied by the utilities of each state), given that the increased probability does not change the income in each state (as it will if the individual pays for the risk reduction; see below).

The importance of the baseline risk for WTP depends on the difference in the marginal utility of income between the sick and the healthy state. Consider again our two health states arthritis and full health. For a given reduction in the risk of the arthritis state of for instance 1 percentage unit, we would expect the WTP to be higher for a reduction from 90% to 89% than for a reduction from 10% to 9% for the same income level (and everything else also held constant) if the marginal utility of income is higher in full health than in the arthritis state. This is because the marginal utility of income that is used to translate the WTP into a change in expected utility is a weighted average of the marginal utility of income in the healthy and the sick state.

Since the weights will be the probabilities of the events, the marginal utility of income will be lower as the initial probability of the arthritis state becomes higher, if the marginal utility of income is lower in that state.

We would also expect the marginal WTP to decrease with the size of the risk reduction. If an individual is risk averse with respect to income, the marginal utility of income increases the more the individual pays (i.e. the opportunity cost in terms of the private goods that have to be given up increases) and the marginal WTP will thus decrease for a larger risk reduction. Since the probability of being in the healthy state also increases with the risk reduction, this will also increase the marginal utility of income if the marginal utility of income is higher in the healthy state than in the sick state. Thirdly, when the individual pays the WTP the income in all states will decrease, and if the marginal utility of income is higher when healthy than when sick the difference in utility between the healthy and sick states will decrease, which decreases the gain in utility of additional risk reductions. These factors will all tend to decrease the marginal WTP for larger risk reductions. This means that we will not expect an individual to be willing to pay twice as much for a reduction in the risk of the arthritis state from 10% to 8% as for a reduction from 10% to 9%.

For a marginal increase in the risk we would also expect the WTS to be higher for a higher baseline risk, if the marginal utility of income is higher when healthy than when sick. For the size of the risk reduction we would, however, expect the reverse pattern compared to WTP, i.e. that the increase in WTS for a marginal risk increase will be greater for a larger risk increase. This is because now the marginal utility of income will decrease with increasing compensation for a risk averse individual. The marginal utility of income also decreases for higher risks if the marginal utility of income is higher when healthy than when sick, which would also increase the WTS. Finally, the difference in utility between healthy and sick states increases with greater compensation if the marginal utility of income is higher when healthy than when sick.

There are thus two factors that will tend to diverge WTP and WTA, namely risk aversion with respect to income and the fact that the marginal utility of income may be lower when sick than when healthy. However, for marginal changes we would expect WTP and WTS to coincide as shown in Johansson (1995) (i.e. for a marginal risk change the same marginal utility of income is used to convert WTP and WTA into the change in expected utility). In practice, however, it is not clear how small a risk change has to be in order to be "marginal".

We can then continue and consider mortality risks. Assume that the individual initially has a 90% chance of being alive (H*) and a 10% chance of being dead (H#) and that a treatment is introduced which increases the probability of being alive to 80%. The WTP of this treatment is then defined by the following equation:

$$0.9*V(Y-WTP,P,H^*)+0.1*V(Y-WTP,P,H\#)=0.8*V(Y,P,H^*)+0.2*V(Y,P,H\#) \quad (7)$$

WTP is the amount of money that if paid for the risk reduction keeps the individual on the initial level of expected utility. If we investigate how WTP and WTS vary with the baseline risk and with the size of the risk reduction the results are similar to the morbidity case, but the predictions are somewhat stronger. If we ignore the issue of altruism (i.e. in this case that individuals may derive utility from leaving money behind, for example to their children), the utility of the dead state will be zero and the marginal utility of income in the dead state will be zero. The marginal utility of income to convert WTP or WTS to a change in expected utility will then be equal to the marginal utility of income when alive multiplied by the probability of being alive (i.e. the weighted average of the marginal utility of income when alive and when dead).

This means that the weighted marginal utility of income would be higher as the baseline risk of dying decreases, and we would thus expect WTP or WTS for an identical risk change to be higher as the baseline risk of dying increases.

We would also expect the WTP for a marginal risk reduction to decrease with the size of the risk reduction, and the WTS to accept a marginal risk increase to increase with the size of the risk reduction. This result is due to both the assumption of decreasing marginal utility of income with income and the higher marginal utility of income when alive than when dead. The higher marginal utility of income when alive than when dead has two effects, as in the morbidity case. It leads to a decrease in the weighted marginal utility of income when the mortality risk is higher. It also leads to the difference in utility between alive and dead states increasing with income.

It should also be noted that if we try to define the WTP to avoid a certain death or the WTS to accept a certain death, WTS would be infinitely high and WTP would be bounded by income (i.e. for the WTP case the WTP is the amount of money that reduces utility to zero both with and without the programme, and for the WTS case it is impossible to equalize the expected utility with and without the programme since the marginal utility of income when dead is assumed to be zero). Hoever, it does not make sense to define monetary measures for the complete loss or gain of life, but what is of interest is the trade-offs between changes in risks and income (see Johansson (1995) for a further discussion on this).

In the mortality case it can also be of interest to define a term called the value of a statistical life. The value of a statistical life is the sum of the WTP for a risk reduction among a group of affected individuals divided by the number of lives saved in that group. If for instance a programme will reduce the mortality risk from 5/100,000 to 4/100,000 among a group of 500,000 individuals and the mean WTP per individual in this group is $10, the value per statistical life is (10*500,000)/5=$1,000,000. The results of empirical studies are often presented in terms of the value per statistical life. On the basis of the above model, it can be expected that the value per statistical life will vary depending on the base-line risk, the size of the risk reduction and the income level of the population.

However, for the same population the value of a statistical life on the margin (i.e. for the same risk reduction) for different types of mortality risk reduction (e.g. traffic versus heart disease) would be expected to be the same on the basis of this model, because no distinction is made between different ways of dying. In principle, however, it would be possible to attach different utilities (or disutilities) for different ways of dying.

When the value of statistical life is compared between studies it is important to keep in mind the fact that they often reflect different populations with different base-line risks, different sizes of the risk reduction, and different socioeconomic characteristics (income, education, etc). Thus on the basis of theoretical expectations there is no reason to believe that the value per statistical life should be the same for these studies.

3.2 Altruistic WTP

So far we have only considered the private WTP of improving one's own health. In economics the standard model of individual behaviour is based on the assumption that the individual (or the household) is only concerned with his/her own well-being. In the health field one argument for public intervention and regulation is based on our concern for the health of other individuals. For altruistic reasons we may thus be willing to pay in order to ensure that a health care programme is implemented even though our own health or income is not affected by the programme. The utility function that we used in the last section can then be expanded in the following way:

$$U(C,H,h) \tag{8}$$

where h is the health status of all other individuals in society. We can then define the altruistic WTP in the same way as we defined the private WTP in the last section using the indirect utility function, which will now have income, prices, own health and the health of other people as arguments. Consider for example a programme that improves the quality of life of persons with arthritis. For an individual without arhthritis the WTP of this programme is defined by the following equation:

$$V(Y-WTP,P,H,h1)=V(Y,P,H,h0) \tag{9}$$

In equation (10) WTP is the amount of money that if paid for the improvement in other people's health from h0 to h1 leaves the utility of the individual at the initial level. WTS could also be defined in an analogous manner by considering a decrease in the health status of other individuals.

The way we have entered altruism in the utility function means that we assume that individuals are concerned only about the health status of other individuals and not about the utility or more general wellbeing of other individuals. This type of altruism is sometimes referred to as paternalistic altruism (Jones-Lee 1991,1992). It is

paternalistic in the sense that it does not accept the preferences of the individual, since we are concerned only about the health status of other individuals rather than their general utility level.

Alternatively we could have entered the utility of all other individuals in society in the utility function instead of the health status of the individuals, based on the assumption that individuals are only concerned about the utility of other individuals and do not care about how the individual derives utility. Such a case can be called pure altruism, since it respects the preferences of other individuals.

It has been shown that if the altruism is pure, we can ignore altruism in a cost-benefit analysis (Jones-Lee 1991,1992). The logic behind this result is that both benefits and costs would be increased by the same proportion if altruism was included and that they would cancel out at the optimum, i.e. both costs and benefits have altruistic components. This result is only valid for the case of pure altruism and only holds if we are close to a welfare optimum (Johansson 1994,1995), i.e. if the income distribution is optimal so that the social marginal utility of income is the same for all individuals.

If the altruism were paternalistic or safety oriented the altruistic values would not cancel out. At the moment, there only appears to be one method which potentially can be used in order to estimate the altruistic components of health programmes; this is the contingent valuation method (see below), where individuals in surveys are asked about their willingness to pay for different goods.

It has been argued that within a contingent valuation framework altruism can be incorporated in two alternative ways. According to the first alternative the respondent is asked about the total WTP for a programme after they have been given information about all the consequences of the programme (e.g. the tax increases for different people and the health effects for different people if a programme is financed by taxation) (Johansson 1994,1995). The sum of the WTP for the gainers and the WTS of the losers would then give the information for the cost-benefit analysis. The rationale of framing the question in this way is that the resulting measure would include the altruistic components of both costs (e.g. the reduced consumption for individuals whose taxes increase) and benefits.

Alternatively (Johansson 1994,1995) we could ask individuals about their total WTP for a health care programme based on the presumption that everybody else is paying or being compensated so that they remain on the initial utility level. The rationale for this question is that it would capture only the paternalistic altruism part because of the changes in health status of other individuals. If the altruism is pure, the WTP should in this case be the same as if only the consequences for the individual are included in the WTP, but if the altruism is paternalistic with respect to health the WTP would increase by comparison with the private case (because even if the utility is unchanged for other individuals, they have increased their health status and decreased their consumption of other goods).

3.3 Conclusions

In this chapter it has been shown that it is possible to define monetary measures of changes in health based on utility theory and individual preferences. The WTP of a health improvement is the sum of money that if paid leaves the individual on the same utility level as before the health improvement. The WTS for a health deterioration is the sum of money that if paid to the individual will keep the individual on the same level of utility as before the health deterioration. These measures can also be straigtforwardly extended to the more realistic case with risks of different health states (including death), where individuals are paying for the decreased risk of morbidity and mortality.

The WTP involves the trade-off which individuals are willing to make between health risks and the consumption of non-health goods. Even though the definitions of the WTP measure in the case of risk were based on the expected utility theory, Jones-Lee (1989) has shown that WTP measures for risk reductions can also be defined on the basis of other theories for decision-making under uncertainty, such as prospect theory (Kahneman & Tversky 1979). It has also been shown that it is possible to define an altruistic WTP for health changes, i.e. individuals may be willing to pay for a health care programme even though their own health or consumption is not affected by the programme.

The measurement of the WTP for health changes can be based on the actual decisions which individuals make in situations involving health, such as the use of seatbelts; alternatively, it can be based on the trade-offs between health and income as expressed in surveys. These approaches to measuring the WTP of health changes are reviewed in chapters 5 and 6 below.

REFERENCES

Hicks JR. The foundation of welfare economics. Economic Journal 1939;49;696-712.

Hicks JR. The four consumer's surpluses. Review of Economic Studies 1941;11:31-41.

Johansson P-O. The economic theory and measurement of environmental benefits. Cambridge: Cambridge University Press, 1987.

Johansson P-O. An introduction to modern welfare economics. Cambridge: Cambridge University Press, 1991.

Johansson P-O. Cost-benefit analysis of environmental change. Cambridge: Cambridge University Press, 1993.

Johansson P-O. Evaluating health risks: an economic approach. Cambridge: Cambridge University Press, 1995.

Johansson P-O. Altruism and the value of statistical life: empirical implications. Journal of Health Economics 1994;13:111-118.

Jones-Lee MW. The economics of safety and physical risk. Oxford: Blackwell, 1989.

Jones-Lee MW. Altruism and the value of other people's safety. Journal of Risk and Uncertainty 1991;4:213-219.

Jones-Lee MW. Paternalistic altruism and the value of a statistical life. Economic Journal 1992;102:80-90.

Kahneman D, Tversky A. Prospect theory: an analysis of decisions under risk. Econometrica 1979;47:263-291.

Kaldor N. Welfare propositions of economics and interpersonal comparisons of utility. Economic Journal 1939;49:549-552.

von Neumann J, Morgenstern O. Theory of games and economic behaviour. Princeton NJ: Princeton University Press, 1947.

Pratt JW. Risk aversion in the small and in the large. Econometrica 1964;32:122-136.

Viscusi WK, Evans WN. Utility functions that depend on health status: estimates and economic implications. American Economic Review 1990;80:353-374.

4. THE RESOURCE CONSEQUENCES OF HEALTH CHANGES

In this chapter we analyse the resource consequences due to changes in morbidity and mortality as a result of a health care programme. The resource consequences are defined as the effect on the consumption of goods and services and the effect on the consumption of leisure. We focus both on the extent to which the WTP for a health care programme can be estimated on the basis of information about the production and consumption of individuals, and the extent to which the external costs can be estimated on the basis of this information. In particular, we examine the relationship between the values produced by the human-capital approach (this approach estimates the value of health care programmes in terms of decreased health care consumption and increased production (Weisbrod 1961)) and the private WTP, and the relationship between the human-capital approach and the external costs.

Since the institutional arrangements in society affect whether some resource consequences are externalities or not, we make a distinction between two different types of society: one society where individuals pay for their own health care, income losses due to health detriments, and consumption during years after retirement (either out of pocket or through insurance where the premium reflects the expected payments) and one society where the state pays for health care, income losses due to health detriments, and consumption during years after retirement through a tax system (Johannesson 1994). The distinction between the institutional arrangements is important since it affects what will be included in the private WTP for a health care programme and what needs to be estimated separately.

In the first section below we describe the model and the prerequisites for the analysis that will follow in rather more detail. The relationship between the private WTP, the human-capital approach and the external costs is then analysed for the private system. This is followed by a section where the same issues are analysed for the public system. We end the chapter with some conclusions.

4.1 The Model and Assumptions

In order to analyse the resource consequences of a health care programme and determine whether they are externalities or not, it is necessary to explicitly consider how health care programmes and income are financed. In this chapter we will therefore expand the model used in chapter 3 somewhat. To the utility function above we also add leisure time, which leads to the following utility function:

$$U=U(C,L,H,h) \qquad (1)$$

In equation (1) the utility of the individual depends on the consumption of non-health goods (C), the consumption of leisure (L), the health status (H), and the health status of other individuals (h). Since the issue of altruism has been dealt with above, we

suppress that argument in the utility function in the rest of this section. In the above model income and health status were assumed to be exogenously given. Here income is defined as being equal to the labour income of the individual before any tax payments, and we also assume that individuals can affect their own health status by investing in various health inputs.

In order to consider the externality issue, the health inputs are further divided into health inputs that are provided free of charge (E) through insurance or public provision, and private health inputs that are bought on the market (F). It is also assumed that the individual will use time (X) in the production of health and that the health status is a function of the initial health status of the individual (H0), which leads to the following production function for health:

$$H=f(X,E,F,H0) \qquad (2)$$

The individual will also have two constraints, one budget constraint and one time constraint. According to the time constraint leisure (L), health production time (X), and working time (Z) must be equal to the total time in the period (T). In the general case with private or public insurance the budget constraint will be the following: WZ-PC-PfF-M+N=0. W is the wage rate, P is the price of non-health goods, Pf is the price of private health inputs, M is taxes paid to the state (or in the private system a private insurance premium), and N denotes public income transfers for income losses due to illness and retirement payments (or in the private case, payments from private insurance against income losses due to illness and payments from a retirement insurance).

The above model is designed to incorporate the most relevant issues concerning the resource consequences of health programmes and will be used to show how the external costs should be calculated under different institutional arrangements. It also makes it possible to show how the human-capital method is related to the willingness to pay of individuals and the external costs. When we assess the private WTP of a health care programme, the difference compared to the previous section is that the private WTP will now reflect not only the change in health status but also changes in the consumption of non-health goods and the consumption of leisure.

The WTP of a health care programme is otherwise defined in the same way as in the previous section, but it will now also include all other changes as well as that in health status (i.e. the utility level before a health care programme is introduced is held constant). In all the examples below we focus on the introduction of a health care programme and the WTP, but the WTS for the withdrawal of a health care programme can be analysed in an analogous manner.

We use the above model to analyse how to handle the resource consequences due to changes in morbidity and mortality under two different institutional arrangements in society. The first case will be the purely private market solution with no external costs,

and the second case will be the "social insurance market" where health care costs, income losses due to disease and retirement payments are paid by general taxation. We make a distinction between the resource consequences due to changes in morbidity and the resource consequences due to changes in mortality.

We also analyse the different cases both when there is certainty and when there is uncertainty (i.e. the probabilities of different events). In the uncertainty case it is assumed that the health care programme changes the probabilities of different health states or changes the probability of survival for another period. In these cases the health status and the length of life should be viewed as random variables, whose probability distributions can be changed by investing in health inputs.

For both the morbidity case and the mortality case we first analyse the resource consequences without including the costs of the programme that leads to the change in morbidity and mortality. The private WTP for the change is thus defined as the WTP at a zero price of the health care programme (gross WTP). For simplicity's sake we also assume that the programmes that are analysed at a zero price do not involve any costs for the individuals (e.g. no time costs), since these costs would otherwise enter the WTP at a zero price. Towards the end of each section we analyse how the programme costs as such should be defined.

Throughout, we will work with the simplest possible case using simple examples to illustrate the results. To simplify things no discounting is assumed and all insurances are assumed to be actuarially fair, and all taxes are assumed to be lump-sum so that no excess burden of taxation arises (for the method of dealing with taxes see Chapter 8 below). In addition it is also assumed that market prices reflect prices under perfect competition (i.e. various sources of market failure and market imperfection are not considered here; see instead the section about market failure in Chapter 2 for a discussion of these issues, and Chapter 7 for the estimation of costs below).

4.2 A Private System

In a private system it is assumed that health care costs, income losses due to health deterioration and retirement payments are not funded publicly, and if insurance exists then this is private insurance where the premium is set exactly at the expected cost for each individual.

4.2.1 Morbidity Changes

We can start by analysing the private case for morbidity and no uncertainty. In this case the budget constraint can be written as follows: WZ-PC-PfF=0. By rearranging the budget constraint we find that the consumption of non-health goods is equal to the labour income minus the consumption of health inputs. For the sake of simplicity we

assume that the time period is one year, and after that year individual A will die with certainty (this means that there are 8760 hours of time in the period). Initially individual A is assumed to suffer from arthritis and to spend $5,000 on health care and $5,000 on other health inputs.

Individual A works 1800 hours in total and the wage rate is $15, leading to a total labour income of $27,000. If individual A was free of arthritis he would work 2000 hours, and the loss of income due to the arthritis is thus $3000. Imagine now that individual A is offered a drug that will cure the arthritis for sure. According to the human-capital approach the value of this drug would be measured as the reduced health inputs and the increased production. In this case the drug leads to an increased production of $3000 (the increased income) and reduced health care costs of $5,000.

It is assumed that even with the drug the individual will spend $5000 on health inputs, since the health inputs include the subsistence consumption (i.e. the minimum consumption needed to stay alive). The human-capital approach leads to an estimate of $8000 for the value of this drug (3,000+5,000). This is also equal to the increase in the consumption of non-health goods by the individual, who would be willing to pay $8000 for this increase. In addition, however, the health status of the individual has also improved and the consumption of leisure has decreased by 200 hours due to the increase in working time. Thus the human-capital estimate only estimates the change in consumption of non-health goods, but ignores the consumption of leisure and the change in health status per se.

An interesting issue is whether the human-capital estimate provides a lower bound on the willingness to pay for the drug. It is not obvious that the change in consumption of non-health goods (the human-capital estimate) provides a lower bound, since the health status improves but the consumption of leisure decreases. Thus for the human-capital estimate to be a lower bound, the utility change due to the health improvement has to be greater than the utility loss due to the decreased leisure.

Since the lost leisure is leisure with arthritis, it may be of low value and it thus seems reasonable to assume that the human-capital estimate is a lower bound at least for significant changes in health status. In principle, in order to obtain the total value of a health care programme in this case one should add the value of the change in consumption of non-health goods, the value of the change in the consumption of leisure and the value of the change in health status.

This is illustrated in Figure 1 using the utility function with respect to non-health consumption. In Figure 1 the individual has a consumption of non-health goods of $17,000 (C0 in the Figure) before the introduction of the drug (the cost of the drug is not included in the figure) and the initial utility function with respect to non-health goods consumption is U (HA,L0). First the change in utility for the decreased leisure is shown as the difference between the initial utility function and the utility function

U(HA,L1) at the same consumption level (C0). This loss in utility is equal to U0-U' in the figure.

Next, the change in utility for the improved health status is shown as the difference in utility between the utility function U(HA,L1) and the utility function U(H*,L1) at the same consumption level (C0). This change in utility is equal to U''-U' in the figure. Finally, the change in utility for the increased consumption of non-health goods is shown as the increased utility for a move from a consumption of $17,000 to $25,000 (C0 to C1 in the Figure) along the utility function U(H*,L1). This gain in utility is equal to U1-U'' in the figure.

This gives the total change in utility in the figure of U1-U0; this could also have been measured directly as the change in utility from the utility function U(HA,L0) at a consumption level of $17,000 (C0) to the utility function U(H*,L1) at a consumption level of $25,000 (C1).

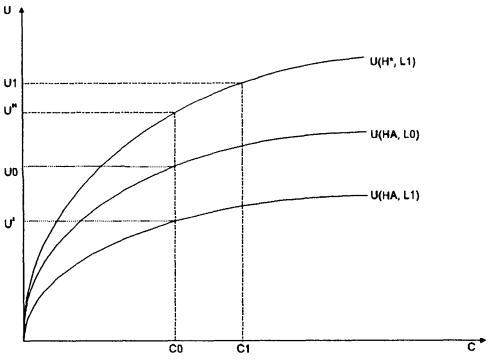

Figure 1. The change in utility due to a treatment

It should be noted, however, that if the willingness to pay for these three components is estimated independently, it may not be possible to simply add them together, since the change in utility of one component may depend on the levels of the other

components, and the marginal utility of income does depend on the level of the other components. However, if the change in non-health consumption is measured independently and the WTP for the combined effect of the changes in leisure and health status is measured independently at the initial consumption level, then these two effects can be added to obtain the total private WTP (this is the same as estimating the WTP for the changes in leisure and health, given that the individual has already paid for the change in consumption).

Note that we did not assume above that the amount of health production time changed due to the introduction of the drug, since if it did it would lead to either a change in non-health consumption (if it affected working time) or a change in leisure time (if it affected leisure time). The changed working time would be included in the human-capital estimate of the value of improved health and the changed leisure time should be included in the overall change in leisure time, but this does not alter the principal conclusions above. Note also that there are no external costs in this case, since there is no change in consumption minus production.

We can then continue our example and add uncertainty to the morbidity case. In the private case it is most likely that individuals would buy private insurance in the uncertainty case (however, as shown in the previous chapter it is not obvious that individuals will insure fully against losses that occur in health states with less than full health, even if they are risk averse). We will therefore assume a case where the individual buys insurance that reduces the price of health care to zero. The individual also buys insurance that fully covers the income losses due to diseases or accidents. For the sake of simplicity it is assumed that both insurances are actuarially fair.

In this purely private case we will also assume that the individual always pays the expected costs as an insurance premium. This means that the budget constraint changes to: WZ-PC-PfF-M+N=0. The labour income plus the reimbursements for income losses from the insurance (N) minus the expenditures on health inputs, the expenditures on non-health goods, and the insurance premiun (M) equals zero.

We assume that the individual has a 50% probability of being in the arthritis health state and a 50% probability of being in the healthy state. The income from labour is $30,000 in the healthy state and $27,000 in the arthritis state, and the sickness insurance benefit is $3000 in the arthritis state leading to the same full income in both states. The individual consumes private health inputs for $5,000 in both states and consumes health care for $5,000 in the arthritis state (for the sake of simplicity we have here assumed that the consumption of health care does not change in spite of the insurance). The insurance premium paid by the individual equals the expected value of $4,000 (0.5*5,000+0.5*3,000), and the insurance premium is the same in both states.

Now assume that a drug is introduced which increases the probability of the healthy state from 0.5 to 0.6. The human-capital method would measure this gain as the

increased production plus the decreased health care costs, which in this case will be $800 (0.1*5,000+0.1*3,000). The insurance premium will also decrease by $800 since the expected loss has decreased by that amount (neglecting the cost of the drug and focusing on gross WTP). The individual will be willing to pay $800 for this decreased premium, since it increases the consumption of non-health goods by $800 irrespective of the health state.

The drug also increases the probability of the healthy state, but also the probability of the state with less leisure. The human-capital estimate will thus be a lower bound if the gain in health status per se exceeds the loss in leisure. The case thus parallells the certainty case. Note also that even though we have introduced insurance for both health care and income losses there are no external costs, since the insurance premium reflects the expected costs.

For persons who are uninsured or not fully insured in the private case, it is not obvious that they will be willing to pay the same amount as the change in the expected consumption of non-health goods. This case is similar to the discussion about expected WTP and ex ante WTP in the section about WTP for health changes, and the relationship between WTP and the expected change in non-health consumption depends on how the marginal utility of income varies with income and health status.

In this private case, the cost of the programme as such also has to be estimated in order to obtain the full resource consequences of a programme. The costs that should be included in this estimation are only those which are not already included in the WTP of the individual at a zero price. For a health care programme this would be the health care costs. Note that if the health programme involves any leisure or working time, these costs will already be included in the gross WTP of the individual and it would be double counting to add them again. The same is true for travel costs due to the programme. In these examples we assumed no time costs or travel costs for the programmes, but they will be included in the changes in leisure and non-health consumption due to the programme.

If the resource consequences are used to estimate a lower bound on WTP or to add to the WTP of the pure health change, then the resource consequences of the programme that are paid by the individual at a zero price, such as leisure time, should be included as well.

One interesting case that could arise is the situation where the health of the individual does not change. Assume for instance that a drug for arthritis does not change the health status, but reduces the in-patient health care costs. In such a case the drug and the in-patient care are alternatives for obtaining the same health gain. In this case it would be possible to estimate the WTP of the individual on the basis of the resource consequences of the programme (i.e. the change in leisure and the change in non-health consumption), and then compare this with the price of the drug. Of course, this only applies if it is possible to estimate the price of the change in leisure or if the

amount of leisure does not change. Such an analysis, where the health is the same for two options (e.g. as in this case the programme VS no programme), is sometimes referred to as a cost-minimisation analysis, where the option with the lowest cost should be chosen.

4.2.2 Mortality Changes

We can continue our example and analyse the case of increased length of life. For the sake of simplicity we now assume that individual A is healthy and thus is working at full capacity, earning an annual income of $30,000. Of this, $5,000 is spent on health inputs (the survival consumption) and the remaining $25,000 is spent on non-health goods. The individual is now offered a drug that increases the survival from one year to two years with certainty. During the second year the production and consumption pattern is assumed to be the same as during the first year.

According to the human-capital approach the value of increased length of life is equal to the increased production, which in this case is $30,000. However, from a theoretical viewpoint it is impossible to know whether this value exceeds or falls short of the WTP of the individual. The individual has to trade off increased length of life against decreased consumption per time unit and it is impossible to know per se how this trade-off will be valued (Rosen 1988).

If the human-capital value was equal to the WTP, the individual would be indifferent between an income of $30,000 and living for one year, and an annual income of $15,000 and living for two years. For this to be the case, the gain in utility of living in year 2 with an income of $15,000 has to equal the drop in utility from the reduced income of $15,000 in year 1. At first glance it may seem reasonable that the human-capital estimate provides a lower bound on WTP if the marginal utility of income decreases at higher income levels.

The problem is that in year 2 some income is needed in order to survive and gain any utility at all. In the example individual A has to spend $5,000 on health inputs in order to reach the survival level. This means that even if the marginal utility of income increases at higher income levels the human-capital estimate will not in theory be a lower bound for the WTP. The importance of the survival level income is easily seen if we assume that there is an endless supply of health programmes which improve survival at exactly the cost of the increased production. If the government implements all these programmes everybody will die, since the consumption per time unit will decrease below the survival level consumption.

In Figure 2 the issue of the survival income is shown by graphing the utility function with respect to income for an individual in full health (H*). The utility only starts to increase when the income reaches a level above the survival level (Y* in the Figure).

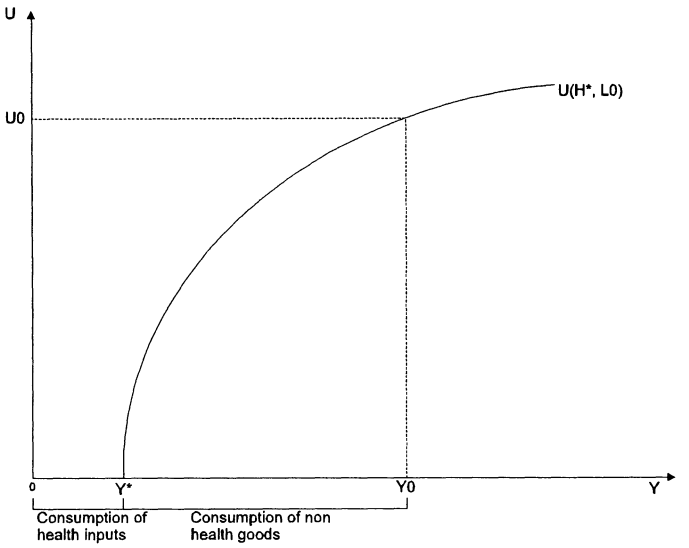

Figure 2. The utility function with respect to income

If an individual can allocate income between the periods so as to even out the marginal utility of income between the periods, the increase in non-health consumption due to the increased survival will be a lower bound on the WTP. Unlike the morbidity case, however, the gain in the consumption of non-health goods will not equal the WTP for this consumption change (unless the utility function is linear in income, i.e. risk neutral with respect to income), since the foregone consumption expressed in the WTP is taken from another consumption level than the consumption gained. The WTP for the gain in consumption of non-health goods (i.e. the WTP for the increased survival) also includes the health and leisure gains, since the utility of the consumption of non-health goods depends on the health status and the amount of leisure. It thus seems meaningless to try and separate the consumption, health and leisure changes in the mortality case. Note also that there are no external costs in this case, since the individual's consumption always equals the production.

In order to introduce uncertainty we can assume that the increased survival is uncertain. Assume that the survival for the first year is certain, but that the probability of survival for the second year is only 0.10. The labour income is $30,000 both years and the individual is in full health and spends $5,000 per year on health inputs. Assume that there is only uncertainty concerning the survival, but not the health state, so that the individual consumes no health care and works at "full capacity" and therefore pays no insurance premium.

Assume that a drug is introduced which increases the survival probability for the second year from 0.10 to 0.15. According to the human-capital approach the value of this would be equal to the value of the increase in expected production. This increase would be \$1500 (0.05*30,000). However, as in the certainty case there is no reason why this should equal the WTP for the increased survival. For small probability changes the WTP and the increased production would be equal if the marginal utility of income in year 1 were equal to the average utility of the income in year 2. Similarly, WTP would exceed the increased production if the marginal utility of income were below the average utility of income, and WTP would fall short of the increased earnings if the marginal utility of income exceeded the average utility of income. It is not the case that a decreasing marginal utility of income will ensure that the human-capital estimate is a lower bound. Again the problem is that the consumption of health inputs during the increased length of life does not directly yield any utility. In the example, the individual has to spend \$5,000 on health inputs in order to survive during the second period.

The increase in the expected consumption of non-health goods will, however, be a lower bound for the WTP provided that the individual can even out the marginal utility of income between the periods by reallocating income. The results are thus the same for the certainty and uncertainty cases. It is not obvious, however, that individuals can reallocate income between different periods so as to even out the marginal utility of income between the periods. If this is not the case, the increase in the expected consumption of non-health goods will not be a lower bound for the WTP even if individuals are risk averse. It is possible for instance that the marginal utility of income at the same income level will differ between periods, due to for instance the health status in the periods, and that the utility of consumption will differ at different ages. In this case again there will be no external costs since the consumption and production of the individual will be the same.

In the uncertainty case it can be of interest to consider one further case with one more type of insurance that is of interest in the mortality case, i.e. a retirement insurance. Assume that the case is the same as before, but with the difference that the individual is now working during the first year and will retire during the second year if he is still alive. The labour income is thus \$30,000 in the first year, and of this income \$5,000 is spent on health inputs. The individual also has a retirement insurance for which he pays a premium that equals his expected cost. The individual pays a premium of \$2,700 initially (0.1*27,000) and the insurance will pay \$27,000 in the second year if the individual survives (so that the non-health consumption is about \$22,000 in both years).

The drug is now introduced and it increases the probability of surviving to the second year from 0.10 to 0.15. The increased production due to a longer life is now zero and the value of the drug according to the human-capital approach is zero. Taking the drug would lead to an increased premium for the retirement insurance of \$1350 (0.05*27,000), which is equal to the increase in the expected total consumption of

$1350 (i.e. the increases in the consumption of health goods and the consumption of health inputs). If the individual agreed to take the drug this would imply that the marginal utility of income in year 1 was lower than the average utility of income in year 2 (i.e. income from the retirement insurance), which is the same condition as for the human-capital estimate to be a lower bound for the WTP.

It is in principle impossible to determine whether the WTP for the drug would exceed zero in this case (although since the individual has a retirement insurance to start with, this indicates that WTP exceeds zero in this example). Even for the private case with a retirement insurance no external costs arise, since the increase in expected consumption during year 2 is exactly offset by reduced consumption during the first year.

It should be mentioned, however, that apart from the retirement insurance no allowances for savings have been made in the above model and examples. If the individual saves money, in order for instance to have a buffer during the retirement years or to pass on to the next generation, and changes in morbidity and mortality lead to changes in private life-time savings, then external costs will arise. The size of the external costs is measured as the change in consumption minus the change in production of the individual due to the programme. For the retirement case above, increased consumption during gained life-years were exactly offset by decreased consumption in other years, but with savings the increased consumption could be financed by decreased private life-time savings leading to an externality.

The estimation of programme costs that will have to be carried out in this case refers only to those costs which are not already included in the WTP of the individual at a zero price. For a health care programme this would be the health care costs. As above, if the health care programme involves any leisure or working time, these costs will already be included in the individual's WTP at a zero price, and it would be double counting to add them again. The same is true for travel costs due to the programme. In these examples we have assumed no time costs or travel costs for the programmes, but if they existed they would lead to changes in leisure and non-health consumption and would thus affect the WTP at a zero price.

If, however, we were using an estimate of the WTP for increased survival that did not include these programme costs, then the programme costs that are paid by the individual at a zero price would have to be subtracted from the WTP or added to the price of the programme. This would also include any adjustments in the use of health inputs or time in the production of health due to the programme. It is for instance possible that part of the "gain" from a health care programme which reduces mortality is the fact that the individual reduces the use of other health inputs and thus increases the consumption of leisure and non-health goods, i.e. a form of off-setting behaviour. These changes should then also be included among the programme costs if the WTP is for the mortality reduction rather than for the whole programme.

4.3 A Public System

We will now continue the analysis for the public system and work through the same cases as above. Here we will especially examine the relationship between the human-capital estimate of improved health and the external costs. The public system case means that we assume that health care costs, income losses due to health detriments, and retirement payments are funded by taxation.

It is furthermore assumed that the tax payments are not related to the expected tax benefits, but that instead the tax is equally distributed among the tax payers in the community (the case thus applies also to a system with private insurance companies, where the insurance premiums are unrelated to the expected costs). For simplicity's sake it is assumed that the total health care costs, income losses due to health detriments and retirement payments are equally divided among the tax payers in the community and that the tax is collected as a lump-sum tax (to avoid an excess burden of taxation; this is instead discussed below in Chapter 8).

Furthermore, for the sake of simplicity it is assumed that the community is sufficiently large so that the increased tax share of an individual which imposes extra costs can effectively be ignored (e.g. if there are 1,000,000 tax payers in the community and an individual tax payer increases his health care costs by $10,000, the increased tax share for the individual would be 10,000/1,000,000=$0.01, but this is neglected below for the sake of simplicity).

The external costs that we will analyse below are defined as the total change in consumption minus production for an individual who receives a health programme. This is the effect on other people's consumption in that community due to the health care programme. If the external costs are positive then other people's consumption decreases, and if the external costs are negative then other people's consumption increases. If the external costs are negative this is thus a benefit of the health care programme. In a publicly funded system the programme costs as such will also be a part of the externality (except if the individual's leisure time or travel costs are part of the programme costs). Thus in order to estimate the total external costs, the programme costs that are not paid by the individual should also be added.

4.3.1 Morbidity Changes

In the public system for morbidity changes we will assume that health care is financed publicly by taxation and that there is a public insurance which covers income losses due to health detriments. This changes the budget constraint to: WZ-PC-PfF-M+N=0, where M is the tax payment (to cover the health care costs and the income losses due to health detriments) and N is the payments received for income losses due to health detriments. The health inputs are now also divided into health care (which is subsidised to a price of zero), and private health inputs.

The new budget constraint means that the consumption of health care and income losses due to health detriments will not affect the consumption of non-health goods. As in the private case, individual A is assumed to suffer from arthritis and to spend $5,000 on health care and $5,000 on other health inputs. The labour income for individual A is $27,000 and individual A also receives $3,000 for the income losses due to the arthritis. We also assume here that individual A pays a tax of $10,000. This means that the consumption of non-health goods initially equals $15,000 (27,000+3,000-5,000-10,000).

Imagine now, as before, that individual A is offered a drug which will cure the arthritis for sure. According to the human-capital approach the value of this drug would be measured as the reduced health care costs and the increased production. In this case the drug leads to increased production of $3000 and reduced health care costs of $5,000. Thus the human-capital approach leads to an estimate of $8000 for the value of this drug. In this case, however, this estimate is not related to the WTP of the individual.

The consumption of non-health goods does not change at all for the individual, since the tax payment does not change and the increased labour income is exactly offset by decreased payments for income losses due to health detriments. In terms of Figure 1, the gain in utility due to increased consumption of non-health goods should not be included in the utility gain for the individual this time, and the utility change due to decreased leisure and improved health status should be evaluated at a non-health goods consumption level of $15,000.

The human-capital estimate is thus not related to the WTP of the individual. In this case, however, the treatment inflicts external costs on the rest of the community and this externality equals the human-capital estimate. The reduced health care costs lead to external costs of -$5,000 and the increased production leads to external costs of -$3,000 (measured as decreased payments for earnings losses due to health detriments). The external costs are thus -$8,000 in total.

The consumption externality is defined as the change in consumption minus production for individual A due to the programme. In this case the consumption was $25,000 initially (the consumption of non-health goods, the consumption of health care, and the consumption of private health inputs), and the production of the individual was $27,000 (the labour income). Due to the programme consumption decreased to $20,000 and production increased to $30,000 and the change in consumption minus production is -$8,000 ([20,000-30,000]-[25,000-27,000]). This decrease in external costs equals the estimate of $8,000 with the human-capital approach.

In this case the change in consumption minus production can be measured directly as the change in health care costs and the change in production. The human-capital

estimate thus has a theoretical basis in this case as a measure of the external costs, but it is only part of the value of a health care programme.

We can then add uncertainty to the morbidity case. Assume that the individual has a 50% probability of being in the arthritis health state and a 50% probability of being in the healthy state. The income from labour is $30,000 in the healthy state and $27,000 in the arthritis state, and the payments for lost earnings due to health detriments are $3,000 in the arthritis state. The individual consumes private health inputs for $5,000 in both states and consumes publicly financed health care for $5,000 in the arthritis state. Since the individual pays $10,000 in taxes in both health states, the consumption of non-health goods is $15,000 in both health states.

Now assume that a drug is introduced which increases the probability of the healthy state from 0.5 to 0.6. The human-capital method would measure this gain as the increased production plus the decreased health care costs, which in this case will be $800 (0.1*5,000+0.1*3,000). For the individual the consumption of non-health goods is the same in both health states and the human-capital estimate does not enter the private WTP. However, the external costs of the drug are equal to the estimate of $800 in the human-capital method. The consumption minus production is -$10,000 in the healthy state and -$2,000 in the arthritis state, and if the probability of the healthy state increases by ten percentage units the increase in the expected production minus consumption is -$800 (0.1*8,000).

The human-capital estimate is thus equal to the external costs, and the uncertainty case is identical to the certainty case. Note also that the fact that uncertainty was introduced does not make the size of the external costs uncertain for the tax payers since the risk is pooled among all of them (Arrow & Lind 1970), and the treatment of the external costs is therefore not affected by risk aversion with respect to income.

In this case with the public system, in order to incorporate the full resource consequences, the programme costs which are not paid by the individual should be added to the external costs as well. If the programme is a health care programme, all the health care costs of the programme should be added so as to get the total external costs of the programme that should be compared to the WTP of the individual. Note that if the programme involves leisure time the change in leisure time is included in the WTP of the individual. For travel costs the same would be true unless they are included among the health care costs and supplied free of charge. If the programme includes working time, this would be included in the external costs if it is assumed that the insurance also covers losses in income due to participation in health programmes. The total external costs of the programme are thus defined as the change in consumption minus the change in production due to the programme.

4.3.2 Mortality Changes

We can continue our example and analyse the case of increased length of life. As in the private case, we now assume that individual A is healthy and is working at full capacity, earning an annual income of $30,000. The individual spends $5,000 on health inputs (the survival consumption) and pays $10,000 in taxes, and the remaining $15,000 is spent on non-health goods.

The individual is offered a drug which increases the survival from one year to two years with certainty. During the second year the production and consumption pattern is the same as during the first year. According to the human-capital approach the value of increased length of life is equal to the increased production, which in this case is $30,000. As in the private case, this cannot be used as an estimate of the WTP of the individual. Furthermore, in the mortality case it is not an estimate of the external costs either.

In order to get an estimate of the external costs, the increased consumption minus the increased production has to be estimated. In this case the external costs of the increased survival are equal to the consumption of $20,000 in the second period minus the production of $30,000. The external costs are thus -$10,000 (i.e. other individuals in the community can increase their consumption by $10,000) and they arise because the individual pays part of the production in tax. Thus in the public system the human-capital estimate for mortality has no theoretical foundation as a basis for either the private WTP or the external costs. In some human-capital studies, however, the value of increased survival is estimated as the production minus consumption during the gained life-years (Ridker 1967). This net-production version of the human-capital approach is an estimate of the external costs. It has to be remembered, however, that this is only a part of the value of increased survival.

We can then introduce uncertainty to the mortality case also. Assume that the survival for the first year is certain, but that the probability of survival for the second year is only 0.10. The labour income is $30,000 in both years and the individual is in full health and spends $5,000 per year in health inputs. The tax is $10,000 each year and the tax covers health care costs, income losses due to health detriments and retirement payments.

Assume that a drug is introduced which increases the survival probability for the second year from 0.10 to 0.15. According to the human-capital approach the value of this would be equal to the value of the increase in expected production. This increase would be $1500 (0.05*30,000). However, as in the certainty case there is no reason why this should equal the private WTP of the increased survival. It does not measure the external costs either. In order to measure the external costs, the consumption minus production during year 2 should be estimated. The net consumption during year 2 is -$10,000 ($20,000-$30,000) and the external costs are equal to -$500

(0.05*-10,000). As in the morbidity case, the introduction of uncertainty does not affect the external costs due to the risk pooling among the tax payers.

As in the section about the private system, it can also be of interest to investigate the retirement case. Assume that the case is the same as before but with the difference that the individual is now working during the first year, but will retire during the second year if he is still alive. The labour income is thus $30,000 the first year, and zero the second year. During the second year the individual receives a retirement payment of $30,000. As before, he pays $10,000 in taxes both years and consumes private health inputs of $5,000 each year. The drug is now introduced and it increases the probability of surviving to the second year from 0.10 to 0.15.

The increased production of longer life is now zero and the value of the drug according to the human-capital approach is zero. As before, this is not an estimate of either the private WTP or the external costs. The consumption minus production is $20,000 during the second year and the external costs are equal to $1,000 (0.05*20,000), i.e. the other individuals in the community would have to decrease their consumption by $1,000 if the individual were treated with the drug.

In order to incorporate the full resource consequences of the programme, the programme costs that are not paid by the individual should also be added to the external costs. If the programme is a health care programme, all the health care costs of the programme should be added so as to get the total external costs of the programme, which should be compared to the WTP of the individual. Note as above that if the programme involves leisure time then the change in leisure time is included in the WTP of the individual, whereas for travel time and working time it depends on whether or not the insurance is assumed to cover these costs. As in the morbidity case, the total external costs of the programme are defined as the change in consumption minus the change in production due to the programme, including also the programme costs not paid by the individual.

4.4 Conclusions

An interesting issue is if the WTP of a morbidity programme can be estimated based on adding the WTP for the changes in leisure, health and non-health consumption respectively. If these amounts are estimated independently then in principle this is not possible, but if they are evaluated sequentially contingent on the fact that the individual has already paid for the prior changes, then they could be added. A practical solution may be to estimate the changes in health and leisure in the form of WTP in a contingent valuation study, based on the initial income level, and then to add the change in non-health consumption so as to get the total WTP. The argument in favour of this approach would be that individuals cannot be expected to estimate the change in their non-health consumption on their own.

However, one problem with this approach is that the change in leisure which is due to a change in the number of hours worked is tied to the change in labour income. It may thus be unrealistic to ask individuals to incorporate the decreased leisure due to the increased number of hours worked, and at the same time ask them to ignore the increased income. In the uncertainty case it also means creating a new possible health state where the earnings and health care costs are unchanged. The most natural way may be to ask the respondents to assume that a social insurance system exists which pays for health care costs and earnings losses due to disease or accidents. However, this would change the initial position of the individual if he/she is not actually fully insured against income losses and health care costs, since the non-health consumption would increase in all health states where health care costs and/or income losses occur. It still seems to be a reasonable approximation, however. Note that within a public system the change in non-health consumption consisting of changes in health care costs and insured earnings losses would not be borne by the individual, so that in this case these changes should anyhow be estimated separately as an externality.

The external costs were defined as the change in consumption minus the change in production for an individual receiving a health care programme. This would thus also include changes in the private lifetime savings of individuals. Some people may argue that changes in private lifetime savings are not an externality, and that they are internalised in the WTP for a health programme. If for instance a household perspective is used instead to define the external costs (i.e. the change in the consumption minus the change in the production of the household due to the health programme), some of the changes in private lifetime savings may be internalised. An alternative way to estimate the external costs could be to estimate them as the net contribution to the tax or insurance system (if insurance does not reflect expected costs) provided that, as assumed here, prices reflect opportunity costs. Such a definition would be similar to the estimates of the external costs of smoking and alcohol consumption that have been carried out in the literature (Manning et al 1991). This is discussed further in the chapter about the estimation of costs.

Another interesting issue is that in the private system it was assumed that the insurance premium perfectly reflects the expected costs of an individual. This seems to be unrealistic, and even in systems based on private insurance externalities are likely to be imposed. The private case with insurances may thus be similar to the public system in reality, and the insured costs should then be treated as externalities. In all societies with taxes, externalities will also be imposed even if health care, income losses due to disease or accidents, and retirement payments are funded by private insurances where the premiums perfectly reflect the expected costs.

REFERENCES

Arrow KJ, Lind RC. Uncertainty and the evaluation of public investment decisions. American Economic Review 1970;60:364-378.

Johannesson M. The cost concept in economic evaluation of health care: a theoretical inquiry. International Journal of Technology Assessment in Health Care 1994;10:675-682.

Manning WG, Keeler EB, Newhouse JP, Sloss EM, Wasserman J. The costs of poor health habits. Cambridge MA.: Harvard University Press, 1991.

Ridker RG. The economic costs of air pollution. New York: Praeger, 1967.

Rosen S. The value of changes in life-expectancy. Journal of Risk and Uncertainty 1988;1:285-304.

Weisbrod B. Economics of public health: measuring the impact of diseases. Philadelphia: University of Pennsylvania Press, 1961.

5. THE REVEALED PREFERENCE APPROACH

If the costs of implementing health care programmes are to be compared directly with health benefits, it is necessary to express health consequences in monetary units. This chapter and the following chapter examine the empirical methods that economists have devised to quantify how much citizens are willing to pay in monetary terms for health effects. There are two principal approaches that can be used to obtain willingness-to-pay estimates of health changes: revealed preference as observed in actual choices or expressed preference as observed in hypothetical choices in surveys. This chapter is devoted to the revealed preference approach and the following chapter is devoted to the expressed preference approach.

In their daily lives, individuals take actions that influence their probabilities of being injured, becoming sick, or dying prematurely. Some homes are located in cleaner and safer neighbourhoods than others. Working in some jobs is more hazardous than others. Driving a big car is safer than driving a small car (at least from the perspective of the relative probability of death or serious injury in case of a crash). A diet high in saturated fat content increases the risk of a heart attack more than does a diet with less saturated fat. Although some risk factors for disease and injury are beyond the individual's immediate control (e.g. family history and the dangers of war), individuals can influence their health risk profile through personal choices about where to live, what job to take, what kind of car to buy, and what types of food to eat.

The health of an individual can thus be considered to be partly "endogenous" to the individual, which means that people can improve their health status through use of various resources such as time and technology. By studying various decisions that involve tradeoffs of health and non-health outcomes, inferences can be made about the willingness to pay for health. Examples of choices that have been studied include whether to use a pedestrian subway instead of crossing a busy street (Melinek 1974), the decision of how fast to drive on a highway (Ghosh et al 1975), the decision to wear automobile safety belts (Blomquist 1979), the decision to purchase a residential smoke detector (Dardis 1980), and the decision whether to take measures to reduce concentrations of radon inside the home (Åkerman et al 1991). But the classic money-risk tradeoff documented in the economics literature is the wage compensation provided to workers who assume hazardous jobs. This chapter starts with a description of wage-risk studies. A section about consumer choices follows and then the chapter ends with some conclusions.

5.1 Wage-risk studies

A high-rise construction worker faces more immediate danger on the job than does a university professor. Since health risks vary by occupation, a testable hypothesis, first suggested by Adam Smith (1776), is that (all else equal) the observed wage should be

higher in riskier jobs. This prediction flows directly from classical economic theory of labour markets.

If an employee would prefer to be healthy rather than injured, then compensation will be required to induce him or her to take a hazardous job rather than a non-hazardous one. From the perspective of the employer, it may be costly or unfeasible to eliminate job-related hazards and thus the employer is prepared to offer a wage premium to employees who assume a hazardous job. Firms will weigh the marginal costs of making jobs safer against the marginal costs of wage compensation for hazardous jobs. The wage premium for hazard that the employer pays should not exceed the marginal cost to the employer of reducing the hazard.

The required compensation for the same job risk may differ between individuals because of heterogeneity in the value of safety (i.e. some workers care more about health and safety than others). Likewise, the marginal costs of providing more safety in the workplace are assumed to differ between firms (e.g. due to differences in the age of a production facility or the average level of experience among employees). Hence, we expect to observe different combinations of wages and risks in real-world labour markets.

The wage-risk combinations that are observed for each job in the market are assumed to represent points where the marginal value of safety for the worker equals the marginal cost to the employer of providing more safety. The aim of a wage-risk trade-off study is to fit a curve through all of these points. The estimated slope of this curve is assumed to reflect both the worker's willingness to pay for a marginal increase in job safety and the worker's willingness to sell (in the form of wage compensation) for a marginal decrease in job safety (Viscusi 1993). For the firm, the slope simultaneously reflects the marginal cost of greater safety and the marginal cost reduction of diminished safety (Viscusi 1993).

The task of estimating the wage-risk tradeoff function is not trivial. Casual observation might lead one to the wrong conclusion, since factory workers earn less than factory managers and poor workers tend to be concentrated in the most hazardous occupations. Since such bivariate comparisons can be misleading, multiple regression analysis is employed to estimate the "hedonic wage function". This means that the researcher tries to explain the variation in observed wages using the job risk measure and other variables that can be assumed to determine the observed wages. Inclusion of the non-risk explanatory variables (e.g. the skill level required in the job) are crucial, since in order to estimate the wage-risk trade-off it is necessary to control all the other factors that affect the wage. A regression equation of the following form is typically estimated:

$$W=\alpha+\beta 0R+\beta iZi+\text{random error} \qquad (1)$$

In equation (1) W is the wage rate, α is a constant, R is the job risk, and Zi are i different variables reflecting characteristics of the worker and the job. By entering the values of the different variables for individual workers, it is possible to estimate the coefficients of the regression equation (α, $\beta 0$ and βi). The coefficient $\beta 0$ then shows the wage compensation for a marginal increase in risk, holding the other factors constant. Note that the linear specification used above implies that the wage-risk trade-off is the same at all levels of risk, while the logarithm of the wage rate, which is often used, implies that the trade-off will vary with the risk level.

In principle, hedonic wage models should include all variables that might explain differences in wage rates among workers. Typically, the explanatory variables include both demographic factors (e.g. age, sex, education and geographical location) and job characteristics (e.g. the speed of work, degree of job security and worker skill requirements). Since insurance coverage for job-related injury might be expected to influence the required wage premiums for job risk, some recent studies have included as an explanatory variable the expected compensation to the worker who is injured or killed on the job (Viscusi & Moore 1987). The choice of explanatory variables is important: collinearity between different explanatory variables can create imprecise coefficient estimates (Atkinson & Crocker 1987) while failure to include relevant explanatory variables (Hwang et al 1992) may cause serious bias in the estimated regression coefficients.

The standard approach can be illustrated with the results of one of the first wage-risk studies (Thaler & Rosen 1976). In a linear hedonic wage equation, the estimated coefficient for the death-risk variable was 0.0352. In this study, the risk variable was measured as the annual risk of death per 100,000 persons while the wage variable was measured as weekly earnings. The risk coefficient implies that a job with an increased risk of 1/100,000 pays \$0.0352 more per week, or \$1.76 per year. If \$1.76 is multiplied by 100,000, we obtain the implied value per "statistical life" of \$176,000 (in 1967 dollars). The results of wage-risk studies are often reported in terms of the value per statistical life (for mortality risks) or the value per statistical injury (for non-fatal injury risks).

Viscusi (1992,1993) summarized the results of 24 wage-risk studies and found that the estimated value of a statistical life varied between \$0.6 million and \$16.2 million (in 1990 dollars). He concluded, on the basis of a technical evaluation of the various studies, that the most reasonable estimates of the value per statistical life are in the \$3 million to \$7 million range. In his summary of the 14 studies addressing non-fatal job injuries, the value per statistical injury varies between \$13,000 and \$130,000 (1990 dollars). It is important to note, however, that the non-fatal injuries in various studies differ with respect to severity.

Since all of us will die sooner or later, it may be more useful to estimate the value of additional life-expectancy rather than the value of a statistical life. Wage-risk studies have explored this issue by examining how wage premiums vary as a function of the

number of life years at risk. In particular, the hypothesis is that the wage-risk trade-off will vary with worker age, with a declining compensation demanded at greater ages. One strategy for addressing this issue is to add another explanatory variable to the wage regression, equal to age multiplied by the job-risk measure (Thaler & Rosen 1976; Viscusi, 1979). Alternatively, the traditional job-risk variable may be replaced by a new variable equal to the number of expected life years lost due to job risk (i.e. life expectancy multiplied by job risk) or the number of discounted life years lost due to job risk (Moore & Viscusi 1988; Viscusi & Moore 1989). From these more refined studies, the willingness to pay for life years saved has been estimated (and the discount rate applied by workers to future life years). Use of a variable reflecting the number of expected discounted life-years lost in order to value life years is complicated by the possibility that intrinsic safety preferences may be dependent on age. If one assumes that safety preferences vary with age independently of differences in life expectancy, then the proper interpretation of these estimates is not obvious, since the estimates assume a constant value per statistical discounted life-year with age.

It may be more straightforward to estimate the value per statistical life-year by dividing the value per statistical life in the wage-risk studies by the average discounted life-expectancy at the age of the sample. By estimating the value of a statistical life as a function of age, the value per statistical life-year can also be estimated for different ages, and may be allowed to vary with age. If the value per statistical life is in the $3 million to $7 million range, as is concluded to be the most reasonable value in the review by Viscusi (1992,1993), this implies a value of a statistical discounted life year of about $150,000 to $350,000 if the average remaining life-expectancy of a worker in the wage-risk studies is 40 years and life-years are discounted by 5%, which is common in cost-effectiveness analyses (if life-years are not discounted the value per statistical life-year is about $75,000 to $175,000).

The main challenge in wage-risk studies is to obtain valid data. Information on individual workers is required, including data on the wage rate, the job risk, and other variables that affect the wage rate. The appropriate dependent variable is the after-tax wage rate, since that is the money that the individual worker receives. Since the after-tax hourly wage rate may not be available, many studies use other proxy income variables.

The key explanatory variable in hedonic wage equations is the one representing job risk. Sometimes job risk is represented as two variables: one for the risk of fatal accidents and one for the risk of non-fatal job accidents. The more complex description of job risk makes it possible to produce separate estimates of the amount of compensation required for marginal increases in fatal and non-fatal risks. If the non-fatal injury variable is not included, then the coefficient estimate for the fatal risk variable may be biased upward (assuming that fatal and non-fatal injury risk are positively correlated). The inclusion of both a fatal and a non-fatal risk variable may, however, cause problems of multicollinearity leading to imprecision in the estimated coefficients for the risk variables.

One of the major problems with measures of job risk is that they tend to be based on actuarial (objective) risk data rather than the dangers perceived by workers. Ideally, the risk measures should be the subjective probabilities of fatal and non-fatal injury as perceived by the workers. In most studies, the (heroic) assumption that workers have perfect information about the actual probablities is made. Viscusi (1993) argues that individuals tend to overestimate low-probability events and underestimate high-probability events. If true, this means that perceived differences in job risks will be understated by workers, leading to an underestimate of the wage compensation required for a specified change in risk.

In several studies the objective measure of risk has been combined with qualitative (self-reported) information on whether or not the worker perceives his or her job to be hazardous (Viscusi 1979). Those who do not perceive their job as hazardous are assigned a value of zero risk (regardless of the objective estimate), while those who perceive hazard on the job are assigned the objective estimate of risk. Although this adjustment for risk perception is very crude, it is a step in the right direction of basing the studies on the subjective probabilities perceived by workers.

Since institutional arrangements in real-world labour markets can depart from perfectly competitive markets, the observed wage premiums for job hazards may not reflect what would be observed in a perfectly competitive market. For example, unionized workers tend to receive larger wage-risk premiums than non-union workers. Workers who do not have significant job choices may also not be able to express their safety preferences. For example, once a worker's pension is vested (i.e., not transferable), the worker may find it too costly to move from one firm to another. Legal arrangements such as minimum wage laws may also reduce the number of job opportunities available to unskilled labourers, thereby distorting equilibrium wage rates. In the presence of labour market imperfections, the observed wage-risk premiums may not be identical to what would be observed in a perfectly competitive labour market.

Another problem in the interpretation of the results of wage-risk studies is that the observed wage premium may represent the preferences of the last worker hired (the "marginal worker"), but other ("inframarginal") workers may receive larger wage premiums for risk than they would have received if their safety preferences had determined the wage rate. In effect the inframarginal workers may receive a "rent" that they would not necessarily require in order to take their job. This observation suggests that wage-risk studies may overestimate the wage-risk premium among many workers included in the studies.

When interpreting the results of wage-risk studies, it is important to observe the absolute levels of risk that are being studied. Imposing an additional mortality risk of 1 in 10,000 per year on workers who face baseline occupational risks of 1 in 1,000 per year may generate different valuations than if these same workers faced baseline

occupational risks of 1 in 10,000 per year. In Chapter 3 it was concluded that we should expect willingness-to-pay and willingness-to-sell values for identical increments in mortality risk to be greater in situations where the individual is already facing a high baseline risk of death. This prediction reflects the plausible assumption that the marginal utility of income is lower when workers are dead than when workers are alive. In other words, workers can be expected to be willing to forego more income for marginal risk reductions in situations where income is less important to them (i.e., when they already face a high baseline probability of death from all causes). Note that this theoretical prediction is valid only for an individual worker or a group of workers who have identical preferences.

In contrast to this prediction, wage-risk studies tend to find larger wage-risk premiums among workers who face smaller risks (Viscusi 1993). The statistical valuation of life has been shown to be higher among groups of workers facing baseline annual risks of 1 in 10,000 than for groups of workers facing baseline annual risks of 1 in 1,000 (Viscusi 1993). This finding is probably due to self selection in the sense that the workers with low valuations of safety are the most likley to accept risky jobs, i.e. the safety preferences differ between individuals at high and low baseline risks.

It is questionable whether the observed trade-off between risk and income in wage-risk studies is representative of the valuations of risk changes in the general population. It seems likely that workers with low valuations of health risks will choose hazardous jobs while those most averse to hazards will choose relatively risk-free jobs. In the many occupations where job-related risks are virtually zero, no information is available about the wage-risk trade-off. This means that wage-risk estimates are probably best suited for measuring the valuation of health risks among blue-collar workers where job risk is a more prominent issue.

5.2 Consumer Valuations Of Safety

Like the worker, the ordinary consumer faces numerous decisions in daily life where trade-offs are made between hazard and money. Examples of goods that are bought directly for safety purposes are smoke detectors and optional airbags. There are also goods such as cars where safety is one attribute that differs between models. It is therefore possible to study the actual purchasing decisions for goods which involve safety and try to infer the value of reduced health risks. While the literature on consumer choices is less extensive than the literature on wage-risk trade-offs, it is based on the same concept of revealed preference and it has attracted considerable interest among economists.

The earliest studies addressed the trade-off between time and risk rather than the trade-off between money and risk. Melinek (1974) noted that pedestrians who cross a busy street rather than taking the extra time to use a pedestrian subway reveal an upper bound on their personal value of safety. Ghosh et al (1975) argued that drivers

reveal their rate of trade-off between safety (which declines at higher speeds) and personal travel time when they decide how fast to drive. Blomquist (1979) examined the trade-off between the time it takes to buckle a seat belt and the risk reduction from using seat belts, suggesting that motorists with lower values of personal safety are less likely to fasten safety belts while those who confront smaller time costs will be more likely to fasten belts.

Studies of the trade-off between consumption of time and safety can be used to place a monetary value on a statistical life, but only if the monetary value of time can be quantified accurately. Using the after tax wage rate of the driver as a proxy for the value of time may be useful in a preliminary analysis, but there is good reason to believe that time is valued very differently by citizens depending upon the context (Mishan 1976). In the case of safety belt use, the discomfort and hassle associated with belt use may be a more significant factor than the time spent in fastening the belt. Since the conversion from time to money adds another uncertain factor to the valuation task, it would be preferable to find a more direct method of inferring the consumer's valuation of safety.

When deciding whether to purchase a small car instead of a large car, consumers implicitly place an upper bound on the value of their safety (and that of those who will use their car). Assume that the average annual probability of death is 15 in 10,000 with a small car and 10/10,000 with a large car, and that the consumers have perfect information about these risks. Consider the consumer who chooses a $12,000 new large car instead of a $8,000 new small car. Note that the $4,000 price difference amounts to an annual payment of $518 per year (assuming a ten-year vehicle life and a 5% real interest rate). Assume for the sake of simplicitly that the consumer is indifferent between the two cars apart from the difference in safety. On the basis of this choice we can then infer that the lower bound on the valuation of a statistical life of the consumer is 518/0.0005=$1,036,000. If the valuation per statistical life by the consumer was below this amount the consumer would have bought the small car instead. This simplified example demonstrates the logic of revealed preference, but in a more realistic situation the analysis would become much more complicated since it would be necessary to take into account all the other differences between cars except safety, e.g. comfort, space and fuel economy. It would also be necessary to estimate the perceived differences in safety between different cars, if consumers do not have perfect information.

Atkinson and Halvorsen (1990) analyzed new car choices using a hedonic price model that is similar in approach to the hedonic wage-risk studies. Housing prices have also been analyzed to infer price-risk trade-offs. For instance, mortality risks from air pollution have been used as an explanatory variable in cross-sectional studies of residential property values (Portney 1981). Not only do these studies of consumer choice suffer from the same basic limitations that plague hedonic studies of job choice (i.e. assumptions about perfect risk perception and a well-specified econometric

model) but there has also been a much less extensive amount of empirical work done on this subject.

Revealed preference has also been used to value risk reductions based on the direct purchase of products that enhance consumer safety (e.g. optional airbags). Markets for safety products are an appealing way to infer risk valuation, since these are cases where consumers directly buy risk reduction. An example of this type of product is smoke detectors, which reduce the probabilities of fire-related fatality and non-fatal injury. Two studies of the value of mortality risk reductions based on the purchase of smoke detectors have appeared in the literature (Dardis 1980; Garbacz 1989). In these studies the price of a smoke detector is compared with the expected risk reduction. In other words, the price of the smoke detector is assumed to be the willingness to pay for risk reduction among those individuals who buy smoke detectors.

Since buying a smoke detector is a discrete decision, precise inferences about maximum willingness to pay are impossible. Instead, the observed transaction price reveals a lower bound on the value of safety for an individual who buys a smoke detector and an upper bound on the value of safety for an individual who does not buy a smoke detector. As long as some consumers refrain from buying the smoke detector, the observed market price is thus not a lower bound on the value of safety in the population.

A difficulty with using smoke detector purchases to infer the value per statistical life is that a smoke detector reduces the risk of both non-fatal and fatal injuries due to fires. The revealed preference will thus be for the total risk reduction. Since the amount of safety offered by smoke detectors may not be accurately perceived by buyers, it is also difficult to interpret the choices of buyers unless risk perception data are collected and analyzed.

In order to infer the maximum willingness to pay for safety from direct purchases of safety products involving discrete choices, it is necessary for the price to vary between individuals or populations. If the price varies widely enough, it is possible to estimate the whole demand curve for a population and the total willingess to pay in the population can be estimated as the area below the curve. This can be compared with the case where the price is fixed and we can only discover one point on the aggregate demand curve.

One revealed preference study where the price varied between individuals is the study of indoor radon control by Åkerman et al (1991). In this study, the decisions of households were scrutinized to determine whether they had taken action to reduce indoor radon concentrations. Elevated indoor radon concentrations are believed to be a risk factor for fatal lung cancer. The cost of measures to reduce radon levels was estimated for each household, generating a distribution of costs for the various houses. On the basis of these data, the mean willingness to pay per household in the sample could be estimated. By using objective risk estimates of risk reduction, the value per

statistical life was estimated. The problem with the risk variable is the same in this study as in the previous ones, but the desired variation in the price variable is achieved in this study.

In the review by Viscusi (1992,1993) the value per statistical life varies between $0.07 million and $4 million in the revealed preference studies that have been carried out outside the labour market (1990 prices). The valuations are thus lower than for the job risks, and one explanation for this which is given by Viscusi (1992,1993) concerns the problems of estimating the maximum WTP for the risk change in these studies.

5.3 Conclusions

Most of the revealed preference studies so far have been on the wage-risk trade-offs on the labour market, but some studies have also been based on the purchase of safety products and other actual decisions which involve risks. The great advantage of revealed preference is of course that it is based on actual decisions, and it should be possible to apply this approach to more fields than those reviewed above. Examples of markets where the application of the revealed preference approach would be of great interest include the market for non-presription drugs and the market for health foods. These are examples of markets where individuals buy perceived health changes directly.

It is difficult to use the revealed preference approach to obtain a direct estimate of the value of health care programmes, since health care is seldom bought directly on a market. However, estimates of the value per statistical life or life-year taken from revealed preference studies can be used to value lives or life-years saved in health care programmes. The disadvantage of such an approach, apart from the estimation problems in revealed preference studies, is that the population which receives the health care programme may differ from the population studied in the revealed preference study. The safety preferences may thus differ between these populations and it is also possible that different types of risk reductions may be valued differently, e.g. reductions in the risk of heart disease death may not be valued the same as mortality risk reductions in traffic.

An alternative to using the revealed preference approach in order to estimate the WTP of health changes is to use expressed preference techniques. The expressed preference approach is discussed in the next section. The advantage of this approach is its flexibility, since it can be designed for the specific issue. On the other hand, it is not based on actual decisions by individuals, and traditionally economists have been very suspicious about using data based on what individuals say that they will do rather than what they actually do.

REFERENCES

Åkerman J, Johnson FR, Bergman L. Paying for safety: voluntary reduction of residential radon risks. Land Economics 1991;67:435-446.

Atkinson SE, Halvorsen R. The valuation of risks to life: evidence from the market for automobiles. Review of Economics and Statistics 1990;72:133-136.

Atkinson SE, Crocker TD. A bayesian approach to assessing the robustness of hedonic property value studies. Journal of Applied Econometrics 1987;2:27-45.

Blomquist G. Value of life saving: implications of consumption activity. Journal of Political Economy 1979;87:540-558.

Dardis R. The value of a life: new evidence from the marketplace. American Economic Review 1980;70:1077-1082.

Garbacz C. Smoke detector effectiveness and the value of saving a life. Economic Letters 1989;31:281-286.

Ghosh D, Lees D, Seal W. Optimal motorway speed and some valuations of time and life. The Manchester School 1975;43:134-43.

Hwang H, Reed WR, Hubbard C. Compensating wage differentials and unobserved productivity. Journal of Political Economy 1992;100:835-858.

Melinek SJ. A method of evaluating human life for economic purposes. Accident Analysis and Prevention 1974;6:103-114.

Mishan EJ. Cost-benefit analysis, 2nd edition. New York: Praeger, 1976.

Moore MJ, Viscusi WK. The quantity-adjusted value of life. Economic Inquiry 1988;26:369-388.

Portney PR. Housing prices, health effects, and valuing reductions in risk of death. Journal of Environmental Economics and Management 1981;8:72-78.

Smith A. The wealth of nations. New York: Modern Library, [1776] 1937.

Viscusi WK. Employment hazards: an investigation of market performance. Cambridge Massachusetts: Harvard University Press, 1979.

Viscusi WK. Fatal tradeoffs: public and private responsibilities for risk. New York: Oxford University Press, 1992.

Viscusi WK. The value of risks to life and health. Journal of Economic Literature 1993;31:1912-1946.

Viscusi WK, Moore MJ. Workers' compensation: wage effects, benefit inadequacies, and the value of health losses. Review of Economics and Statistics 1987;69:249-261.

Viscusi WK, Moore MJ. Rates of time preference and valuations of the duration of life. Journal of Public Economics 1989;38:297-317.

6. THE EXPRESSED PREFERENCE APPROACH

Expressed preference means that the data about WTP are not based on actual decisions, but on the preferences of individuals as expressed in hypothetical surveys. The method of measuring WTP or WTS in surveys is usually called the contingent valuation (CV) method.

The CV method has been developed principally in the area of valuing environmental benefits. The first person to use the method was Davis (1963). Davis interviewed a sample of 121 hunters in the Maine woods when he was trying to tease out the value of a recreational area. Ridker (1967) used the CV method to study the costs of air pollution. In the 1970s a number of applications appeared; the most notable of these were perhaps Darling's valuation of the amenities of public parks in Chicago (1973), Cicchetti & Smith's (1973) valuation of congestion costs in recreation areas, Acton's valuation of risk reductions from mobile coronary care units (1973), Hammack & Brown's valuation of the rights concerning the hunting of waterfowl (1974), Randall et al's valuation of air visibility (1974), and Hanemann's valuation of increased water quality (1978). This method is now the one most often used to value environmental benefits; for an extensive review of the environmental literature see Cummings et al (1986) and Mitchell & Carson (1989).

In the first section of this chapter open-ended CV questions are described. This is followed by a section about binary (yes/no) CV questions. Different forms of potential bias in a CV study are then discussed. After that we devote a section to the NOAA (National Oceanic and Atmospheric Administration) panel report and the NOAA proposed regulations for the measurement of non-use values of the environment. We then review the CV studies in the health area and also describe in detail a health care application of the CV method. The chapter ends with some conclusions.

6.1 Open-Ended CV Questions

CV questions can be divided into open-ended and binary valuation questions. In open-ended valuation questions the researcher tries to measure the maximum willingness to pay of each respondent, and in binary valuation questions the respondent accepts or rejects only one price (bid) level for the good.

Interviews (face-to-face or on the telephone) or mail questionnaires can be used for eliciting willingness to pay with open-ended questions. It is possible to ask the respondent directly for the maximum WTP in an open-ended question, but usually some kinds of aid are used to make it easier for the respondent to answer the valuation question. The reason for using these aids is that it has been found to be difficult to get respondents to state a maximum WTP directly, leading to problems of non-response.

One aid that is used in interviews is the so-called bidding game, introduced by Randall et al (1974). A bidding game resembles an auction. A first bid is made to the respondent, who accepts or rejects the bid, and then the bid is raised or lowered depending on the answer. The process goes on until the respondent's maximum willingness to pay is reached.

Payment cards can also be used, which include a range of values for the respondent to choose from. It should be mentioned that it is also possible to use bidding-game or payment-card type questions in mail questionnaires, where the respondent is provided with a number of different amounts to choose between. The major problem when using a bidding game is that it involves the risk of starting-point bias (see the discussion about different forms of bias below), which means that the respondent's willingness to pay is affected by the first bid made. Payment cards can lead to similar problems, since the amounts which the respondent has to choose between may affect the valuation. The problems with starting point bias in CV studies led researchers to experiment with binary valuation questions, where the respondent is only given one bid to accept or reject.

6.2 Binary CV Questions

An alternative to the open-ended approach is to use binary valuation questions. Binary CV questions are of the yes/no type, which means that the respondent accepts or rejects a bid. By varying the bid in different sub-samples it is possible to calculate the proportion of respondents who are willing to pay as a function of the price (bid), and this curve can be interpreted as an aggregate demand curve for the good if the proportion is multiplied by the number of respondents. Binary contingent valuation questions are sometimes also referred to as the referendum approach, since the question in some cases can be phrased as a vote on a referendum.

Binary valuation questions were first used in the classic study by Bishop & Heberlein (1979) concerning the value of goose-hunting permits, and this is now the most commonly used elicitation technique. An advantage of this approach is that it resembles a market situation for the respondent, since individuals are accustomed to deciding whether or not to buy a good at a specific price. The binary approach also avoids the problem of starting point bias, since the individual is only given one bid.

The disadvantage of the binary approach is that the information received from each respondent is less than in the case of open-ended questions. With a binary question the only information that is received from each respondent is whether the WTP is greater or smaller than the bid. In order to estimate the mean WTP on based the basis of binary CV data, the relationship between the bid and the proportion of respondents who are willing to pay has to be estimated.

In Figure 1 this is exemplified by showing the proportion of respondents who are willing to pay as a function of the bid (price). In Figure 1 the mean WTP is the area below the curve and the median WTP is the price where the proportion of acceptance is 0.5, i.e. the price where 50% would answer yes and 50% would answer no to the CV question.

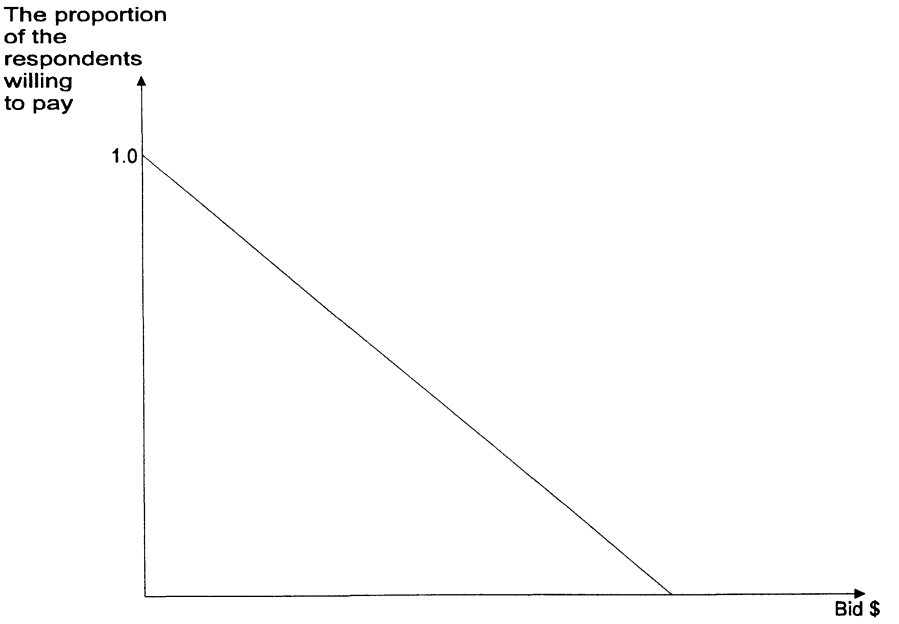

Figure 1. The proportion of respondents willing to pay

In order to estimate the curve in Figure 1, either regression analysis or non-parametric methods can be used. If regression analysis is chosen then the answer to the binary CV question is used as the dependent variable and the bid is used as the explanatory variable together with other explanatory variables of interest. Since the answer is in the form of a yes/no answer this is not a continuous dependent variable and therefore regression techniques for dummy dependent variables (i.e. 0/1 variables) should be used.

The most common regression technique which has been used to analyse binary CV data is logistic regression (Hanemann 1984; Cameron 1988). With logistic regression the probability of acceptance of a bid is calculated as a function of the bid, i.e. the curve in Figure 1 is estimated. The logistic function has the following form:

$$P=1/(1+e^{-X}) \qquad (1)$$

where X is the regression equation to be estimated and P the probability of acceptance of the bid. In the simplest case with only the bid as the explanatory variable, X in Equation 1 is $\alpha+\beta 0$BID+error term, where α is the constant and $\beta 0$ is the coefficient for BID. In this case the mean WTP is equal to $-\alpha/\beta 0$, and the mean is also equal to the median.

This can be seen since the logistic function in Equation (1) can be rewritten as $\ln[p/(1-p)]=\alpha+\beta 0$BID (the mean of the error term is zero, so it is suppressed here). By setting $p=0.5$ the equation can be rearranged to yield BID$=-\alpha/\beta 0$. The bid where 50% of the respondents would say yes is the median WTP and in this case the median and the mean coincide so that this gives the formula for both the mean and the median WTP.

If more explanatory variables are included as well as the bid, they can be entered in the regression as their deviations from the mean of the variable in the sample, and the mean can still be estimated using the above simple formula. If the explanatory variables are entered directly without this adjustment, the mean (and median WTP) can be estimated by setting $p=0.5$ and solving for BID at the mean values of the explanatory variables.

It should be mentioned that using the logistic function in the above way does not rule out negative WTP. The above formula can, however, be adjusted to rule out negative WTP and instead allow some individuals to have a zero WTP. In that case the formula for the mean WTP becomes: $-(1/\beta 0)*\ln(1+e^{\alpha})$, which corresponds to integrating the logistic function from 0 to infinity rather than from negative to positive infinity as in the estimation above (Johansson 1995).

In some studies ln BID is also entered as an explanatory variable instead of the BID and this will rule out negative WTP. As long as the coefficient for ln BID is below -1, the formula for the mean WTP in that case becomes: $e^{-\alpha/\beta 0}*[(-3.1416/\beta 0)/(\sin(-360/2*\beta 0)]$ (Hanemann 1984, Johansson 1995). If the coefficient of ln BID is not below -1 the integral does not converge and the mean WTP becomes infinitely large (see Hanemann (1984) for details). Due to the problems of non-convergence and extremely high mean WTP estimates based on the ln BID model, it can be useful to plot the proportion who are willing to pay as a function of BID (i.e. to plot Figure 1) and to integrate the area below the curve between zero and what is considered to be the maximum possible WTP. Since the maximum WTP is often uncertain, sensitivity analysis can be used to investigate how the mean WTP changes with different upper limits of integration. The similarities between the computation of willingness to pay using binary data and the computation of life expectancy should be noted. The function that is calculated can be viewed as a survival function with respect to willingness to pay (bid). It is also possible to normalize the estimated cumulative distribution function to reflect the truncation at the highest WTP

amount, and to estimate the mean WTP on the basis of the normalized function (see Boyle et al (1988) for details). Such an approach is based on the assumption that the proportion of individuals who accept the maximum WTP in the study have the same mean WTP as the other individuals in the study, rather than assuming that their WTP is equal to the maximum WTP.

Since the mean WTP based on binary CV questions can be sensitive towards the choice of the functional relationship between the bid and the probability of accepting the bid (Hanemann 1984; Boyle & Bishop 1988; Bowker & Stoll 1988; Kriström 1990), it can also be useful to estimate the mean WTP using a non-parametric method (Kriström 1990). The non-parametric method was developed by Kriström (1990) and it is a simple way of estimating the mean WTP. In this approach the proportion of yes answers at each of the different bid levels is used to estimate the curve in Figure 1.

The use of the non-parametric method is illustrated with a hypothetical data set in Table 1. In Table 1 it is assumed that we have a data set with 1,000 respondents. The 1,000 respondents are divided into 10 sub-samples with 100 respondents in each. The first column shows the bid, which was varied between $0 and $900. The second and third columns show the numbers of yes and no answers at each bid level. The proportion of yes answers at each bid level is used to estimate the mean WTP on the basis of the non-parametric method. This is shown in column four. However, if the proportion of yes answers increases at any bid level, the mean of the proportion of yes answers at two or more bid levels should be estimated until a non-increasing proportion of yes answers is obtained.

Table 1. Illustration of the non-parametric method

Bid ($)	Yes answers	No answers	Proportion yes	Non increasing proportion yes
0	100	0	1.00	1.00
100	90	10	0.90	0.95
200	100	0	1.00	0.95
300	85	15	0.85	0.85
400	70	30	0.70	0.70
500	50	50	0.50	0.50
600	40	60	0.40	0.40
700	25	75	0.25	0.25
800	10	90	0.10	0.10
900	0	100	0.00	0.00

In Table 1 we can see that the proportion of yes answers is 0.9 for the $100 bid and that it increases to 1.0 for the $200 bid. The data should then be combined for the two bid levels to calculate the proportion of yes answers. This has been done in the fifth column in the table, where the proportion of yes answers is 0.95 for both the $100 bid

and the $200 bid. This process is continued until a non-increasing proportion of yes answers is achieved. In the table the non-increasing proportion of yes answers is shown for the hypothetical data set. The proportion of yes answers is then plotted as a function of the bid, and linear interpolation is used to connect the different points on the curve. This has been done in Figure 2, for the hypothetical data set.

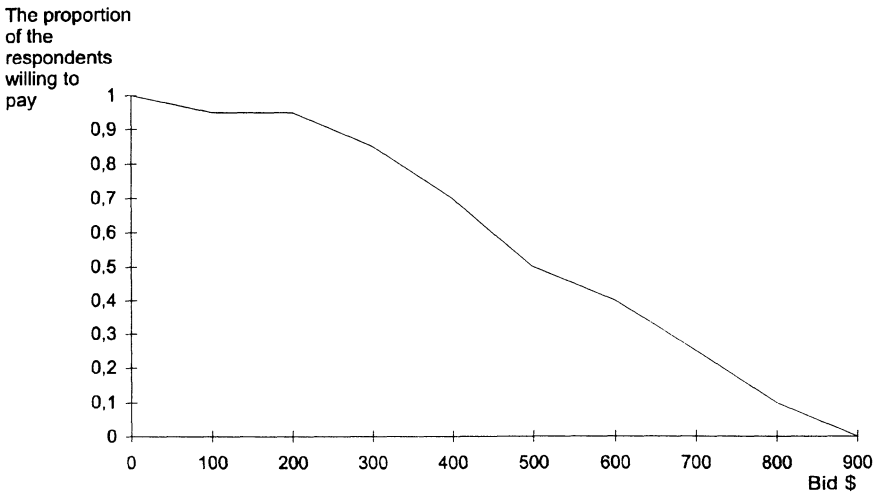

Figure 2. Estimation of the mean WTP

In Figure 2 the mean WTP based on the hypothetical data set can be estimated as the area below the curve. The median WTP is equal to the bid where the proportion of yes answers is 0.5 (i.e 50% answer yes and 50% answer no). In this hypothetical example we assumed that no respondent was willing to pay the highest bid. If this is not the case, we have to make an assumption about the maximum WTP in order to estimate the mean WTP; as above, sensitivity analysis can be used to test the sensitivity towards this assumption.

In some cases it may happen that some respondents are not willing to pay anything for the good, i.e. they have a zero WTP. If we have information about the proportion of respondents who are not willing to pay anything, this can be incorporated into the estimation of mean WTP by using a spike at zero, e.g. if 30% have a zero WTP the curve in Figure 2 would start at 70% for a zero bid. The advantage of the non-parametric approach is that it does not impose a specific functional form on the

relationship between BID and the proportion who are willing to pay. It is also easy to use and it makes the assumptions in the calculation of the mean WTP explicit.

The disadvantage of the non-parametric approach is that it is not possible to calculate the effect of different explanatory variables on willingness to pay, e.g. to test whether WTP is related to income. In practice it is therefore useful to employ both the non-parametric method and regression analysis in order to analyse the data from a binary CV study.

In a binary CV study the design of the bid vector (i.e. the set of bids that is used) is very important. It is important to get information about the whole distribution of WTP, so as to decrease the uncertainty in the estimation of mean WTP; for example, if some respondents answer yes to the highest bid in the study then some assumption has to be made about the maximum WTP of this group, and if the group that answers yes to the highest bid is large this will lead to great uncertainty in the estimation of the mean WTP. It is thus important that the bids should cover the whole distribution of WTP, and in order to obtain information about the distribution of WTP it is important to carry out pilot studies. The bid vector can also be designed so that the maximum possible information is received for a given sample size (See Minkin (1987) and Nyquist (1992)). If possible, it is also useful to design the bid vector so that it is updated during the study. For instance, one bid vector can be used in the first half of the study, and then updated in the second half if necessary.

It has been noticed that in regression analysis of binary CV data, the median willingness to pay obtained from binary CV data is relatively stable with regard to assumptions concerning the functional form of the relationship between the bid and the proportion who are willing to pay, whereas the mean is highly sensitive towards assumptions about the functional form (Hanemann 1984; Mitchell & Carson 1989; Kriström 1990). This has led some researchers to recommend the median rather than the mean as the welfare measure in cost-benefit analysis using CV data (Hanemann 1984). However, the mean is the correct welfare measure according to cost-benefit analysis, since if the mean WTP is multiplied by the number of persons in the population who receive the health care programme, this yields the total WTP of this population (Johansson et al 1989). There are therefore statistical advantages to using the median, but the mean is the strictly correct welfare measure in cost-benefit analysis.

The mean often exceeds the median since the distribution of WTP is skewed to the right by some individuals who are willing to pay very high amounts. One argument for using the median could be that this skewed distribution is considered to be an artifact of CV studies, and that the median WTP is a better estimate of the real mean than the mean observed in a CV study.

The general recommendation in the literature at the moment is to use binary CV questions rather than open-ended CV questions. This is because the binary questions

avoid the problem of starting point bias and more closely resemble a market situation for the respondents. It has also been argued that binary CV questions give less incentives for strategic answers (see below) than open ended questions (Mitchell & Carson 1989; Carson 1991). It should be noted that the problem with starting point bias can be avoided if the respondent is asked directly for the maximum WTP without payment cards or a bidding game. Such a question is, however, perceived as being very difficult to answer, and this leads to problems of non-response; it is also perceived as unrealistic.

In studies that have compared binary and open-ended CV questions, the mean WTP has been higher with the binary approach (Kriström 1989,1990; Sellar et al 1985; Johannesson et al 1991). Since open-ended and binary CV questions have not yet been calibrated against real decisions, it is impossible to say with certainty that binary CV questions are more valid than open-ended CV questions, but according to the current recommendations of practitioners the binary CV approach is to be preferred. In CV studies of health programmes it thus seems most appropriate to use the binary approach until further evidence is available that may change this recommendation. To clearly resolve the issue, however, studies are needed where both types of question are compared with real money transactions.

Recently some researchers have also started to use double-bounded binary CV questions (Hanemann et al 1991). In this approach an individual first responds yes/no to one bid; then if the answer is yes the respondent is given one higher bid, and if the answer is no the respondent is given one lower bid. Each respondent is thus given two bids, and the idea is to increase the amount of information from each respondent. The problem with this approach is that it introduces the potential for starting-point bias (see below), in the same way as a bidding game (i.e. the approach is a bidding game with only two bids). It thus has the same problems as the bidding game and at the same time less information about the WTP is received than in a bidding game. Thus if this approach is to be favoured, it has to be argued that starting point bias is less of a problem with only two bids than with more than two bids. We are not aware of any study that tests the presence of starting point bias in the double-bounded approach or compares this approach to the use of only one binary question.

The survey instrument in a CV study is of the utmost importance. When constructing the instrument there are a number of sources of potential bias that it is important to bear in mind. In the next section we will review these biases, using the bias typology developed by Mitchell & Carson (1989).

6.3 Potential Bias In A CV Study

The sources of potential bias in a CV study can be divided into five main areas according to the bias typology of Mitchell & Carson (1989): incentives to misrepresent responses; implied value cues; scenario misspecification; sample design and execution

biases; inference biases. The first three areas concern the design of the CV instrument, whereas the other two areas concern more general issues to do with the sampled population and the way the results are used. We will focus mainly on the design issues below. All biases relate to validity issues, i.e. systematic errors. Reliability, i.e. random variance in answers, can be assessed through replications of a CV study.

Incentives to misrepresent responses can be divided into strategic bias and compliance bias. Strategic bias means that the respondent feels that it is in his/her self-interest to give a lower or higher valuation than the true value. The well-known free-rider problem in economics is an example of strategic behaviour, where individuals try to avoid paying for public goods since they cannot be excluded from consumption of the public good. Strategic bias can produce both higher and lower valuations than the true value. If the respondent thinks that the provision of the good is dependent upon the valuation of the good, but that his/her level of payment for the good is independent of the value given, there is an incentive to exaggerate the valuation. If the respondent on the contrary feels that the provision of the good is independent of the value given, but that the level of payment is dependent on his/her valuation (bid), there is an incentive to understate the valuation.

In the empirical studies that have tried to establish whether strategic bias exists, it has never been clearly demonstrated that strategic bias is present to a large extent (Rowe et al 1980; Bohm 1972; Scherr & Babb 1975; Brookshire et 1976; Smith 1977,1979; Milon 1989). It has also been argued that binary CV questions are not as susceptible to strategic bias as open-ended questions. The reason for this is that in a binary CV question, the respondent only answers yes or no to one bid level, and it is therefore more difficult to have a large effect on the mean WTP compared with an open-ended CV question. However, if all the respondents answered CV questions in a strategic way and, for instance, gave a WTP which was 50% lower than their true valuation, then the results of open-ended and binary CV studies would be affected to the same extent.

Some researchers have claimed that the binary CV question not only reduces the strategic bias problem, but also provides the respondent with incentives to give a true valuation (Mitchell & Carson 1989; Carson 1991). The reasoning behind this argument is that if the binary CV approach is phrased in terms of a public referendum, it is optimal to answer yes if the WTP exceeds the bid and it is optimal to answer no if the WTP is below the bid. However, for this to hold, the respondent has to treat the survey as a real referendum and believe that if the project is carried out, the actual payment for the respondent will equal the bid in the CV study. On the other hand, so long as CV studies are hypothetical there is no reason to believe that the binary approach is incentive-compatible in this respect. In general the incentives for strategic answers seem to be the same for both the binary and the open-ended approach, but the possibility of a single individual affecting the result of a CV study by answering in a strategic way appears to be less for binary than open-ended CV questions.

The most well-known type of compliance bias is interviewer bias. Interviewer bias is a well-known phenomenon in survey research and means that the respondent overstates or understates his/her true valuation in an attempt to please the interviewer. However, those few studies that have compared willingness to pay elicited through, on the one hand, mail questionnaires and, on the other, interviews using the same questions, have found no statistically significant difference (Schulze et al 1983; Sorg & Brookshire 1984). Sponsor bias is another form of compliance bias, indicating that the respondent gives a valuation that differs from his/her true valuation in an attempt to comply with the presumed sponsor of the investigation, e.g. if the respondent thinks that the study is sponsored by Greenpeace the respondent may try to give the answer that he/she thinks Greenpeace wants.

Another major area of bias is implied value cues, which are a form of anchoring. Implied value cues can be divided into starting-point bias, range bias, relational bias, importance bias and position bias. Implied value cues exist when some information in the CV instrument implies a certain value for the good. Starting-point bias is present when the respondent's valuation is affected by some potential willingness-to-pay amount that is given in the study. The standard case is that the respondent is affected by the first bid in a bidding game. Starting-point bias has been shown to be an important problem in CV studies using bidding games, and that is why this approach is seldom used any more (Rowe et al 1980; Brookshire et al 1980,1981; Thayer 1981; Boyle et al 1985).

Range bias is similar to starting-point bias, with the difference that there is a range of potential willingness-to-pay amounts that is introduced in the survey and affects the value given by the respondent. So-called payment cards are the typical example where the respondent can choose one amount from a range of values. Relational bias means that the description of the good being valued includes information about its relationship to other goods, which influences the amount given by the respondent. Importance bias can arise if the valuation instrument suggests to the respondent that the good being valued is particularly important and valuable. This may lead the respondent to exaggerate his/her true valuation. If the ordering of questions suggests to the respondent some value of the good that influences the valuation, there is a case of position bias.

Scenario misspecification occurs when the respondent does not respond to the correct contingent scenario, i.e. the CV question is formulated incorrectly. This group of biases can be divided into theoretical misspecification bias, amenity misspecification bias, and context misspecification bias. Theoretical misspecification occurs if the scenario is not correct with regard to economic theory; the scenario can for instance be incompatible with the policy change being valued.

Amenity misspecification occurs when the good being valued by the respondent differs from the one intended by the researcher, i.e. the respondent misunderstands the question. The respondent can for instance express an attitude or belief instead of a

valuation for the good since the question is perceived in a symbolic way. Another possibility is that the respondent values a different quantity or quality of the good than that intended by the researcher.

Context misspecification bias is also a case of misunderstanding of the question, but in this case it is not the good, but the context of the market, that is perceived in a way that differs from what was intended. Parts that can be perceived in a manner not intended by the researcher are for instance property rights and the payment vehicle (i.e. the way in which it is stated that the respondent will pay for the good, e.g. through an increased tax or a direct user charge).

Studies that have compared different payment vehicles support the hypothesis that the payment vehicle affects the valuation (Randall et al 1978; Rowe et al 1980; Brookshire et al 1980,1981; Daubert & Young 1981; Greenley et al 1981; Cronin 1982). It is not obvious, however, that individuals are indifferent as to the choice of payment vehicle. The payment vehicle can therefore be viewed as a part of the good, which would mean that payment vehicle bias only arises if the payment vehicle is misperceived by the respondent or valued in a way not intended by the researcher. Context misspecification bias is also present if the order of the questions affects the valuation or if the respondent gives a "reasonable" valuation instead of the greatest willingness to pay.

Sample design and execution biases are mainly statistical problems. This type of bias will arise, for instance, if the population chosen does not correspond to the population to whom the benefits will accrue, or if the sample is non-random. The other forms of bias in this group concern the problem of non-respondents. There will be biased estimates of willingness to pay if respondents' willingness to pay is not representative of that of the sampled population. If there is non-response in a CV study, some assumption has to be made concerning the WTP of the non-respondents. If the assumed WTP of the group of non-respondents differs from the actual WTP of the non-respondents then the WTP of the CV study will be biased.

The last bias group according to this typology is inference bias. Inference bias can arise if preferences have changed from the time of the CV survey to the time the results are used for decision making. There will also be inference bias if there is an adding together of the willingness to pay for each of a number of different goods in order to evaluate a policy package, but these goods were valued independently of each other. As was stated in Chapter 3, the WTP for goods that were evaluated independently cannot in theory be added so as to obtain the total WTP for the goods.

The above bias typology differs from the one that was used by researchers prior to the publication of the Mitchell & Carson (1989) textbook about CV, see for instance Cummings et al (1986). The most important change is the inclusion of information bias and hypothetical bias, which were not mentioned in the typology presented above. It is doubtful whether what used to be called information bias can in fact be termed a bias, since the answer to a question in a CV study is contingent upon the information

given. If the information changes, the good that is valued changes. Individuals may have preferences about the payment vehicle, about who provides the good, about how it is provided, about who else is paying etc (Mitchell & Carson 1989).

Mitchell & Carson (1989) also argued that the hypotheticalness of the situation in a CV study can increase the variance in the answers, but that there is no evidence from CV studies that the hypotheticalness of the situation would bias the estimates in any systematic way. This would mean that the hypotheticalness of the situation is a random error and not a systematic bias.

In the last few years some evidence has surfaced, however, that the WTP amounts based on hypothetical CV studies are higher then the WTP amounts based on actual decisions (see the discussion about the NOAA panel report and the NOOA recommendations below). Thus if individuals systematically overestimate the WTP in hypothetical CV studies, this may be a form of hypothetical bias that is systematic. Hypothetical bias should therefore perhaps be reintroduced as a form of bias in the bias typology of CV studies, unless this possible overestimation can be explained by some of the other bias forms above such as strategic bias.

Carson (1991) has put together five conditions for a valid CV scenario. These conditions should not be viewed as sufficient to guarantee the validity of a CV scenario, but it may be useful to think about them in the design of a CV study. These criteria are that the scenario should be theoretically accurate, policy relevant, understandable by the respondent as intended, plausible to the respondent, and meaningful to the respondent. Perhaps the most important feature of these criteria is that they illustrate the importance of a plausible and meaningful scenario to enhance validity. The first two criteria can be fulfilled by specifying the problem correctly, and for the other three criteria to be fulfilled pilot testing of the instrument is of the utmost importance.

In the field of environmental economics, there has been a big controversy regarding the use of the CV method to measure existence values. Existence values are values of environmental resources that are not directly related to any use of the resources, e.g. if someone is willing to pay to preserve a wilderness area even though he/she never ever expects to visit the area. Existence values can be viewed as a form of altruism, and the issue is thus similar to the issue about altruistic WTP for health programmes.

After the Exxon-Valdez oil spill in Alaska, the state of Alaska appointed some CV practitioners to assess the loss of existence values due to the oil spill using CV. The intent was to use these values as part of the damage assessment in court. On the other hand, Exxon-Valdez appointed a group of economists to show that the CV method could not be used to assess existence values. A very fierce debate followed between these different groups.

Following this debate, the government agency responsible for damage assessments in connection with oil spills, National Oceanic and Atmospheric Administration (NOAA), appointed a contingent valuation panel of economic experts to evaluate the use of the CV method in determining existence values. This panel was co-chaired by Kenneth Arrow and Robert Solow. The report of this panel and the proposed regulations from NOAA for natural resource damage assessments are reviewed below.

This does not directly concern the value of health programmes, but many of the issues concerning the validity of the CV method that are discussed in the report (National Oceanic and Atmospheric Administration 1993) and in the proposed regulations (National Oceanic and Atmospheric Administration 1994) are also of great interest for the use of the CV method to value health changes.

6.4 The NOAA Panel Report

The NOAA panel report (National Oceanic and Atmospheric Administration 1993) identified a number of problems with the CV method, but at the same time it stated that the CV method seemed to yield some useful information. One of the problems raised was the fact that there are some indications that hypothetical WTP exceeds real WTP. According to the panel, in the few studies that have been able to compare real and hypothetical WTP of environmental goods in experimental studies, it appears as though hypothetical WTP exceeds real WTP.

An example of such a study was done in Norway by Seip & Strand (1992). In this study the CV method was used to assess the WTP for membership in a Norwegian organization devoted to environmental affairs, and the WTP in the CV study was compared with the WTP when some of the same individuals were offered the possibility of actually joining this organization. The study found that hypothetical WTP was lower than actual WTP.

A similar study was carried out by Duffield & Patterson (1991), who studied the WTP for the maintenance of stream flow in two Montana rivers that provided spawning grounds for two rare species of fish. In two parrallel samples the hypothetical and real WTP was investigated, and it was found that hypothetical WTP exceeded real WTP.

The panel also stated that some studies have also been carried out on ordinary market goods with similar results (e.g. Dickie et al 1987). On the basis of the results of these studies comparing real and hypothetical WTP, the panel concluded that CV studies tend to overestimate WTP, but that the overestimation may be systematic so that the WTP based on CV studies can be discounted for this overstatement.

Another problem raised by the panel was the lack of sensitivity towards the size of the environmental programme. In some studies the WTP is about the same irrespective of the size of the programme (e.g. the WTP for cleaning up one lake in an area versus the

WTP for cleaning up all lakes in an area) (Kahnemann 1986; Kahnemann & Knetsch 1992; Diamond et al 1993; Desvousges et al 1993). On a theoretical basis we would expect the WTP to increase with the size of the programme, and if such an increase cannot be demonstrated in CV studies the method would be of no use in cost-benefit analysis.

This problem is sometimes also referred to as embedding, since the respondents do not differentiate between the scales of the programmes. A related problem raised by the panel is the so called "warm-glow" effect. It has been argued by some critics of the CV method that what individuals express in CV studies is not a valuation of the environmental change, but that instead responses serve the same functions as donations to charities (Diamond & Hausman 1993). A CV response would then express some kind of general approval of the environmental programme rather than a valuation. This could be an explanation for the fact that the WTP does not vary with the size of the programme.

In terms of the bias typology of Mitchell & Carson (1989), the problems of demonstrating that the WTP varies with the size of the programme and the "warm-glow" effect would be examples of scenario misspecification, i.e. the respondent answers some question other than that intended by the researcher.

A further problem in CV studies which was discussed by the panel is the absence of a meaningful budget constraint. The problem here is that individuals may answer CV questions without thinking about where the money would come from, i.e. what alternative consumption they would reduce. The panel criticises CV studies for not reminding respondents about their budget constraints, i.e. the fact that they would have to reduce some other spending in order to pay for the environmental good.

The panel also sets out a number of guidelines for CV studies that will be used to provide information to a damage assessment process. They recommend the use of the binary referendum approach to eliciting WTP in surveys. Thus they do not recommend the use of open-ended questions, since these are viewed as difficult to answer and unreliable. They also recommend the use of interviews in CV studies rather than mail surveys. It is stated that face-to-face interviews are the preferred approach, but that telephone interviews may also be acceptable in some cases. They also say that the budget constraint should be imposed on the respondents by reminding them of alternative spending possibilities and making it clear that the payment will reduce the consumption of other goods.

They further recommend that a no-answer (don't know) option should be added to the binary CV question, so that individuals are not forced to state a yes or no response. It is also recommended that the answers to the valuation question are followed up with a question about the reason for the response. The follow-up question could serve as a way to sort out the responses that reflect valuations of the environmental good from

protest bids and other types of answer that do not reflect a valuation of the environmental good.

If the reason for a no-answer is that "It's not worth it", this might for instance indicate a valuation, whereas if the stated reason is for instance that "The oil companies should pay", this would indicate that the no-answer does not reflect a valuation of the good. By including the question where the respondent explains the choice it may also be possible to divide the "don't know" answers into those which are indifferent at that bid level and those which are for other reasons. The panel also recommends what they call a conservative design of CV studies. This means that when there is ambiguity about some aspect of the survey or the analysis of the data, the option that will tend to underestimate WTP should be chosen. This also means, for instance, that WTP rather than WTS should be used in CV studies.

Following the report of the NOAA panel on contingent valuation, NOAA has also issued proposed regulations concerning the use of CV studies as a basis for natural resource damage assessment (National Oceanic and Atmospheric Administration 1994). In this document it is stated that existence values (also called passive use values) should be included as a component of damage assessment and that CV studies can be used to value existence values for damage assessment as long as the CV studies follow the standards in the proposed regulations.

The standards in the proposed regulations rely heavily on the report of the expert panel. NOAA recommends the use of face-to-face interviews and the use of the binary referendum CV format. NOAA also proposes that the response rate should be at least 70% in a CV study in order to be considered as a basis for damage assessment. Furthermore it is proposed that the sensitivity to the size of the programme should be demonstrated in the CV study using different subsamples, i.e. it must be demonstrated using different subsamples that the WTP increases with the size of the environmental programme. It is also proposed that the WTP from a CV study should be reduced by 50%, due to the indications that hypothetical WTP exceeds real WTP. These are the most important of the proposals in the proposed NOAA regulations. A lot of the recommendations are of the "rule of thumb" type, e.g. a 70% response rate and that WTP should be reduced by 50% due to hypothetical bias.

A lot of the issues raised in connection with the NOAA panel report and the proposed regulations by NOAA are also of interest for the valuation of health programmes. For instance, it seems to be a minimum requirement that if the results of a CV study are to be used in a policy context then it must be demonstrated that the WTP increases with the size of the programme, e.g. the size of the risk reduction in a health programme. It also seems useful to reinforce the budget constraint by reminding individuals that the payment has to be taken from their own disposable income, and to enquire into the reasons for different types of responses in order to try and sort out invalid responses.

What seems most important in CV research at the moment, however, is to compare hypothetical payments with real money payments so as to analyse the relationship between hypothetical and real WTP. Even if hypothetical WTP overestimates real WTP they may be systematically related, so that real WTP can be predicted from hypothetical WTP. However, recommending the use of a 50% reduction of hypothetical WTP, as in the proposed regulations by NOAA, seems somewhat arbitrary.

6.5 CV Studies Of Health Changes

Although the number of CV studies of health changes is small compared to the number of applications of environmental changes, several studies have been carried out. Below we will briefly review these studies, before we examine one study in the health care field in more detail.

The first study of the CV method in the health field is the study by Acton (1973). Acton investigated the WTP for mobile coronary care units which would decrease the risk of dying after a heart attack. Jones-Lee (1976) also carried out an early study of the value of airline safety. Both these early studies were explorative in nature and used very small samples. Two later and much larger studies investigating the value of mortality risk reductions are those of Jones-Lee et al (1985) and Smith and Desvousges (1987).

Jones-Lee et al (1985) used the CV method to investigate the value of reductions in the risk of traffic deaths in the UK. A random sample of 1150 individuals was used, and 1103 interviews were carried out. A follow-up study was also carried out of a sample of 210 individuals from the original sample, to test the reliability and stability of the answers. The results showed a value of a statistical life of about 1.5 million pounds (1982 prices) with only WTP for own risk reduction included, and about 2.0 million pounds when the WTP for reductions in other peoples' risks was included. The authors also noted that the results as a whole conformed to the theoretical predictions, and that individuals seemed to be able to understand probabilities.

Smith and Desvousges (1987) used the CV method to value risk changes from hazardous waste. A representative sample of 720 households in Boston, USA was used in the study and 609 interviews were carried out with these households (i.e. a non-response of about 15 %). The risk of dying from exposure was divided into two parts: the risk of exposure and the risk of dying given that one is exposed. The sample was divided into eight different subsamples confronted with different risk reductions and different base line risks. The variation in WTP between the different subsamples was small; thus the WTP did not increase with the size of the risk reduction. This study demonstrates the problem discussed on the NOAA panel, that the WTP does not increase with the size of the programme. However, within each subsample the WTP did increase with the size of the risk reduction.

Contrary to the theoretical predictions, the WTP for an identical risk reduction was also found to decrease with the level of the risk, i.e. according to an expected utility model we would expect the WTP for a marginal risk reduction to increase with the baseline risk (see Chapter 3). The regression analysis also indicated that individuals exaggerate small risks. Moreover, the WTP for a marginal risk reduction was greater than the WTP for avoiding an identical marginal risk increase.

The results were thus contrary to most theoretical predictions, which contrasts with the results and conclusions of the study by Jones-Lee et al (1985). However, the fact that the risk was divided into two components (exposure and mortality risk if exposed) may have led to problems for the individuals when interpreting the actual risk reductions. In comparing the results from Smith and Desvousges (1987) and Jones-Lee et al (1985) it also seems as if results are consistent within samples, i.e. the WTP increases with the risk reduction within a sample, but not between samples. (The latter was not studied in Jones-Lee et al [1985]). This fact may indicate that individuals have difficulties in differentiating between risk reductions involving small probabilities unless they are given some kind of reference (anchoring) point, as when the WTP is elicited for several risk reductions within a sample. To test if the WTP increases with the size of the risk reduction, different subsamples should be used as in the study by Smith & Desvousges (1987), so that the result is not just an artefact due to anchoring.

Both the studies by Jones-Lee et al (1985) and Smith and Desvousges (1987) concerned mortality risks, but CV studies of morbidity risks are also starting to appear in the literature, e.g. the study by Viscusi et al (1988) about the WTP to reduce the risk of insecticide inhalation skin poisoning, the study by Evans & Viscusi (1991) about the WTP to reduce the morbidity risks of pesticides and toilet bowl cleaners, and the study by Viscusi et al (1991) about a reduction in the risk of chronic bronchitis.

In his review of the value of risks to life and health Viscusi (1992,1993) lists six CV studies on mortality risks and in these studies the value per statistical life varies between $0.1 million and $15.6 million (1990 dollars). He notes that the value per statistical life in the large scale CV studies conforms quite well with the range of $3 to $7 million given for the most reasonable estimates from the wage-risk studies. A similar conclusion is also reached in the review of the literature by Fisher et al (1989). Viscusi (1992,1993) also includes eight CV studies on valuations of nonfatal health risks.

The studies included in the review of Viscusi (1992,1993) includes mainly studies of environmental health risks. Studies using the CV method has, however, also started to appear in the health care field. Examples of this apart from the early study by Acton (1973) mentioned above are Thompson et al's (1984) and Thompson's (1986) studies concerning the WTP for a hypothetical cure for arthritis, the study by Berwick & Weinstein (1985) on the WTP for ultrasound in normal pregnancy, the study by Reardon & Pathak (1989,1990) on the WTP for antihistamine drugs, the study by Donaldson (1990) on the WTP for nursing home care for the elderly, and the studies

by Johannesson et al (1991,1993) on the willingness to pay for antihypertensive drug treatment. The studies in the health care field have largely been feasibility studies testing the acceptability of the CV approach in the health care field. Below we will review one of these studies in some detail.

6.6 A Health Care Application of the CV method

Johannesson et al (1993) analysed the WTP for antihypertensive drug therapy. This was a follow-up study to the one by Johannesson et al (1991) which compared open-ended and binary CV questions in a group of hypertension patients. In the latter study, the open-ended question did not work well due to problems of non-response and implied value cues, so the follow-up study used only the binary approach.

In order to estimate the willingness to pay for antihypertensive therapy, a mail questionnaire was used. The questionnaire was mailed to 525 patients being treated for hypertension at a primary health care centre in Sweden at the beginning of 1991. The patients were asked to hand the questionnaire in next time they attended the health care centre. The patient population was randomly divided into 15 subsamples in which the monthly price for the treatment was varied between SEK 50 and SEK 1500. The patients were asked whether they would continue their current treatment at the specified price. Instead of using the usual yes/no response alternatives, the respondents were allowed to choose between five different responses: yes, definitely; yes, probably; no, probably not; no, definitely not; and don't know. The CV question that was used was worded in the following way:

This question concerns how you value your treatment for high blood pressure in economic terms. Assume that you would have to pay a greater share of the treatment cost than today, and that the fees for drugs and physician visits were raised.

Would you choose to continue your current treatment for high blood pressure if your fees for the treatment were SEK 500 per month?

YES, DEFINITELY
YES, PROBABLY
NO, PROBABLY NOT
NO, DEFINITELY NOT
DON'T KNOW

The bid was varied between SEK 50 and SEK 1500 in 15 subsamples. In the questionnaire, data were also collected about the perceived health with and without treatment by using a visual analog scale (VAS), where the respondents were asked to indicate their perceived health status with and without treatment between worst possible health state (0 cm) and best possible health state (15 cm). Data about the

following socio-economic background variables were also collected in the questionnaire: taxable household income, education, age, sex and household size.

Of the 525 mailed questionnaires 335 were returned, i.e. the response rate was about 64%. The item non-response rate was about 5% on the WTP question, i.e. 5% of the individuals who returned the questionnaire did not answer the CV question. About 16% of the answers to the WTP question were "don't know" answers.

Logistic regression analysis estimated by maximum likelihood methods was used to analyse the binary CV data. Table 2 shows the results of the logistic regression analysis, with yes definitely and yes probably both coded as yes, and no probably not and no definitely not both coded as no. The results are shown for the 210 individuals where data were available on all the explanatory variables (the bulk of the non-response on the explanatory variables was on the question about perceived change in health status). In one of the regressions in Table 2 only the bid variable was entered as the explanatory variable.

Table 2. Logistic regression coefficients, t-ratios within parentheses

Regressor variable	**Equation**	
	1	**2**
Intercept	1.97[a]	2.06[a]
	(7.49)	(7.44)
Bid	-.0023[a]	-.0024[a]
	(-5.48)	(-5.36)
ΔVAS		.12[b]
		(2.42)
Sex		.12
		(.34)
Age		-.021
		(-1.13)
Income		.000016
		(.75)
Education		.12
		(.44)
Household size		.086
		(.43)
n	210	210
Log-likelihood	-108.98	-104.05
Goodness of fit		
Individual prediction	76.67%	79.52%
Likelihood ratio index	.15	.18
WTP:		
Mean	857	858
Standard error	98	100

[a,b]: Significant at 1% and 5% levels.

The bid was highly significant and the mean WTP was estimated by dividing the intercept by the coefficient for bid as shown above (1.97/0.0023), which yielded a mean WTP of SEK 857. In the other regression in Table 2 all the other explanatory variables were added as well. In this regression the change in perceived health status was statistically significant on the 5% level, but none of the socioeconomic variables were statistically significant. In this case, the mean WTP was again estimated in the same way as above, since the added explanatory variables were measured as their deviations from the mean. The mean WTP according to the second regression was thus equal to SEK 858 (2.06/0.0024). A logistic regression analysis was also run for all the respondents who answered yes/no to the CV question with only bid as explanatory variable (260 respondents), yielding a mean WTP of about SEK 770 (not shown in the table).

The logistic regression analysis was also carried out on two subsets of the data: a sample that only included the responses of those who definitely were willing or not willing to pay the suggested sum of money ("certain" respondents), and a sample that included only the probable yes/no responses ("uncertain" respondents). The two subsamples were used to test if there was any behavioural difference between those delivering certain responses to the yes/no question and those delivering uncertain ones.

There was a great difference between the goodness of fit of the regression for the sample of certain respondents and that for the sample of uncertain respondents (a difference in the likelihood ratio index of 0.66 versus 0.07 and a difference in individual prediction of 94% versus 71% for the regression with all the explanatory variables). This indicates that the type of binary CV question used discriminated between the certain and uncertain respondents. The mean WTP was similar in both the regression with certain and the regression with uncertain respondents.

The mean WTP was also calculated with the non-parametric method developed by Kriström (1990). If the proportion of acceptance is not 0% at the highest bid level, the upper WTP has to be decided in order to calculate the mean WTP. In this study the proportion of acceptance was 24% at the highest bid level (SEK 1500). The mean was therefore calculated with SEK 1500, 2000 and 2500, respectively, as the highest WTP. The mean WTP according to this method was SEK 735, SEK 795, and SEK 855, respectively, using the three aforementioned assumptions about the highest WTP. This was similar to the mean according to the logistic regression analysis. For these calculations of the mean to be representative for all respondents, the mean WTP of the respondents giving "don't know" answers has to be the same as for those giving yes/no answers. A possible interpretation of the "don't know" answers given in the study is that respondents are indifferent to the proposed bid. Since in that case the mean WTP of these respondents would be about SEK 600 in the study, the mean WTP of all respondents would only change slightly.

One test of validity in a CV study is to assess whether the hypothesized theoretical relationships are supported by the data (Mitchell & Carson 1989). The estimated logistic equations in the study were in accordance with the theoretical predictions. The bid variable was highly significant, showing that an increased price reduces the demand for antihypertensive treatment. The perceived change in the expected health status (ΔVAS) was also significant, indicating as expected that an increased difference in health status between the treatment and no treatment cases increases the WTP.

The results of the study thus gave some support for the validity of the CV method using binary responses. On the other hand income was not statistically significant, as could have been expected. An interesting feature in the study is the comparison of the certain and uncertain respondents. The results of the regressions with these subsamples showed that the fit of the regression improved greatly in the sample with only certain responses, which is what could be expected if the uncertainty in the answers is reduced. One possible hypothesis could also be that only the individuals who answer yes, definitely in a CV study would actually pay in a situation with real money transactions. No estimates of the WTP based on such an interpretation were done in the study, but this would give a conservative estimate of WTP. However, in order to test this hypothesis it would be necessary to carry out comparisons between hypothetical and real WTP.

One problem in interpreting the WTP in the study is that the perceived risk reduction from hypertension treatment among the hypertension patients may differ from the objective risk reduction. Ideally the respondents should have been given information about the likely risk reduction from hypertension treatment or data should have been collected about the perceived risk reduction. The mean WTP amounts in the study should thus be interpreted cautiously.

6.7 Conclusions

CV studies generally measure the WTP rather than the WTS even for a deterioration in the health or the environmental quality. The reason for this is that the studies which have compared WTP and WTS for the same change have noted unexpectedly large differences between the amounts (Hammack & Brown 1974; Bishop & Heberlein 1979; Rowe et al 1980; Brookshire et al 1980; Knetsch & Sinden 1984). Even though it has been argued on the basis of the theory that we can expect large disparities in some situations (Hanemann 1991), it seems hard to explain the observed differences solely on the basis of differences in the marginal utility of income (if WTP and WTS measure the same change in utility, the difference between them is that they are converted to money using different marginal utilities of income).

This has led to a recommendation from most researchers in the area to use WTP rather than WTS, since WTS has been perceived as being less reliable (Mitchell & Carson 1989). One reason for preferring WTP questions in CV studies is that

consumers are more used to buying (i.e. expressing their WTP) than selling goods (i.e. expressing their WTS). The NOAA panel also recommended the use of WTP rather than WTS, since it gives a more conservative estimate.

In spite of the large number of CV studies that have been carried out, mainly in the environmental area (for example, Carson et al (1993) lists more than 1,400 studies related to the method), the validity of the method is still unclear. The number one research priority therefore seems to be to test the relationship between hypothetical and real money payments in experimental studies. It is important that such studies should also be carried out on health care goods, since it is possible that the validity of the method differs for different types of goods. It may also be easier to find cases where real payments can be compared with hypothetical payments in the health area than in the environmental area, due to the private goods nature of health care.

For current CV applications it seems to be important to demonstrate that the WTP increases with the size of the health change, e.g. the size of the risk reduction, and to achieve a high response rate. Whether the WTP increases with the size of the health care programme can be tested by randomly assigning the respondents to different subsamples with different sizes of the health change. If the WTP is assessed for a specific treatment it would also be possible to collect information about the perceived health gain for different respondents, in order to analyse whether the WTP increases with the perceived health gain. To achieve a high response rate it may be necessary to use some form of interview (i.e. face-to-face or telephone interviews) rather than mail questionnaires. If interviews are used it also seems easier to explain the nature of the market and the good that is sold to the respondent.

REFERENCES

Acton JP. Evaluating public programs to save lives: the case of heart attacks. Santa Monica: RAND Report R-950-RC, 1973.

Amemiya T. Qualitative response models: a survey. Journal of Economic Literature 1981;19:1483-1536.

Berwick DM, Weinstein MC. What do patients value? Willingness to pay for ultrasound in normal pregnancy. Medical Care 1985;23:881-893.

Bishop RC, Heberlein JA. Measuring values of extra market goods: are indirect measures biased? American Journal of Agricultural Economics 1979;61:926-30.

Bohm P. Estimating demand for public goods: an experiment. European Economic Review 1972;3:111-130.

Bowker JM, Stoll JR. Use of dichotomous choice nonmarket methods to value the whooping crane resource. American Journal of Agricultural Economics 1988;70:372-381.

Boyle KJ, Bishop RC. Welfare measurements using contingent valuation: a comparison of techniques. American Journal of Agricultural Economics 1988;70:21-28.

Boyle KJ, Bishop RC, Welsh MP. Starting point bias in contingent valuation bidding games. Land Economics 1985;61:188-194.

Boyle KJ, Welsh MP, Bishop RC. Validation of empirical measures of welfare change: comment. Land Economics 1988;64:94-98.

Brookshire DS, d'Arge RC, Schulze WD, Thayer MA. Experiments in valuing public goods. In: Smith VK (Ed.). Advances in applied microeconomics. Vol 1. Connecticut: JAI Press Inc, 1981.

Brookshire DS, Ives BC, Schulze WD. The valuation of aesthetic preferences. Journal of Environmental Economics and Management 1976;3:325-46.

Brookshire DS, Randall A, Stoll JR. Valuing increments and decrements of natural resource service flows. American Journal of Agricultural Economics 1980;62:478-488.

Cameron TA. A new paradigm for valuing non-market goods using referendum data: maximum likelihood estimation by censored logistic regression. Journal of Environmental Economics and Management 1988;13:255-268.

Carson RT. Constructed markets. In Braden JB, Kolstad CD (Eds.). Measuring the demand for environmental quality. Amsterdam: Elsevier/North Holland, 1991.

Carson RT, Wright J, Alberini A, Flores N. A bibliography of contingent valuation studies and papers. La Jolla CA.: Natural Resource Damage Assessment, 1993.

Cicchetti CJ, Smith VK. Congestion, quality deterioration, and optimal use: wilderness recreation in the Spanish Peaks primitive area. Social Science Research 1973;2:15-30.

Cronin FJ. Valuing nonmarket goods through contingent markets. Report to the U.S. Environmental Protection Agency, Richland, Washington D.C., 1982.

Cummings RG, Brookshire DS, Schulze WD (Eds.). Valuing environmental goods. New Jersey: Rowman and Allanheld, 1986.

Darling AH. Measuring benefits generated by urban water parks. Land Economics 1973;49:22-34.

Daubert JT, Young RA. Recreational demands for maintaining instream flows: a contingent valuation approach. American Journal of Agricultural Economics 1981;63:666-76.

Davis RK. Recreation planning as an economic problem. Natural Resources Journal 1963;3:239-249.

Diamond PA, Hausman JA. On contingent valuation measurement of nonuse values. In Hausman JA (Ed.). Contingent valuation: a critical assessment. New York: North-Holland, 1993.

Donaldson C. Willingness to pay for publicly provided goods: a possible measure of benefit? Journal of Health Economics 1990;9:103-118.

Duffield JW, Patterson DA. Field testing existence values: an instream flow trust fund for Montana rivers. Paper presented at the annual meeting of the American Economic Association, New Orleans, 1991.

Desvousges WH, Johnson FR, Dunford RW, Boyle KJ, Hudson SP, Wilson KN. Measuring natural resource damages with contingent valuation: tests of validity and reliability. In Hausman JA (Ed.). Contingent valuation: a critical assessment. New York: North-Holland, 1993.

Diamond PA, Hausman JA, Leonard GK, Denning MA. Does contingent valuation measure preferences? Experimental evidence. In Hausman JA (Ed.). Contingent valuation: a critical assessment. New York: North-Holland, 1993.

Evans WN, Viscusi WK. Estimation of state dependent utility functions using survey data. Review of Economics and Statistics 1991;73:94-104.

Fisher A, Chestnut LG, Violette DM. The value of reducing risks of death: a note on new evidence. Journal of Policy Analysis and Management 1989;8:88-100.

Greenley DA, Walsh AG, Young RA. Option value: empirical evidence from a case study of recreation and water quality. Quarterly Journal of Economics 1981;95:657-673.

Hammack B, Brown GM Jr. Waterfowl and wetlands: toward bioeconomic analysis. Baltimore: The Johns Hopkins University Press for Resources for the Future, 1974.

Hanemann MW. A Methodological and Empirical Study of the Recreation Benefits from Water Quality Improvement. Ph.D. Dissertation, Harvard University, 1978.

Hanemann MW. Welfare evaluations in contingent valuation experiments with discrete responses. American Journal of Agricultural Economics 1984;66:332-341.

Hanemann MW. Willingness-to-pay and willingness-to-accept: How much do they differ? American Economic Review 1991;81:635-647.

Hanemann MW, Loomis JB, Kanninen B. Statistical efficiency of double-bounded dichotomous choice contingent valuation. American Journal of Agricultural Economics 1991;73:1255-1263.

Johannesson M, Jönsson B, Borgquist L. Willingness to pay for antihypertensive therapy: results of a Swedish pilot study. Journal of Health Economics 1991;10:461-474.

Johannesson M, Johansson P-O, Kriström B, Gerdtham U-G. Willingness to pay for antihypertensive therapy: further results. Journal of Health Economics 1993;12:95-108.

Johansson P-O, Kriström B, Mäler KG. Welfare evaluations in contingent valuation experiments with discrete response data: comment. American Journal of Agricultural Economics 1989;71:1054- 1056.

Jones-Lee MW. The value of life: an economic analysis. London: Martin Robertson, 1976.

Jones-Lee MW, Hammerton M, Philips PR. The value of safety: Results of a national sample survey. Economic Journal 1985;95:49-72.

Kahnemann D. Comments. In Cummins RG, Brookshire DS, Schulze WD (Eds.). Valuing environmental goods. Totowa New Jersey: Rowman & Allanheld, 1986.

Kahnemann D, Knetsch JL. Valuing public goods: the purchase of moral satisfaction. Journal of Environmental Economics and Management 1992;22:57-70.

Knetsch JL, Sinden JA. Willingness to pay and compensation demanded: experimental evidence of an unexpected disparity in measures of value. Quarterly Journal of Economics 1984;98:507-521.

Kriström B. Discrete and continuous valuation questions; do they give different answers? Swedish University of Agricultural Sciences in Umeå, Department of Forest Economics, Working paper 90, 1989.

Kriström B. A non-parametric approach to the estimation of welfare measures in discrete response valuation studies. Land Economics 1990;66:135-139.

Milon JW. Contingent valuation experiments for strategic behavior. Journal of Environmental Economics and Management 1989;17:293-308.

Minkin S. Optimal designs for binary data. Journal of the American Statistical Association 1987;82:1098-1103.

Mitchell RC, Carson RT. Using surveys to value public goods: the contingent valuation method. Washington D.C.: Resources for the Future, 1989.

National Oceanic and Atmospheric Administration. Report of the NOAA panel on contingent valuation. Federal Register 1993;58:4602-4614.

National Oceanic and Atmospheric Administration. Natural resource damage assessments: proposed rules. Federal Register 1994;59:1062-1191.

Nyquist H. Optimal designs of discrete response experiments in contingent valuation studies. Review of Economics and Statistics 1992;74:559-563.

Randall A, Grunewald O, Pagoulatos A, Ausness R, Johnson S. Reclaiming coal surface mines in central Appalachia: a case study of the benefits and costs. Land Economics 1978;54:472-489.

Randall A, Ives BC, Eastman C. Bidding games for valuation of aesthetic environmental improvements. Journal of Environmental Economics and Management 1974;1:132-149.

Reardon G, Pathak DS. Assessment of a contingent valuation technique with utility estimation models. Journal of Research in Pharmaceutical Economics 1989;1:68-89.

Reardon G, Pathak DS. Segmenting the antihistamine market: an investigation of consumer preferences. Journal of Health Care Marketing 1990;10:23-33.

Ridker RG. Economic Costs of Air Pollution. New York: Praeger, 1967.

Rowe RD, d'Arge RC, Brookshire DS. An experiment on the economic value of visibility. Journal of Environmental Economics and Management 1980;7:1-19.

Scherr BA, Babb EM. Pricing public goods: an experiment with two proposed pricing systems. Public Choice 1975;23:35-48.

Schulze WD, Cummings RG, Brookshire DS, Thayer MA, Whitworth R, Rahmatian M. Methods development in measuring benefits of environmental improvements: experimental approaches to valuing environmental commodities. Vol 2. Draft manuscript of a report to the Office of Policy Analysis and Resource Management, U.S. Environmental Protection Agency, Washington D.C., 1983.

Seip K, Strand J. Willingness to pay for environmental goods in Norway: A contingent valuation study with real payment. Environmental and Resource Economics 1992;2:91-106.

Sellar CJ, Chavas JP, Stoll JR. Validation of empirical measures of welfare change: a comparison of nonmarket techniques. Land Economics 1985;61:156-175.

Sellar CJ, Chavas JP, Stoll JR. Specification of the logit model: the case of valuation of nonmarket goods. Journal of Environmental Economics and Management 1986;13:382-390.

Smith VK, Desvousges WH. An empirical analysis of the economic value of risk changes. Journal of Political Economy 1987;95:89-115.

Smith VL. The principle of unanimity and voluntary consent in social choice. Journal of Political Economy 1977;85:1125-1139.

Smith VL (Ed.). Research in Experimental Economics. Vol 1. Connecticut: JAI Press, 1979.

Sorg C, Brookshire DS. Valuing increments and decrements of wildlife resources: further evidence. Report to the Rocky Mountain Forest and Range Experiment Station, U.S. Forest Service, Fort Collins, Colorado, 1984.

Thayer MA. Contingent valuation techniques for assessing environmental impacts: further evidence. Journal of Environmental Economics and Management 1981;8:27-44.

Thompson MS. Willingness to pay and accept risks to cure chronic disease. American Journal of Public Health 1986;76:392-396.

Thompson MS, Read JL, Liang M. Feasibility of willingness to pay measurement in chronic arthritis. Medical Decision Making 1984;4:195-215.

Viscusi WK. The value of risks to life and health. Journal of Economic Literature 1993;31:1912-1946.

Viscusi WK, Magat WA, Huber J. An investigation of the rationality of consumer valuations of multiple health risks. Rand Journal of Economics 1987;18:465-479.

Viscusi WK, Magat WA, Huber J. Pricing environmental health risks: survey assessments of risk-risk and risk-dollar trade-offs for chronic bronchitis. Journal of Environmental Economics and Management 1991;21:32-51.

7. THE ESTIMATION OF COSTS

After estimating the benefits, the costs also have to be estimated. In Chapters 3 and 4 dealing with the definitions of costs and benefits, we used the term costs to reflect the resource consequences of a treatment. The resource consequences were defined as the effect on the consumption of goods and services of a health care programme and the effect on the consumption of leisure. The resource consequences were also divided into those that are paid by the recipients of a health care programme and those that are paid by other people in society. The costs that are carried by other people were defined as the external costs of a programme.

In principle the cost estimations that should be carried out in a cost-benefit analysis are the resource consequences that are not already included in the WTP of the individuals who receive the health care programme. What is included in the WTP of the individuals will depend on the institutional arrangements in society as shown in Chapter 4, and if the WTP is based on a CV study it will also depend on how the CV question is phrased.

Below we will focus on how to estimate the value of different resource consequences without considering the institutional arrangements in society and the way the CV question is posed. The reader should keep in mind, however, that the estimations which should be carried out in an individual cost-benefit analysis depend on what is included in the WTP of the programme. For the mortality costs, the estimation of the external costs is also the only issue considered, since this seems to be the only meaningful estimation of costs that can be carried out in the mortality case.

We will make a distinction below between programme costs, morbidity costs and mortality costs. Programme costs are defined as the inputs that go into a health care programme (i.e. increased costs), morbidity costs are defined as the resource consequences due to changes in morbidity (i.e. decreased costs if the morbidity decreases), and mortality costs are defined as the resource consequences due to changes in mortality. The distinction between these different types of cost is most straightforward for different types of preventive programmes, where the risks of different morbidity and mortality events are reduced, e.g. a programme that reduces the risk of heart attacks through lowering cholesterol.

For health programmes that lead to improved quality of life, the distinction may be less useful, e.g. a new treatment that improves the quality of life of arthritis patients. In such cases the distinction between programme costs and morbidity costs becomes blurred, e.g. the new arthritis treatment may reduce the use of some other treatments for arthritis; it then seems rather arbitrary whether this cost reduction is treated as a reduction in the programme costs or in the morbidity costs. The principles for the calculations of the programme and morbidity costs are, however, the same and thus the main focus below will be on the estimation of the programme costs.

In many studies a distinction is made between direct and indirect costs. Direct costs are then usually defined as those resources that are used directly in an activity and indirect costs are defined as those resources that do not arise due to an activity or a disease. Indirect costs are then usually used to depict production losses or production gains and sometimes also gains and losses in leisure time. The distinction between direct and indirect costs is, however, somewhat arbitrary, e.g. if a patient takes time off work to participate in a prevention programme this time is used directly in the treatment programme.

In studies of health care programmes, the distinction between direct and indirect costs can be useful from a practical standpoint, since indirect costs then usually refer to costs outside the health care sector. Below we will not use the direct versus indirect cost terminology, but we will discuss the estimation of gains and losses in working and leisure time for the recipients of a health care programme.

7.1 The Estimation Of Programme Costs

Following our model in Chapter 4, the programme costs can be divided into three different items. The first item is what was referred to as health inputs, which are all the resources that go into the treatment programme except for the patient's time. An increase in health inputs either reduces the non-health consumption of the recipients of the health care programme or reduces the total consumption of other individuals in society (i.e. for other individuals in society the decrease in consumption could be non-health goods, health inputs, and even leisure).

The second item is the market production of the individual who receives the health programme, and a decreased market production will either decrease the consumption of the recipient of the health care programme or decrease the consumption of other individuals in society. The third item is the leisure of the individual. Reduced leisure will reduce the consumption of leisure for the individual. We will consider the estimation of each item separately below. As an illustrative example we will use the costs of treating hypertension.

The first, and in most cases the most important part of the programme costs, are the health inputs. The estimation of costs can be divided into two steps. The first step is to estimate the quantity of inputs used and the second step is to estimate the unit costs (prices) of the inputs used. The costs are then estimated as the unit costs multiplied by the quantities.

The quantities of inputs can be defined in different ways, which will affect how the prices are estimated. A physician visit can for instance be a quantity used, and the unit cost per physician visit can be used in order to estimate the cost of physician visits in a programme. However, the physician visit can also be disaggregated into time for the

physician, time for administrative personnel etc, and the unit costs for these different components are then used.

In order to estimate the costs of health inputs, we should start by identifying the quantities of health inputs that are used in the health care programme. The quantities that we are interested in are the increase in health inputs compared to the comparison alternative. In principle, all increases in quantities of resources which can be attributed to the health care programme should be estimated.

The data on quantity can be collected from different types of source. For health care costs the data can be collected retrospectively by going through the patient records of a sample of patients, or they can be collected prospectively by following up a group of patients and recording the quantities of health care resources used. Some health care centres may also have data about the quantities of resources used in treating specific patient groups that can be used.

A health care programme may also lead to behavioural changes for an individual, which will have resource consequences. If a treatment is introduced, the individual may spend less time and resources on health production carried out by the individual himself. This could lead to a change in the quantity of health inputs consumed outside of health care. However, it would be very difficult to collect data about these changes unless individuals were followed prospectively and the quantities of all health inputs were registered. In practice it may also be difficult to determine whether some types of consumption should be registered as health inputs or non-health consumption. For instance, it is difficult to determine whether the purchase of a pair of training shoes should be defined as consumption of health inputs or as consumption of non-health goods.

We can use hypertension treatment to illustrate the collection of data about the quantities of health inputs. Assume that we want to estimate the cost of treatment of hypertension compared to no treatment of hypertension in a group of patients with mildly elevated blood pressure. In this case the quantities of health care inputs can be divided into the number of physician visits, the number of nurse visits, the number of laboratory tests of different kinds, and the dose of different types of antihypertensive drugs used.

In addition to these health care inputs, the number of trips to the health care centre using different transport modes should be included. Furthermore, some patients may use relatives to get to the health care centre and then the number of hours of leisure time and the number of hours of working time for the relatives should also be included in the quantities of health inputs.

Data on these quantities in the hypertension case can be estimated for instance through a retrospective analysis of the patient records (physician visits, nurse visits, laboratory tests, drugs) and a questionnaire to the patients (number of trips, hours of leisure and

working time for the relatives). Data can also be collected prospectively by following up a group of patients and recording all the quantities of health inputs used.

It is also possible that the hypertension treatment causes behavioural changes so that the amounts of some health inputs that are not health care are reduced (or increased) due to the treatment. An individual may for instance buy less healthy food. This should then in principle also be included, but would be very difficult to measure.

The second step in determining the cost of health inputs is to estimate the unit costs or prices of the different inputs. The cost should represent the value of the input in its best alternative use, and in a market this would be the minimum price required to use the input in its current use rather than in its alternative use. Sometimes it is possible to use market prices. However, sometimes there may not be any prices for some health inputs. In studies of health care, different types of information may be used as prices. If for instance the number of physician visits is the quantity, one possibility would be to use the charge or fee that physicians use for private visits or as a basis for reimbursement by the state or an insurance company. Another possibility is to use the cost per physician as estimated by some health care centre or hospital, based on accounting principles. A third possibility is to carry out an estimation of the cost per physician visit based on the quantities and prices of inputs that are used to produce a physician visit.

It may be problematic to use charges, since the health care market is not a competitive one and the charges may be a poor reflection of the costs (Finkler 1982). It is also possible that the charges differ between payers, e.g. a large insurance company pays a smaller charge than private patients. In a system with uninsured patients the charges may also be set so high that they compensate for losses on uninsured patients. Thus charges are not in general a good approximation of the underlying costs.

The second alternative was to use some estimate of the accounting cost of a physician visit. For instasnce, a health care centre or a hospital may have estimated the cost per physician visit as a basis for the charges they use. One problem with such estimations is that since health care centres and hospitals are often not profit maximizing they may have no incentives to estimate the true cost of a physician visit. It is also likely to reflect an average cost for all the physician visits and this average cost may not be representative for the patient population in the study. The principles for allocation of overhead costs may also be unclear.

The third approach is to try and carry out an estimation of the cost per physician visit. In a sense this means attempting to make a better estimation than the accounting cost above, which may be more relevant for the patient population studied. This resembles a whole cost estimation, since now different quantities of inputs and the prices of these inputs have to be determined. The inputs that go into the physician visit, for instance, are the physician's time, and for this input the salary cost per time unit could be used. However, there are also a lot of costs that may be difficult to attribute directly to the

physician visit, such as the cost of facilities, administration etc. Costs like this which cannot be attributed directly to a physician visit (or some other unit being used) are often referred to as overhead costs. It would therefore be necessary to determine some principle for allocating overhead costs to the physician visit. For example, overhead costs could be attributed to visits according to the time the physician spends on a visit. For more on the different principles for allocating overhead costs see Drummond et al (1987).

One relevant issue in the case of overhead costs is whether they should be included or not in principle. The important thing for an economic evaluation is the change in costs due to the programme. It could therefore be argued that the overhead costs are not affected by whether more physician visits are carried out, i.e. the cost of the facilities and the administration will be the same irrespective of whether some extra visits are carried out or not. Some of the overhead costs may also be fixed, such as the cost of the facilities, but this is not necessarily the case for all the overhead costs, e.g. the amount of administrative personnell at the health care centre.

The overhead costs which are not fixed represent increased costs due to the programme, since these resources would have been used for something else in the absence of the programme (unless there is excess capacity at the health care centre so that these resources have no opportunity cost). Even if some of the overhead costs are fixed, it does not necessarily follow that they should not be included in the estimation of the costs. An example of a fixed cost may be the cost of the facilities of the health care centre. However, if the health care centre is operating at full capacity, adding more physician visits may lead to some other visits not being carried out due to space limitations. The facilities thus have an opportunity cost that should be included among the costs.

Some costs may be fixed in the short run but not in the long run; for example, when a health care centre is rebuilt the size will depend on the current or predicted capacity at the health care centre. It may thus be argued that the fixed costs should be included in the estimation of the cost per visit. In our opinion the treatment of fixed costs differs according to the situation. For instance, one case where it seems appropriate not to include the cost of the facilities is if a treatment programme utilizes the facilities at a time when they are not otherwise in use and the opportunity cost is zero, e.g. during the evenings if the facilities are not normally utilized at this time.

As was noted in Chapter 2, there are a number of market failures which may cause the market price to deviate from the marginal cost, and these are of interest when input prices are used to estimate costs. Perhaps the most well-known cause of market failure is monopoly. If there is a monopoly, price is likely to exceed marginal cost (although it is not totally obvious, see the discussion about contestable markets in Chapter 2). This means that in the case of a monopoly the price used in a cost-benefit analysis should be adjusted downwards so as to reflect the marginal cost.

The part of the price that exceeds the marginal cost for a monopoly can be denoted a transfer payment. Transfer payments are transfers of money from one group of individuals to another. Transfer payments can be neglected in a cost-benefit analysis, since the costs and benefits of the transfer payments in a sense cancel out. In the monopoly case the part of the price that is above the marginal cost represents a transfer of money from the consumer of the monopoly good to the owners of the monopoly firm (e.g. the shareholders). If the transfer is included on both the cost and the benefit side the correct result will be achieved in a cost-benefit analysis, since the costs and benefits net out.

This means that if in the monopoly case we use the monopoly price in the analysis and then also add the profit for the owners of the monopoly, we will get the same result as if we used the marginal cost of the good. However, in cost-benefit analysis the standard method is to try and adjust the prices so that they reflect the marginal cost. Note, however, that if we wanted to discuss not only the net benefits of a project, but also the distribution of benefits and costs, the transfer payments should be included on both the cost and benefit sides; for example, in the monopoly case the transfer from the consumer to the owners of the firm may reflect a transfer from poor people to rich people and this could be of interest in making a decision about a project. The same holds true if we want to weight benefits and costs differently in different groups (see the discussion about social welfare functions in Chapter 2).

The monopoly case is very relevant for the cost of drugs. New drugs (i.e. new chemical entities) are protected by patent for a number of years. This means that the producer of the good has a form of monopoly until the patent expires. Since the development of new drugs is costly, the patent system is a way of enabling the drug companies to receive a price above the marginal production cost so as to finance research and development of new drugs. The research and development costs of the patented drug are sunk costs in the sense that these resources have already been used and the size of these sunk costs should not affect the decision on whether or not to use the drug.

Note that this case is somewhat different to the discussion above of fixed costs for physician visits. The fixed costs above were fixed costs in the production of the physician visits. This distinction is important for two reasons. Firstly, the fixed costs such as the facilities may have an opportunity cost. Secondly, in the long run the addition of more physician visits will affect the size of the fixed costs, i.e. the fixed costs are not fixed in the long run. The research and development case is different since these resources have been used without producing something that may have an alternative use, and the size of the research and development costs for a drug is not affected by the size of the production of the drug in the future.

In principle therefore, the prices of drugs should be adjusted so as to reflect the marginal cost of producing and administering the drug. The alternative would be to use the price of the drug and also add the excess of price over marginal cost as a benefit to the owners of the drug company. If the excess of price over marginal cost is

used to finance research and development in new drugs, then in a sense the benefit that should be added is the expected consumer surplus of investment in research and development of new drugs (i.e. the difference between the WTP for the new drug and the production costs of the new drug; if this "consumer surplus" exceeds the investments in research and development then the investment is socially profitable). One argument for using the monopoly price of new drugs without adding any benefit for the investment in new drugs could then be that it is assumed that new drugs will not produce any consumer surplus. There seems to be no basis for such an assumption, however.

Another argument that can be put forward is that for drugs which are imported into a country, the price of that good is the relevant opportunity cost for the country, i.e. the country has to pay the import through increased exports to the same value. In that case the benefit of the price that is above marginal cost goes to another country, and it can be argued that this should be neglected if the cost-benefit analysis is restricted to people in the original country.

Another case where the monopoly argument may be of relevance is the market for the supply of physicians. Since entry to the physician market is restricted, i.e. the number of places at medical school is limited, the market for physicians is not a competitive one, and the restrictions on entry into the market may lead to the price of physicians exceeding the marginal cost.

Other cases where the price may exceed the marginal cost are if there are positive or negative externalities in the production of a good. An example of such an externality is pollution due to the production of a good. It is harder to find any obvious cases where the externalities in production would be relevant for the cost of health inputs than it is for the monopoly argument.

Another cause of market failure is decreasing marginal costs in production. If the marginal cost is decreasing in production, no firms will operate in the market if the price is set equal to the marginal cost, since this would mean making a loss. This case seems to be relevant for health care programmes. The number of in-patient days at a hospital is an obvious example where the marginal cost can be assumed to be decreasing with the number of days. The average cost per in-patient day may thus be a poor measure of the cost per in-patient day if the number of in-patient days differs from the average. There may be decreasing marginal costs in many other areas also, such as the number of operations performed etc, since the more operations that are performed, the better and more efficient may the surgeons become in performing the operations. The decreasing marginal cost argument depends on the existence of economies of scale in some way (i.e. that less inputs are needed for every additional unit, e.g. a operation) and the fact that a production volume has not been reached, so that decreasing returns to scale are experienced. In many cases the marginal cost may also be approximately constant (i.e. constant returns to scale).

A related issue touched upon above concerns the so-called learning economies, which means that the cost of a new technology decreases over time when people learn to use it in the appropriate way (the example with operations above is a form of learning economy, since the surgeons become more experienced with the number of operations carried out and therefore the cost per operation decreases). Learning economies may be of importance for many types of health care. However, it is very difficult to take into account these types of changes in costs over time due to "learning economies" in an economic evaluation, since the cost changes are hard to predict. Economic evaluations can also be carried out on different phases of the use of a new technology so as to incorporate the cost changes over time. However, if the economic evaluation is used as a basis for deciding whether or not a new technology should be introduced, this is not possible since if the costs exceed the benefits the technology will not be implemented, and thus the cost changes cannot be observed. More studies would therefore be of interest to determine how the cost of different types of new technologies changes with time, in order to enable predictions to be made about the impact of "learning economies" on the cost of a new technology.

In a sense, these types of cost reductions are an additional benefit of using a new technology, since it may reduce the cost so that future users become economically motivated to use the technology, or the new technology may become less costly for future users who would otherwise have used it at the current cost level. Thus this is a kind of investment effect which in principle should be added to the analysis. Part of the investment effect will, however, be offset by discounting (see the next chapter about the discounting of costs and benefits).

The next type of health input that may go into a health care programme is the labour production of the recipients of a health care programme. This item is relevant for instance in health care programmes which involve patients' time; this would apply to most or even all health care programmes. The patients' time may be working time, in which case there will be a loss in labour production. This quantity can be measured as the increased number of hours of working time devoted to health production due to the health care programme. The price that should be used is then the market value of the marginal output per hour that the worker produces.

Note that the reduced working time due to a health care programme increases the time spent on health production, and does not affect the consumption of leisure. The use of the market value of the production per hour is based on the assumption that the worker is indifferent between working time and health production time per se (i.e. indifferent as to whether the salary of the working time or the health effects of the health production time are ignored). If this is not the case, the difference in the monetary value in time spent working and time spent on health production should also be added to the price per hour (if for instance the job is risky, the individual may prefer the health production time, see for example the section on wage-risk studies in Chapter 5).

To estimate the market value of the production, the salary cost of the individual worker plus any taxes on the output produced, such as sales taxes, can be used (see also the section about taxes in Chapter 8). This represents the value which the consumers put on the output if a competitive labour market is assumed, so that the salary cost for a firm reflects the value of the production for the firm's margins. To return to our example about the costs of hypertension treatment, the quantity of lost working time is the amount of working time lost because of visits to the health care centre. This could be measured through a questionnaire issued to a group of patients.

Finally, if the health care programme involves an increase in the amount of health production time, this may be lost leisure rather than lost working time. The quantities of lost leisure can be measured as the number of hours lost because of the health care programme. It may not be very easy to set a price on the lost leisure. In this case, the price of the lost leisure should be equal to the difference in value between spending time on leisure and on health production per se, i.e. the compensation needed for the individual to swap an hour of leisure time for an hour of health production time, neglecting the health effects of the health production time. This price is not registered on any market. One market for leisure time is the job market. In a perfectly competitive market the individual will work until his wage (i.e. the after-tax wage rate) equals the difference in value between working time and leisure time (i.e. equals the value of leisure time if the value of working time per se is set at zero).

In the simple theoretical model in Chapter 4, no distinction was made between working time and health production time (i.e. they were assumed to have no utility or disutility in themselves), and according to that model the after-tax wage rate for the individual could then be used as the price of the leisure time which is given up for health production. In a sense this entails making the same assumption as if the value of the production had been used as the price of working time compared to health production time. However, in reality it seems probable that the values of time for health production and working per se might well differ, and that it would depend on the type of work and the type of health production activity. Thus the after-tax wage rate for the individual may not be a good measure for the price of leisure time that is given up for health production.

Another reason why the after-tax wage rate may not be a good approximation is also that there may be restrictions on how many hours an individual is allowed to work, so that the individual may not be allowed to adjust his working time until the wage rate equals the opportunity cost of leisure time. It is, however, possible to try and estimate the opportunity cost of leisure time in empirical studies by analysing the choices which individuals make with respect to time. Such studies have been most prevalent in the transport field, where they have been used to estimate the cost of travel time. These estimates are also of interest for health care programmes, since some of the leisure lost due to a health programme may be travel time (i.e. the travel time to get to a health care centre for a physician visit). Using an estimate of the cost of travel time could then be a starting point for valuing lost leisure time due to an increase in health

production time. Sensitivity analysis could then also be used, including for instance the after-tax wage rate.

In our example above about the costs of hypertension treatment, the quantity of lost leisure time is the amount of leisure time lost because of visits to the health care centre. This could be measured through a questionnaire to a group of patients. In addition, the treatment could also lead to other losses in leisure time, for purchasing and taking the drugs etc, but these quantities may be more difficult to measure. It may be possiple to capture these losses in leisure time also, using a survey to the patients. It is also possible that, due to the treatment, individuals will alter their behaviour and reduce (or increase) the amount of time devoted to health production, for instance by taking exercise (i.e. the pharmacological hypertension treatment and health production measures undertaken by the individual may be substitutes in the production of health). Such time changes should then also be included in principle, but they may be very difficult to measure.

In order to illustrate the estimation of programme costs, we will summarise below a study which attempted to estimate the programme costs of antihypertensive drug treatment for a patient population in Sweden (Johannesson et al 1991). The purpose of this study was to calculate the treatment cost per patient in a primary care setting. In this study, the treatment costs were divided into drug cost, consultation cost (including the cost of laboratory tests), travel cost and time cost for the patient.

To calculate the costs of antihypertensive drug therapy, an empirical investigation was carried out at a primary health care centre. The study population consisted of all patients on the hypertension register at a primary health care centre in Sweden. The register consisted of 632 patients (all patients who were undergoing pharmacological therapy for hypertension at the health care centre), of whom 22 were excluded because their medical records were not available at the time of the investigation.

The calculation of drug cost was based on the latest prescribed drug and dosage according to the medical record and the official drug retail prices in Sweden. The consultation cost was calculated as the cost per visit for different types of visit, to the physician, the hypertension nurse or the diabetes nurse (the patients with both hypertension and diabetes went to the diabetes nurse rather than the hypertension nurse), multiplied by the number of visits for each patient. To calculate the cost of different types of visit (i.e. the price per visit), the total cost of the health care centre was divided into directly attributable costs and overhead costs. The directly attributable costs are those costs which can be directly attributed to different types of visit, e.g. physicians' wages. Overhead costs are those costs which cannot be directly attributed to visits: administration, facilities, laboratory equipment, etc.

To obtain the directly attributable cost per visit for each type, the total wage cost in 1988 for the type of personnel receiving the visits (e.g. physicians) was divided by the number of visits that year. Overhead costs were then allocated to different types of

visit, based on the directly attributable cost per visit. The cost of laboratory tests for hypertension was included in overhead costs since it gave similar results as when calculated separately. Visits for each patient were obtained by examining their medical records.

When a patient who was being treated for hypertension had another diagnosis as well (245 patients), it was assumed that half of the consultation, travel and time costs were attributable to hypertension. It was also assumed that a physician visit for a hypertension patient corresponds to an average physician visit.

A questionnaire was used to calculate the time and travel cost of the visits to the health centre. The time cost consisted of costs of lost working time and costs of lost leisure time. The average wage cost per hour for an industrial worker was used as a proxy for lost working time. To value lost leisure time, 35% of the gross wage rate (i.e. the pre-tax wage rate) was used as a proxy, based on estimates of the cost of travel time used for cost-benefit analyses of road investments in Sweden. The travel cost was defined as the transport cost and the time cost of relatives visiting the health care centre with the patient. Time costs for relatives were valued in the same way as lost working time, and for valuing the travel cost, the costs of the different transport modes used by the patients were estimated.

On average a hypertension patient at the primary health centre made 0.59 physician visits per year and 2.20 nurse visits per year. The cost (price) of a physician visit was estimated to be SEK 541, a visit to the hypertension nurse was estimated to cost SEK 190, and a visit to the diabetes nurse was estimated to cost SEK 335 (SEK=Swedish Crowns; 1988 prices). The result in terms of annual treatment costs of hypertension in the study is shown in Table 1. The annual treatment cost is shown per patient in different age groups and for men and women, and is divided into drug cost, consultation cost, and travel and time cost.

Table 1. Treatment cost per patient in different age groups

Age	Drugs		Consultations		Travel and time		Total cost	
	M	W	M	W	M	W	M	W
30-39	1116	948	878	647	396	219	2390	1814
40-49	1553	1037	586	628	369	425	2508	2090
50-59	1590	1071	665	596	404	226	2659	1893
60-69	1397	1207	553	610	171	178	2121	1995
70-79	1469	987	725	607	198	204	2392	1798
80-	987	876	657	537	265	345	1909	1758
All	1420	1067	639	604	251	213	2310	1884

The consultation cost per patient was about SEK 600 for both men and women and it did not vary much between different age groups. The drug cost accounted for the major proportion of the treatment cost, on average about SEK 1200 per patient per year. There was a large difference between the drug costs for men and women. Men had a drug cost that was approximately SEK 350 higher than for women. This difference was present in all age groups and was due to two factors. The first is that a larger proportion of the female patients were treated with cheaper drug alternatives. The second is that the cost for all drug alternatives, except ACE inhibitors, was lower for women. The part of the difference (SEK 353) that depended on dose was SEK 87. The major part (SEK 266) of the difference was therefore due to the fact that on average women were treated with cheaper drug alternatives. The drug cost was lowest in the oldest and the youngest age groups, i.e. 80 years and older and 30-39 years. In the age interval 40-79 years the drug cost was relatively constant. The travel and time cost was on average about SEK 230 per patient per year. The difference between men and women was small. The travel and time cost was higher in the younger age groups, due to higher production losses in these age groups. In the oldest age group (80+) the travel and time cost increased again, due to higher travel costs (i.e. more of these patients needed help from a relative or friend to get to the health care centre for the visits).

This study highlights a number of important issues in the calculation of programme costs. One issue is the problem of joint costs, i.e. when patients are treated for more than one disease or health problem at the same time and the treatment costs for the different diseases cannot be separated. In this case a number of the hypertension patients were also treated for other diseases at the health care centre during the same visits. For instance, a large number of patients were treated for both diabetes and hypertension. In such a case it is difficult to know how large a part of the consultation and travel and time cost is due to hypertension and how much is due to diabetes. The correct cost of the hypertension programme is the additional costs that are due to the addition of the hypertension programme to the costs of treating the other diseases. It is, however, very difficult to know how large these extra costs are.

In this study it was assumed that, for patients treated for more than one disease during the same consultations at the health care centre, half of the consultation cost and the travel and time costs were due to hypertension. This assumption was somewhat arbitrary, however. In the extreme case it could be argued that the additional consultation and travel and time costs were zero for this group, since these costs would have been incurred anyway. On the other hand it seems likely that if hypertension is added, both the number of visits and the length of the visits may increase. To estimate the extra costs a study would be needed where the change in the number of visits and the length of each visit was measured, when for instance a hypertension programme was added for patients already being treated for diabetes. In the above study no such comparison was possible.

Another important issue is the allocation of overhead costs. In the above estimation, the overhead costs were allocated to different types of visit on the basis of the wage cost per visit for the personnel receiving the visits (e.g. the wage cost for the physician time involved in a physician visit or the wage cost for the nurse time involved in a nurse visit). This allocation of overheads assumes that the overhead costs are proportional to the wage cost directly attributable to a visit. This is not obvious either, and in principle it would have been possible to carry out a more detailed study to try and attribute more costs directly to the different types of visit. This could have been done for instance by measuring the amount of time used by other types of personnel in connection with the different visits, e.g. the time used by supporting personnel or administrative staff for each visit. However, studies of this type would require a significant work effort and it is not always obvious that the improved precision in the estimates will be worth the extra resources involved.

The estimation of the cost of lost leisure time in the study is also uncertain. 35% of the gross wage rate was used as the cost of lost leisure time, based on studies of the cost of transport time. It is of course possible that the cost of travel time differs from the cost of time devoted to visits at a health care centre. However, since some of the time lost because of visits to the health care centre is travel time, it may not be unreasonable to use the cost of travel time. The cost of the lost leisure time is also a relatively small fraction of the total treatment cost and different assumptions will therefore have a small impact on the total treatment cost.

A third issue of interest when using these treatment costs in an economic evaluation is that care has to be taken if the treatment costs are extrapolated to other patient groups outside the health care centre studied. It is possible that the patients treated for hypertension in the health care centre studied are not representative of the patients treated for hypertension in another setting. It is also possible that the care of the hypertensives in the health care centre is not representative of the care in another setting.

Further, the study did not take into account any changes in costs due to possible behavioural changes as a result of the treatment. Individuals receiving treatment may for instance devote less time and money to health production, e.g. through diet and exercise. It would however be extremely difficult to estimate these cost changes.

7.2 The Estimation Of Morbidity Costs

The principles for the estimation of morbidity costs are the same as for the estimation of programme costs. The morbidity costs can also be divided into the three components: the change in the consumption of health inputs, the change in market production, and the change in the consumption of leisure. As was mentioned above, in some cases it may not be important to differentiate between morbidity and programme costs. Such a case could be a new treatment for arthritis, where the resource

consequences can be measured directly as the change in the consumption of health inputs, the change in market production, and the change in the consumption of leisure. However, in general the programme costs would represent increases in costs and the morbidity costs would represent decreases in costs. The easiest way to think about the morbidity costs is for programmes that reduce the risk of some morbidity event, e.g. a programme that reduces the risk of non-fatal fractures among women with low bone mass.

The first item to be considered is the change in the consumption of health inputs due to a health programme. If a morbidity case is avoided then the expected consumption of health inputs due to morbidity will decrease. The avoided costs of health inputs can then be estimated as the difference in the consumption of health inputs in the morbidity health state and the consumption of health inputs if the morbidity event did not occur. This gives the reduced consumption of health inputs for avoiding one morbidity case.

We can take a programme that reduces the risk of fractures among women with low bone mass as an example, and assume here for the sake of simplicitly that the programme does not affect the risk of fatal fractures. The programme is assumed to reduce the risk of two different types of fracture. The first type of fracture only leads to a temporary health impairment and after one year the individual returns to full health. The second type of fracture leads to a chronic health impairment. The two different types of fracture are referred to below as minor and severe fractures, respectively. The survival is assumed not to be affected by the fractures.

To estimate the reduced consumption of health inputs for avoiding a minor fracture, we can compare the consumption of health inputs for a group of individuals with minor fractures for the year after the injury with the consumption of health inputs for a group of identical individuals without a minor fracture. Assume that the group of individuals with the minor fracture consumes health inputs for $10,000 during the year after the fracture, and the group of individuals without a minor fracture consumes health inputs for $5,000 during one year. The reduced consumption of health inputs for avoiding one minor fracture is then $5,000.

The reduced consumption of health inputs for avoiding a severe fracture can be estimated in a similar way, by comparing the consumption of health inputs for a group of individuals with a severe fracture to the consumption of health inputs for an identical group of individuals without a severe fracture initially (of course, they may still suffer a severe fracture in the future). In this case, however, the consumption of health inputs has to be compared for the whole life-time, since the severe fracture can be assumed to be associated with an increased consumption of health inputs for the rest of the life. Assume that the life-time consumption of health inputs is $100,000 for the group of individuals with a severe fracture and the life-time consumption of health inputs is $50,000 for the group of individuals initially free from the severe fracture (discounting of costs is ignored here for the sake of simplicity; see Chapter 8 for

discounting of monetary costs and benefits). The reduced cost of avoiding one severe fracture is then $50,000.

The measurement issues for the morbidity case are the same as for the measurement of programme costs, and the quantities of different health inputs have to be multiplied by the respective unit costs (prices) to get the costs. The issues concerning the unit costs used are also the same as for the case with programme costs.

The difficulty in practice seems to be the estimation of the extra quantity of health inputs in the morbidity states compared to the health state if the morbidity state is prevented. For instance, if a group of individuals is followed up after a severe fracture to estimate the consumption of health inputs, it is difficult to know what the costs would have been if these persons had not sustained the fracture. The quantity of health inputs could be compared with the quantity of health inputs consumed by the same individuals before the fracture. The problem with this is that even without the fracture, the consumption of health inputs could increase, for instance as a result of increasing age. An identical comparison group is therefore needed who are initially free from the fracture, but it is difficult to get such a group in practice. It is also difficult to obtain data on the costs in the long run in a morbidity state, and therefore the costs are often extrapolated to future years. For instance, in a chronic health state a measurement of the difference in costs may be carried out for one year, and this difference is then assumed to persist for future years also.

In the medical field, data on the consumption of health inputs for different treatment alternatives can sometimes be collected in connection with clinical trials, which could lead to high quality data on the consumption of health inputs due to morbidity. However, if the clinical trial concerns the prevention of events that can be assumed to lead to increased costs for the rest of the lifetime (e.g. heart attacks), data from clinical trials will have to be supplemented, since they do not have a follow-up for the rest of the lifetime.

In practice most estimates of reduced morbidity costs are based on follow-up studies after the disease events. There is a risk that such studies will exaggerate the savings due to prevention of different events, since the state that the individual will be in if the event is avoided is also associated with some consumption of health inputs. In most cases, the most important difference in the consumption of health inputs between health states is the consumption of health care, but other resources such as travelling, the time of relatives, and the use of social services can also differ.

The next item is market production. If a morbidity event is avoided, this is often associated with increased production. Data on this can be collected in the same way as for the data on the consumption of health inputs, and the same issues will apply. To illustrate the estimation of increased production, we can use the same example as above for the prevention of non-fatal fractures. To estimate the increased production due to avoiding a minor fracture, we can compare the production for a group of

individuals with minor fractures for the year after the injury with the production for a group of identical individuals without a minor fracture. Assume that the group of individuals with the minor fracture has a market production of $30,000 in the year after the injury, and the group of individuals without a minor fracture has a market production of $32,000. The increased production for avoiding one minor fracture is then $2,000.

The increased production due to avoiding a severe fracture can be estimated in a similar way, by comparing the production for a group of individuals with a severe fracture to the production for an identical group of individuals without a severe fracture initially. In this case the production has to be compared for the whole lifetime, since the severe fracture can be assumed to be associated with a decreased production for the rest of the life. Assume that the lifetime production is $600,000 for the group of individuals with a severe fracture and $1,000,000 for the group of individuals initially free from the severe fracture (discounting of costs is ignored here for the sake of simplicity; see Chapter 8 for the discounting of monetary costs and benefits). The increased production due to avoiding one severe fracture is then $400,000.

One important difference compared to the consumption of health inputs is that both the quantity of working time and the price per hour can differ between different health states. As in the case of programme costs, the market value of the production per hour should be used as the price per hour. To estimate the market value of the production, the salary cost of the individual worker plus any taxes on the output produced such as sales taxes can be used.

The final resource consequence due to a change in morbidity is the effect on the consumption of leisure. If the number of hours worked differs between the disease state and the state without the disease, this means that preventing the event reduces the consumption of leisure for the individual. This would then be a loss in leisure due to the prevention of the disease event. This item will not be insured so it will always be part of the consequences that falls on the individual who receives the health programme. It is thus only relevant to estimate this item if the WTP for the health change is estimated holding the leisure and income constant, i.e. using the WTP for the "pure" health change.

The quantity of lost leisure time due to increased working time can easily be estimated if the increased working time is estimated, since the lost leisure then equals the increased working time. However, valuing the lost leisure time is more difficult. How the lost leisure time should be valued depends on how the WTP for the pure health change is estimated (see Chapters 3 and 4 for a discussion about adding together different WTP amounts). If the WTP for the pure health change (or a change in the risk of a health state) is estimated, this means that the WTP has to be estimated by holding the consumption of leisure constant, both in the health state that is prevented and in full health (assuming that the individual will be in full health if the event is prevented). This means that the leisure time could be held constant either at the level

in the prevented health state or at the level in full health, and this has implications for choosing the health state for which the value of leisure should be estimated.

If the leisure is held constant at the level in the prevented health state, then the value of leisure time should be based on the difference between the values of leisure time and working time in full health (and it should in principle be based on the income level when the individual has paid the WTP for the health change; see Chapter 4). If on the other hand the WTP is based on the leisure level in full health, then the value of leisure time should be based on the difference between the values of leisure time and working time in the prevented health state (and the WTP for the health change should in principle be based on the income level when the individual has paid the WTP for the change in leisure; see Chapter 4). The two WTP amounts will not be identical if the difference between the utility of working and the utility of leisure differs between the health states.

It is possible for the value of leisure time to be different in different health states. The value of the lost leisure time that is used for working is equal to the minimum wage that the individual needs in order to work during these hours rather than enjoying leisure. As discussed above, one possible approximation for the value of leisure time is the after-tax wage rate, but this can differ between the disease state with and without the event. This is also based on the assumption that the individual works the optimal number of hours, and it may also exaggerate the value of leisure time for non-marginal changes. It may also be impossible to define the after- tax wage rate in the disease state with the event, since working may not be a possibility and the value of leisure time may be low.

The most straightforward way to try and incorporate the loss in leisure time, if it is estimated separately, may be to carry out the measurement of the WTP for the pure health change, so that the value of leisure time should be based on the healthy state. It may thus be possible to argue that the loss in leisure time as valued by the after-tax wage rate should be added in as a loss of the change in morbidity (or the difference between the values of leisure time and working time per se if the individual is not able to work the optimal number of hours). Alternatively it may be argued that the WTP should be estimated so that the value of leisure time should be based on the poor health state, and that the change in leisure time can then be ignored, since it can be assumed to be low in the poor health state. It is not recommended to try and separate the change in leisure from the change in pure health, however. It seems better to try and incorporate the change in leisure into the WTP for the health change.

If the changes in the individual's consumption of leisure and non-health goods are used to provide a lower bound of the WTP for the change in morbidity, the issue is a little different. In this case it is enough to know that the individual prefers to be in full health and working, rather than to be in the health state with the event and having leisure, to decide that this estimate is a lower bound of the private WTP, i.e. the time per se working in full health should be valued higher than the leisure time in the

health state with the event. This could be tested by asking a number of people in the health state of interest.

However, the increase in the number of hours worked is not the only change in leisure due to a morbidity change. The number of hours for health production may change as well. It seems likely that the number of hours for health production will decrease if the consumption of health inputs decreases (e.g. if the number of health care visits decreases, the time for these visits will decrease). This change in leisure may be extremely difficult to quantify, but the measurement of its value involves the same issues as above with one exception: the difference in time values would now be between leisure and health production time. This change would also, at least in part, offset the change in leisure due to the increase in working time.

Before moving on to an application of the estimations of the change in morbidity costs due to a health programme, it can be useful to say something about the estimation of external costs and their relation to the estimations detailed above. The external costs will be equal to the part of the decreased consumption of health inputs that is not received by the individual, plus the part of the increased production that is not received by the individual. It is assumed here that the private life-time savings of the individual or his household are not affected by the change in morbidity; otherwise this could complicate matters (see below in the section on mortality). The change in the consumption of leisure does not affect the external costs, since the change in leisure is not insured.

As an example here, we will use a study that estimated the treatment and morbidity costs of two alternative treatment options after an acute myocardial infarction. The aim of the study by Olsson et al (1987) was to analyse the economic implications of adding beta-blockers to the current treatment after an acute myocardial infarction, for patients below the age of 70. The alternatives that were compared were thus traditional treatment versus traditional treatment with the addition of a beta-blocker.

The study was based on data from the Stockholm metoprolol study, a randomised clinical trial of metoprolol versus placebo after acute myocardial infarction (Olsson et al 1985). 301 patients aged below 70 years were included in the trial; this was 66% of all the patients under the age of 70 years who survived a myocardial infarction in the population covered by the hospital during the time period included in the trial. The patients were randomised to a double blind treatment with 100 mg of either metoprolol or placebo twice daily. The treatment was initiated one to two weeks after the onset of the myocardial infarction. The patients were seen by a physician after 0, 6, 12, 18, 24, 30 and 36 months and by a specially trained nurse after 9, 15, 21, 27 and 33 months. All the patients were followed up for three years.

Mortality and morbidity were analysed according to the intention-to-treat principle, i.e. according to the initial randomisation irrespective of whether or not they continued the randomised treatment. A reduction in both mortality and morbidity with

metoprolol was observed in the study, as well as an improved quality of life (Olsson et al 1985,1986).

On the basis of the clinical trial, a cost analysis was carried out by Olsson et al (1987). The aim of the cost analysis was to see if the treatment with beta-blockers led to increased or decreased costs. If the costs decreased it would be obvious that the treatment was beneficial from an economic viewpoint, since the treatment improved the health of the individuals. Both health care costs and the change in market productivity were included in the study. Only the change in productivity due to the change in morbidity was included.

The health care costs were divided into the cost of drugs, the cost of out-patient care and the cost of in-patient care. The productivity was measured as the loss of production due to morbidity. All the costs were discounted with a 5% discount rate (for discounting see Chapter 8) and estimated in 1985 prices. Only the costs during the three years of the trial were considered, and no effects on the costs after these three years were included.

The drug cost was based on the use of metoprolol, digitalis and diuretics as recorded in the clinical trial, and the official Swedish retail prices. The cost of in-patient care was based on the number of in-patient days for coronary heart disease during the trial and the number of coronary bypass operations. The cost per in-patient day was based on an estimate by the Swedish county councils of SEK 115 per day, and the cost per bypass operation was set at SEK 65,000 for the operation and SEK 30,000 for the follow-up, based on an estimation in a Swedish hospital.

The cost of out-patient care was based on the number of out-patient visits recorded in the trial and a cost per visit of SEK 370, as estimated in another study. The estimation of loss of production was based on the number of patients who had returned to work at the different follow-up visits, and the loss of production was defined as the production before the myocardial infarction minus the production after the myocardial infarction. The cost of lost production was based on the average salary cost in Sweden for an industrial worker. The result of the cost analysis is shown in Table 2.

The cost of drugs was SEK 2,280 per patient in the metoprolol group compared to SEK 480 in the placebo group. This difference was due to the increased use of metoprolol in the metoprolol group and represents the increase in programme costs.

The total number of in-patient days for coronary heart disease was 1,032 in the placebo group and 638 in the metoprolol group. The number of bypass operations was 9 in the placebo group compared to 3 in the metoprolol group. This led to a total cost for in-patient care of SEK 5,970 per patient in the metoprolol group and SEK 12,660 in the placebo group, i.e. a cost reduction in the metoprolol group of SEK 6,690 per patient. The difference in costs between the groups for out-patient care was small. In

Table 2. The cost per patient after myocardial infarction

Cost	Metoprolol	Placebo	Difference
Health care costs			
Drugs	2280	480	+1800
In-patient care	5970	12660	-6690
Out-patient care	4060	3980	+80
Total health care costs	12310	17120	-4810
Loss of production	106300	120100	-13800
Total costs	118610	137220	-18610

Source: Olsson et al (1987).

total the health care costs were SEK 12,310 in the metoprolol group and SEK 17,120 in the placebo group, i.e. a reduction in health care costs of SEK 4,810 in the metoprolol group.

The proportion of patients who returned to work was 30% after 9 months, 31% after 18 months, 28% after 24 months, 29% after 30 months and 25% after 36 months in the placebo group. The corresponding figures in the metoprolol group were 40%, 38%, 37%, 34% and 34%. On the basis of these figures, the loss of production due to morbidity was estimated to be SEK 106,300 in the metoprolol group and SEK 120,100 in the placebo group. The treatment with metoprolol thus led to an increase in production of SEK 13,800 (i.e. a reduced loss of production according to the terminology in the study). In total the costs decreased by SEK 18,610 in the metoprolol group during the three years. The authors concluded that the addition of beta-blockers after myocardial infarction is highly cost-effective, since it led to both increased health effects and reduced costs.

The result of this study is very clear, since metoprolol led to both decreased costs and increased health effects. The study was also based on solid data from a randomised clinical trial. The analysis shows that it is sometimes enough to carry out an analysis of the costs, and it is not necessary also to quantify the health effects in monetary terms.

Since both health care costs and earning losses due to illness are covered by public insurance in Sweden, the cost estimation can be viewed as an estimate of the external costs of the treatment and the change in morbidity. The study did not include any costs

due to the change in mortality (except the medical costs of coronary disease due to the increased survival in the beta-blocker group during the three years of follow-up, which was included because of the way these costs were calculated).

It could be argued that these costs should also have been estimated, in order to capture the total external costs. However, the results were very clear-cut and it is unlikely that this would have changed the conclusion (i.e. it is possible that the external costs of mortality would have been negative in this age-group, but all empirical studies on the WTP for increased survival indicate that the WTP far exceeds the consumption during the gained life years, which would be the upper bound on the increased costs of adding the external costs due to mortality).

The decreased costs due to the change in morbidity are probably underestimated as well, since a time-horizon of only three years was used. It is also possible that the treatment after a myocardial infarction involves some costs for relatives and friends, but these would also probably be lower in the beta-blocker group than in the placebo group.

The increased labour in the beta-blocker group would also be associated with a decrease in the consumption of leisure, which would be partially offset by an increased leisure due to less time being spent on health production (i.e. fewer in-patient days at hospital). Since the quality of life was improved among the patients, this indicates that the gain in reduced morbidity more than offset this possible loss in leisure time. The study did not include travel costs either, but these would not differ between the groups for the out-patient visits and would be lower in the beta-blocker group due to the decrease in in-patient care.

7.3 The Estimation Of Mortality Costs

Decreased mortality will increase the market production of the individual and it will also increase the consumption of health inputs, non-health goods and leisure. As was shown in Chapter 4, dealing with the resource consequences of health programmes, the only meaningful way of estimating resource consequences for changes in mortality appears to be to estimate the change in consumption minus the change in production due to the decreased mortality, as a measure of the external costs of the change in mortality.

A starting point for this estimation could be to try and estimate the mean consumption and production for different age groups. This could be done by following up a group of representative individuals and recording all their consumption and production. The difference between consumption and production in different age groups could then be used to estimate the external costs of increasing the probability of being alive at different ages. For instance, if for 50-year-old men the average consumption is $40,000 and the average market production is $50,000, the external cost of increasing

the probability of being alive at age 50 for a man is equal to the increase in the probability multiplied by -$10,000 ($40,000-$50,000).

If the survival time gained as a result of the health programme does not represent "average" survival, it is then possible to add on the extra consumption or production at different ages for the group whose survival is extended by the health programme. This may be relevant for instance if the survival gained is in a poor health state, with lower than average market production and higher than average health care costs. Of course, the total consumption and production for the group of individuals covered by the health programme could also be measured directly by following up a number of individuals in that group.

The above approach, however, neglects the dynamics of the change in mortality. It seems to be best suited for the case where health care costs, income losses due to disease and retirement payments are all covered by social insurance, or by insurance where the premium does not reflect the expected costs for the individual. In principle, when the mortality increases the individual may change his consumption per time unit, and so this change in consumption should also be added to the external costs (in principle, the production per time unit could also change, in which case this should also be added).

The above would for instance be true for the purely "private" case that was considered in Chapter 4, where the change in consumption minus production for an individual would be zero for the change in mortality. However, it may prove very difficult to measure these changes in consumption and production per time unit (it might be possible in a randomised clinical trial of a programme which increases survival). However, even in the purely private case, a change in mortality could lead to a change in the private lifetime savings of the individual (which was not included in the theoretical model in Chapter 4), resulting in an externality.

For instance, if elderly people increase their length of life and their consumption they may finance this by decreased savings. According to the general definition of the external costs, this should be included. Some may argue, however, that this externality is internalised in the individuals' decisions. If a household perspective is used to define the external costs (i.e. the change in consumption minus the change in production of the household), the change in savings which affects other household members' consumption will not be included in the external costs.

One alternative, instead of estimating the external costs as above by simply taking the difference between consumption and production at different ages, could be to look at the net contribution to the tax system and any insurances where the premium does not reflect the individual's expected costs. The external costs could then be estimated as the payments received from the tax and insurance systems and the consumption of publicly insured goods, minus the payments to the tax and insurance systems (for the insurances included in the estimation) at different ages and for different populations,

depending on the health programme being studied. This would measure the change in the net contribution to the insurance and tax systems due to the change in mortality. This would be similar to the estimation of the external costs of smoking and alcohol consumption which have appeared in the literature (Manning et al 1991).

Such an estimation would assume that prices of production and insured consumption (if the prices that the insurance company has to pay for the insured consumption are used) reflect opportunity costs, and it would also ignore any changes in private lifetime savings. However, it would be possible to adjust the measure for differences in these prices and opportunity costs. For example, if health care costs are insured and the price of health care in general exceeds opportunity costs by 20%, then the consumption of health care costs based on the prices paid by the state or the insurance company can be decreased by 20%.

7.4 Conclusions

In this chapter the estimation of costs in a cost-benefit analysis was considered. The cost estimations were divided into the programme costs, the morbidity costs and the mortality costs. The most important of these distinctions is that between the programme and morbidity costs and the mortality costs.

It was argued that the only meaningful way of estimating the costs due to changes in mortality would be to try and estimate the external costs of the change in mortality. For the morbidity case, it would also be possible to estimate the costs borne by the individuals who receive the health care programme, if the WTP does not already include these costs. For the programme costs, all costs could be estimated unless they are included in the WTP of the programme (if the WTP is the WTP of the programme at a zero price, then in most cases the majority of the programme costs would not be included in the WTP, even if it included all the changes of the programme).

The estimation of three types of cost was considered according to the theoretical model in Chapter 4. These three types were for the changes in consumption of health inputs (e.g. health care), the changes in market production, and the changes in consumption of leisure. Of these three items, the estimation of the changes in leisure was shown to be the most difficult since the price of leisure is not readily observable on the market (unless we assume that individuals work the optimum number of hours), and it may also differ depending on how the time is used (i.e. the price of leisure is really the difference between the values of leisure time and working or health production time per se, and this may differ depending on the type of work and the type of health production activity).

For morbidity changes, the value of leisure may also differ depending on the health state in which it is evaluated, and the definition of the WTP for the pure health change determines the health state in which the value of leisure should be evaluated. This

would indicate that it is preferable to try and include the changes in leisure in the WTP of the programme. Since the change in leisure is related to the individual's production (i.e. more hours worked leads to less leisure), the change in production that falls on the individual should ideally be incorporated in the WTP as well (an exception could be the leisure and productivity costs of the programme as such; it seems to be easier to estimate and to separate these from the WTP of the programme).

It was argued that the external costs due to morbidity and mortality changes can be estimated by trying to estimate the change in market consumption minus the change in market production. Alternatively, the external costs could be estimated by measuring the change in the payments from the tax and insurance systems and the consumption of publicly provided and insured goods, minus the payments to the tax and insurance systems (including only those insurances whose premiums do not reflect the expected costs, see Chapter 4). If the latter approach is used, this could also be corrected for possible differences between prices and opportunity costs. Measuring the net contribution to the tax and insurance systems would mean that changes in private lifetime savings are not included in the external costs.

REFERENCES

Drummond MF, Stoddart GL, Torrance GW. Methods for the economic evaluation of health care programmes. Oxford: Oxford Medical Publications, 1987.

Finkler SA. The distinction between costs and charges. Annals of Internal Medicine 1982;96:102-109.

Johannesson M, Borgquist L, Jönsson B. The costs of treating hypertension in Sweden: an empirical investigation in primary health care. Scandinavian Journal of Primary Health Care 1991;9:155-160.

Manning WG, Keeler EB, Newhouse JP, Sloss EM, Wasserman J. The costs of poor health habits. Cambridge MA.: Harvard University Press, 1991.

Olsson G, Levin L-Å, Rehnqvist N. Economic consequences of postinfarction prophylaxis with beta-blockers: cost-effectiveness of metoprolol. British Medical Journal 1987;294:339-342.

Olsson G, Lubsen J, van Es G-A, Rehnqvist N. Quality of life after acute myocardial infarction: effects of chronic metoprolol treatment on mortality and morbidity. British Medical Journal 1986;292:1491-1493.

Olsson G, Rehnqvist N, Sjögren A, Erhardt L, Lundman T. Longterm treatment with metoprolol after myocardial infarction: effect on 3 year mortality and morbidity. Journal of the American College of Cardiology 1985;5:1428-1437.

8. ADDITIONAL ISSUES IN COST-BENEFIT ANALYSIS

In this chapter we deal with two additional issues concerning the estimation of costs and benefits. The first issue is how to adjust benefits and costs according to their timing (i.e. the discounting of costs and benefits) and the second issue is how to deal with taxes in a cost-benefit analysis.

8.1 The Discounting Of Costs And Benefits

In a cost-benefit analysis it is common to adjust costs and benefits according to the time period when they occur. This is usually referred to as discounting. First it is important to state that discounting has nothing to do with inflation, and concerns the adjustment of benefits and costs expressed in real money terms at different points in time. Discounting means in practice that costs and benefits in the future are given a lower weight compared to costs and benefits at present.

There are two arguments in favour of discounting. The first argument is the growth argument. Resources available now can be invested and grow in value over time. According to this argument, the discount rate should reflect the real rate of return on investments. If for example resources invested now grow in value with 5% per year in real terms, the real rate of return is 5%. The opportunity cost of consuming $1 this year is thus the $1.05 in consumption next year that could have been achieved if the $1 had been invested rather than consumed. This means that $1 this year is worth the same as $1.05 next year, since the $1 this year can be converted to $1.05 next year.

In the absence of market imperfections, taxes and uncertainty (and transaction costs of borrowing and lending money) the real interest rate (i.e. the interest rate adjusted for inflation) will also equal the marginal real rate of return on investments. For example, if the real interest rate fell short of the real rate of return on investments, profits could be made by borrowing money and investing it, and the interest rate would fall until it equalled the real marginal rate of return on investments. The real interest rate could then be used as the discount rate to convert future costs and benefits into present value terms.

The second argument behind discounting is based on the time-preference argument and the marginal rate at which individuals are willing to substitute consumption now for consumption in the future. Pure time-preference is the preference for consumption now relative to the future at the same consumption level now and in the future (and everything else held constant). It is often assumed that individuals have a positive pure time preference, i.e. they intrinsically prefer consumption now rather than consumption in the future. According to the time preference argument, the discount rate should be based on the rate at which individuals are willing to trade present consumption for future consumption (i.e. the time preference at the current

consumption level in each period; this need not equal the pure time preference unless the consumption is the same in each period).

An individual who is faced with an interest rate and the ability to borrow and lend money will allocate the consumption between periods, so that the marginal rate at which he is willing to substitute present for future consumption will equal one plus the real interest rate. If the interest rate is 5% the individual will allocate consumption between the first year and the second year until $1 of consumption this year is valued the same as $1.05 of consumption next year (the marginal rate of substitution between year 1 and year 2 is thus equal to 1.05). If for instance the individual initially had a marginal rate of substituion between year 1 and year 2 of 1, he would borrow money and increase the consumption in period two until the marginal rate of substitution between the periods was equal to 1.05. In the perfect market case outlined above, the time preference between consumption now and consumption in the future for all individuals will thus equal the real interest rate (e.g. even an individual who has a pure time preference that is negative will adjust the time preference to the ruling market interest rate). The time preference will also equal the real rate of return on investments, and the two arguments for discounting yield the same discount rate.

If taxes on income are introduced this is no longer the case, however, and taxes will lead to the real rate of return on investments exceeding the time preference. This is because individuals will adjust the time preference to the after-tax real interest rate; for investments to yield the same after-tax return, the real rate of return on investments has to be higher (i.e. with a 50% tax, a real return on investments of 5% is needed in order to yield an after-tax return of 2.5%).

There has been some debate over whether an individual's time preference rate (the real interest rate after tax) or the real rate of return on investments (the pre-tax interest rate) should be the basis for the discount rate in cost-benefit analysis. The time preference seems to be a natural candidate, since it converts future consumption to present consumption for an individual. Ideally, each individual's time preference should be used to convert costs and benefits to present value terms, since the time preference may differ due to market imperfections, different taxes etc.

However, if a project is financed by reduced private investments, the opportunity costs of these investments may diverge from the time preference for consumption. If a project is financed by reducing private investment, the real rate of return on investments may thus be appropriate and when the project is financed by reduced private consumption, the time preference may be appropriate (Johansson 1991).

If the project is financed by a combination of reduced private investments and reduced private consumption, this would mean using a weighted average of the pre-tax and after-tax real interest rates. An alternative way of saying this is that the rate of time preference should always be used to carry out the discounting, but the cost of reduced private investments should be added to the cost of the project. In other words, the

difference between the rate of return on private investments and the rate of time preference should be added as a cost for the private investments which are crowded out.

In addition to the problem of whether the pre-tax or after-tax real interest rate should be used, a further complication is that there is more than one interest rate on the market. A number of studies have tried to estimate the real interest rate for different time periods. Perhaps the most interesting of these studies is one by Barro & Martin (1990). These authors estimated an index of real expected interest rates (based on the expected inflation rate) and real actual short-term interest rates (based on the actual inflation rate) for nine major industrialised countries for the time period 1980-1989. The average real expected pre-tax interest rate varied between 0.0% and 4.2% in the different time periods studied, and the real pre-tax actual interest rate varied between -1.0 and 5.3%.

In economic evaluations in the health care field, a discount rate of 5% is often used in the base-case analysis. This rate seems high both in comparison with the real interest rate in the Barro & Martin (1990) study and in comparison with the current growth rate of most economies in the world. In his recent book about environmental economics, Freeman (1993) concluded that a discount rate of 2-3% seemed to be appropriate for cost-benefit analysis. A useful recommendation for practitioners may thus be to use a discount rate of 2-3% in the base-case analysis and complement this with a sensitivity analysis where the discount rate is varied between 0% and 5%. This should be viewed as the discount rate for cost-benefit analysis of projects where the costs and benefits accrue to the present generation. It is not obvious that the same discount rate should be used for costs and benefits for future generations (Freeman 1993).

There are a number of formulae which are useful for discounting costs. In the standard case we would calculate the present value of costs and benefits. A future cost or benefit can then be converted to the present value by multiplying it by the discount factor. The discount factor is given by the following formula:

$$1/(1+r)^n \qquad (1)$$

In equation (1) r is the discount rate and n is the number of years from now. Note that the first year of a programme can be defined as zero years from now or 1 year from now, depending on whether the costs (benefits) are incurred at the beginning of each year or towards the end of each year. Formula (1) can also be illustrated with an example. Assume that a cost of $10,000 is incurred 10 years from now and that we want to estimate the present value of this cost and that the discount rate is 2%. The discount factor is then 0.82 and the present value of the $10,000 is $8,200 (10,000*0.82).

If we have a recurring annual cost or benefit, then instead of estimating the discount factor for each year it is possible to use a simpler formula. The annual cost (benefit) can simply be multiplied by the annuity factor to get the present value of the cost (benefit) stream. The annuity factor is given by the following formula:

$$[1-1/(1+r)^n]/r \qquad (2)$$

In equation (2) n is the number of years that the annual cost (benefit) is incurred and r is the discount rate. The above formula for the annuity factor assumes that the costs (benefits) are incurred towards the end of each year. If instead the costs (benefits) are incurred at the beginning of each year, the annuity factor should be derived for n-1 years and 1.0 should be added to the annuity factor. The use of the annuity factor for present-value calculations of annual cost (benefit) streams can also be illustrated by an example. Assume that an annual cost of $10,000 is incurred towards the end of each year for 10 years and that the discount rate is 2%. The annuity factor then becomes 8.98 and the present value of the cost stream becomes $89,800 (8.98*10000).

Sometimes it is preferable to convert costs and benefits into annual terms, instead of calculating the present value of costs and benefits. An example of this is when a programme has a large investment cost in year zero and then yields annual costs and benefits for a number of years. The annuity factor can then be used to convert the investment cost to an annual cost, and then the total annual costs and benefits can be compared in order to determine whether benefits exceed costs. To convert an investment cost to an annual cost, the investment cost is divided by the annuity factor. Assume that we have an initial investment cost of $100,000 and that we want to convert this initial investment cost to an annual cost which is incurred towards the end of each year for 10 years with a discount rate of 2%. The annuity factor then becomes 8.98 and the annual cost becomes $11,136 (100,000/8.98).

8.2 The Treatment Of Taxes

In a cost-benefit analysis it is important that benefits and costs are compared at the same prices. In such a comparison, the presence of taxes creates a problem. Since the WTPs of individuals are based on the consumer prices including taxes (i.e. both taxes on inputs that are passed on to the consumers and sales taxes), it is necessary for the costs also to be based on the same consumer prices, so as to make them comparable.

The key to understanding how to incorporate taxes is to use the concept of opportunity cost. For inputs that are used in a health care programme, the taxes on the input should be included in the cost if the alternative use of the input is also taxed, i.e. we are comparing the benefits of the health care programme with the benefits of the alternative use of the resources in consumer prices.

This means that if an input is taxed only when it is used in the programme, but not in its alternative use, then the tax should not be included. Assume for example that a health care programme leads to 10 more nurses being employed, and that these nurses were previously working in other occupations. Assume further that the employer has to pay an extra tax for the nurses of $50 per month and that this extra tax was not imposed in the previous occupations of the nurses. In this case the tax should not be included in the cost of the nurses, since the alternative use was not taxed. We could also change the example and now assume that no extra tax is paid for the nurses, but that a tax of $50 was paid for them in their previous occupation. In this case the tax should be added to the cost, since the alternative use was taxed.

In general, in the case where the tax of an input is taxed in all its uses the price including the tax should be included, as long as the programme does not lead to an increase in the total supply of the input. If the supply of the input increases, the increased supply should be assessed at the price without the tax. Again, the rationale for this is easily seen if we consider the labour case. Assume now that the tax on labour is the same in all occupations. If in the example above the total supply of labour does not increase when the 10 nurses are employed in the health care programme (i.e. they are taken from other occupations), the tax should be included in the cost of the nurses. If on the other hand the total supply of labour increases by 10 persons (i.e. these persons did not work previously), then the tax should not be included in the cost of the nurses. This is because the cost of employing the nurses is now the leisure time that they have to give up and the leisure time is not taxed.

The value of the leisure time for the nurses would be their reservation wage for accepting the job (i.e. the pre-tax wage rate). If, however, the nurses were unemployed and received unemployment benefits then the after-tax benefits should be deducted from the reservation wage, since the value of the leisure time would be the increased income that would induce the nurses to work rather than to continue to be unemployed (i.e. the after-tax wage minus the after-tax unemployment benefits).

It is often the case that no sales tax or value added tax is included in the cost of health care. Since the sales tax is imposed on the alternative production (i.e. the goods produced by the inputs in their alternative uses), however, it seems appropriate to add the sales tax to the health care costs in a cost-benefit analysis (unless the health care programme under study has increased the supply of the inputs, e.g. by employing persons who were previously unemployed).

The same principles as for inputs apply also to outputs. If a project increases the output of some goods sold on the market, these goods should be evaluated at the market price including all taxes on the good. If for instance a health care programme leads to some patients returning to work and increases their production, this increased production should be valued at market prices including all taxes. The market price will then be the total price that the firm pays for the worker, assuming that the firm pays on the margin the same price as it gets for the output of the worker on the market, plus any

additional taxes paid directly by the consumers, such as sales taxes on the final product that the worker produces. Similarly, if a project leads to reduced production then this reduced production should be valued at the market prices, including taxes on that production.

One input whose price should not be adjusted for taxes is leisure. No tax should be added on decreased leisure, since leisure is not taxed, and for the same reason no tax should be added to increased leisure. The key to dealing with taxes is to remember at all times that the benefits of the programme, valued in terms of market prices, should be compared with the benefits in terms of the value of the resources in the alternative use, again valued in terms of market prices.

Taxes also impose an additional problem, since increased taxation is normally assumed to lead to welfare losses, i.e. so-called deadweight losses (Boadway & Bruce 1984). In the definitions of costs (i.e. resource consequences) in Chapter 4, lump-sum taxes were assumed which did not lead to any inefficiences. Normally, however, it is not possible to collect taxes as a lump sum, and taxes will then lead to inefficiences. This can be seen by considering a tax on labour income, for instance. If labour income is taxed, the worker adjusts to the after-tax wage rate and the firm adjusts to the pre-tax wage rate. If for instance the firm pays $100 per hour for the worker and the worker pays 50% of this in tax, the worker will choose the number of hours so that for the last hour worked, the value of the leisure time given up equals the after-tax wage rate of $50. Even though the value of the production exceeds the cost of the leisure time of $50 for additional hours worked, this economically motivated exchange will not take place, and the tax thus creates a welfare loss.

This means that the welfare losses (the deadweight losses) due to increased taxes of a health care programme should be included among the costs in a cost-benefit analysis. In practice it may be very difficult to estimate the size of the welfare losses due to taxation, and they are also likely to differ for different types of tax, i.e. a tax on labour income versus a sales tax.

8.3 Conclusions

This chapter considered the discounting of costs and benefits and how to deal with taxes. The two arguments for discounting costs and benefits were reviewed: the growth argument and the time preference argument. It was argued that the 5% discount rate used currently in many economic evaluations in the health care field seem too high. Instead it may be more appropriate to use 2-3% in the base-case analysis, as proposed by Freeman (1993), and to vary this rate in a sensitivity analysis.

We also showed how to deal with taxes in a cost-benefit analysis. Since the WTP of a health care programme is based on the consumer prices including all taxes, the costs have to be adjusted to include taxes as well. Consequently if the input of a health care

programme is taxed in its alternative use, the tax should be included in the price of the health input. The same is also true for any taxes on the output that are produced by the input in its alternative use (e.g. a sales tax). If a health care programme increases the output of some good (e.g. increased production due to reduced morbidity), the increased output should be valued in the market prices including all taxes.

In principle, if a health care programme leads to increased taxation, the welfare losses due to the increased taxation as such should be included as well, but it may be difficult to estimate these deadweight losses of increased taxation.

REFERENCES

Barro RJ, Martin XS. World real interest rates. NBER Macroeconomics Annual 1990;5:15-59.

Boadway RW, Bruce N. Welfare economics. Oxford: Blackwell, 1984.

Freeman AM. The measurement of environmental and resource values: theory and methods. Washington D.C.: Resources for the Future, 1993.

Johansson P-O. An introduction to modern welfare economics. Cambridge: Cambridge University Press, 1991.

9. COST-EFFECTIVENESS ANALYSIS

This chapter is devoted to cost-effectiveness analysis. In a cost-effectiveness analysis the costs are measured in monetary terms and the health effects are measured in non-monetary terms, e.g. the number of life-years gained. The ratio between costs and health effects is then estimated as e.g. the cost per gained life-year. Cost-effectiveness analysis of health care programmes was developed in the medical field, as a response to the criticism about the use of the human-capital approach for valuing health changes in monetary terms, and the first study in the health care field by Klarman et al appeared in 1968.

The underlying rationale for cost-effectiveness analysis is to try and maximize the health effects for a given amount of resources (Weinstein & Stason 1977; Weinstein 1990). Below we will show the decision rules that cost-effectiveness analysis is based on, i.e. the decision rules that can be used to maximize the health effects for a fixed budget. Thereafter we discuss the choice of effectiveness measure in a cost-effectiveness analysis. We then discuss the relationship between cost-effectiveness analysis and cost-benefit analysis. It is argued that cost-effectiveness analysis may best be viewed as part of a cost-benefit analysis where the cost of producing health effects is estimated, and in order to use the analysis for societal decision-making it has to be complemented with information about the WTP for the health effects. We also discuss the role of discounting within cost-effectiveness analysis, since the discounting of health effects is controversial. The chapter ends with an application of cost-effectiveness analysis and some conclusions.

9.1 Maximization Of Health Effects

In this chapter we demonstrate the decision rules of cost-effectiveness (C/E) analysis (Weinstein & Zeckhauser 1973; Weinstein 1990; Johannesson & Weinstein 1993). We make a distinction between independent programmes and mutually exclusive programmes. Two programmes, A and B, are defined as independent if the costs and effectiveness of programme A (B) are not affected by whether programme B (A) is implemented or not. The two programmes are viewed as applying to two different populations; an example would be the treatment of ulcer patients and the treatment of arthritis patients.

Two programmes, A and B, are mutually exclusive if implementing programme A (B) means that programme B (A) cannot also be implemented. If the programmes A and B can both be implemented physically, but implementing A (B) means that the costs and/or effectiveness of B (A) change, then the programmes should be defined as mutually exclusive. In this case we have three alternatives: carrying out only A, carrying out only B, or carrying out both A and B. Mutually exclusive programmes can be viewed as programmes for the same population, e.g. two alternative drugs for ulcer patients.

Before we continue with the decision rules, it can be useful to make another distinction which will be used below, namely that between average cost-effectiveness ratios and incremental cost-effectiveness ratios; the latter are sometimes also referred to as marginal cost-effectiveness ratios. An average cost-effectiveness ratio is equal to the cost of a programme divided by the effectiveness of the programme compared to "doing nothing" (i.e. the base-case alternative). An incremental cost-effectiveness ratio is the incremental cost of a programme divided by the incremental effectiveness compared to the next most effective programme. If only one programme is compared with "doing nothing" then the incremental C/E-ratio will be the same as the average C/E-ratio.

9.2 Independent Programmes

In this section we will analyse the decision rules for cost-effectiveness analysis when only independent programmes are compared. The decision rules show how the health effects can be maximised for a given budget. Either a fixed budget can be specified, which is used to maximise the health effects (which will implicitly yield a price per unit of health effects on the margin), or a price per unit of health effects can be specified (which will implicitly yield a budget).

The decision rule for cost-effectiveness analyses of independent programmes is that they should be ranked according to their C/E-ratios. With a fixed budget, programmes should then be implemented in order of their cost-effectiveness ratios until the budget is exhausted. With a price per unit of health effects, all programmes with a cost-effectiveness ratio below or equal to the price/cut-off should be implemented.

The decision rules can be illustrated with a simple example. Assume that there are three different independent health care programmes available, i.e. three health care programmes for three different patient groups. Assume further that each programme is the only one available for that patient group, so that the three programmes are compared with doing nothing for each patient group (i.e. the average and incremental cost-effectiveness ratios coincide for each patient group).

The costs and effects (in terms of life-years gained) of each programme are shown in Table 1, where the programmes have also been ranked according to their cost-effectiveness ratios. The cost per patient of programme 1 is $1,000 and the number of life-years gained per patient is 0.1, yielding a cost-effectiveness ratio of $10,000. The cost of programme 2 is $4,000 per patient and the number of life-years gained is 0.2 per patient, yielding a cost-effectiveness ratio of $20,000. Finally, the cost per patient of programme 3 is $9,000 and the number of life-years gained per patient is 0.3, yielding a cost-effectiveness ratio of $30,000. These cost-effectiveness ratios can be viewed as a ranking of the different programmes in terms of their desirability, i.e. it is possible to say that programme 1 should be given priority over programme 2 since it has a lower cost-effectiveness ratio. However, in order to

determine which of the available programmes should be implemented, we have to know the size of the budget or the price per life-year gained.

Table 1. Example of three independent programmes

Programme	C	E	C/E
1	1000	0.1	10000
2	4000	0.2	20000
3	9000	0.3	30000

Assume that there are 100 patients in each patient group and that this number is exhaustive. First consider the case of a fixed budget. In Table 2 the choice for different sizes of budget is shown. If we have a budget of $100,000 only programme 1 will be carried out, since this programme will maximise the effects for the given budget (the cost per patient for programme 1 multiplied by the number of patients (100) yields $100,000).

If the budget increases to $500,000, both programmes 1 and 2 will be carried out, since this maximises the effectiveness for that budget. Finally, if the budget increases further to $1,400,000 all the programmes 1, 2 and 3 will be carried out. The decision rule is thus to start implementing the programme with the lowest cost-effectiveness ratio and then to add additional programmes according to their cost-effectiveness ratios until the budget is exhausted.

Table 2 only shows the budgets which lead to all patients in a patient group being given the treatment or not being given the treatment. It is also possible to have a budget so that some patients in a patient group receive the treatment and other patients in the patient group do not receive the treatment. If in the example the budget had been below $1,400,000 and not $500,000 or $100,000, then some patients in a patient group would have received treatment and some patients would not. This is sometimes called a mixed solution, since not all patients in a patient group are given the same treatment. If for example the budget had been $300,000 then programme 1 would have been implemented for all patients, but programme 2 would only have been implemented for 50 patients in patient group 2.

Note also that the different budgets yield an implicit price that we are willing to pay per life-year gained on the margin. For instance, if a budget of $500,000 is used this implies that we are willing to pay $20,000 per life-year gained on the margin (i.e. it is the price we are paying per life-year gained for the programme with the highest cost-effectiveness ratio that is implemented).

Table 2. The choice of independent programmes for different sizes of the budget

Budget	Choice of Programmes
100000	1
500000	1+2
1400000	1+2+3

We can then also consider the decision rule when we know the price per life-year gained that we are willing to pay. Table 3 gives the choice of programmes for different prices per life-year gained. If we know the price that we are willing to pay per life-year gained, we just implement all programmes with a cost-effectiveness ratio below or equal to this price.

With a price below $10,000 no programme will be implemented, with a price between $10,000 and $19,999 programme 1 will be implemented, with a price between $20,000 and $29,999 programmes 1 and 2 will be implemented, and with a price of $30,000 or higher programmes 1, 2 and 3 will be implemented. Note also that the different prices will implicitly yield a budget, e.g. if a price of $10,000 is used this will yield a budget of $100,000 (i.e. only programme 1 will be implemented with this price and the total cost of programme 1 is $100,000).

Table 3. The choice of independent programmes for different prices

Price/cut-off	Choice of Programmes
<10000	NO PROGRAMME
10000-19999	1
20000-29999	1+2
30000-	1+2+3

The decision rules, for ranking programmes in terms of their cost-effectiveness ratios and implementing programmes until the budget is exhausted, are based on the assumption that the cost-effectiveness ratio of a programme is independent of the size of the programme. This means that if there are 100 patients in a patient group, the cost-effectiveness ratio of a treatment for this patient group should be the same irrespective of whether 1 patient, 10 patients or 100 patients are treated (for mutually exclusive programmes, see below, the incremental cost-effectiveness ratio should be the same). We will refer to this assumption as constant returns to scale. Note, however, that in principle this assumption requires both constant returns to scale in the usual sense, i.e. that the amount of inputs needed to produce an additional unit is constant with the scale, and also that the prices of inputs are constant with the scale.

One obvious case where this will not be true is for indivisibilities of a programme, e.g. an investment cost that is needed to start a programme and is independent of the size of the programme. This can be illustrated using the above example and assuming that part of the cost of programme 1 is an investment cost of $75,000; then it also costs $250 per treated patient (yielding a cost-effectiveness ratio of $10,000 for treating 100 patients). Assume then that the budget is only $80,000.

If the budget is spent on programme 1, only 20 patients in that patient group (the budget minus the investment cost divided by the marginal cost per patient) can be treated, yielding 2 gained life-years. If the 80,000 are spent on programme 2, 20 patients in patient group 2 (the budget divided by the cost per patient) can be treated, yielding 4 gained life-years. The budget should thus be spent on programme 2, since this maximizes the health effects, rather than on programme 1, in spite of the higher cost-effectiveness ratio of programme 2. However, if the cost-effectiveness ratio varies with the scale of a programme then the different scales should in principle be defined as mutually exclusive programmes. The decision rules for mutually exclusive programmes are explored below.

9.3 Mutually Exclusive Programmes

We can then continue and consider the decision rules for mutually exclusive programmes. As for independent programmes, either a fixed budget or a price per unit of effectiveness can be used as the decision rule (i.e. to decide which programme should be implemented). In this case the mutually exclusive programmes should first be ordered according to effectiveness. The incremental cost-effectiveness ratios should then be estimated. After this has been done, dominated alternatives should be excluded. The incremental cost-effectiveness ratios should then be re-calculated without the dominated alternatives being included.

This process continues until a number of programmes with increasing incremental cost-effectiveness ratios remain. With a fixed budget we then start by considering the least effective programme, and then continuously switch patients to more effective programmes until the budget is exhausted. With a price/cut-off we implement the programme with the highest incremental cost-effectiveness ratio that is equal to or below the price.

This can also be illustrated with an example. Assume that there are four different alternative programmes (e.g. four different lipid lowering drugs) for treating a patient group consisting of 100 patients. The costs and effects per patient are shown in Table 4.

Note first that since the four programmes are mutually exclusive, it is only possible to carry out one of the programmes for each patient. Column 1 of the table shows the costs per patient, and the second column shows the effects per patient. The third

Table 4. Example of four mutually exclusive programmes

Programme	C	E	C/E	INCR.C/E All	INCR.C/E without 4C
4A	500	0.1	5000	5000	5000
4B	2000	0.2	10000	15000	15000
4C	9000	0.3	30000	70000	-
4D	10000	0.4	25000	10000	40000

column shows the average cost-effectiveness ratios, i.e. each programme compared to the do-nothing base-line alternative. However, the average cost-effectiveness ratios are misleading and should not be used.

The first step in using cost-effectiveness analysis for mutually exclusive programmes is to rank the programmes according to their effectiveness, which has been done in Table 4. Then the incremental cost-effectiveness ratios should be calculated for each programme. For 4A the incremental cost-effectiveness ratio is equal to the costs of 4A compared to doing nothing (500) divided by the effects of 4A compared to doing nothing (0.1), which gives an incremental cost-effectiveness ratio of $5,000. This is the same as the average cost-effectiveness ratio for 4A, since it is the least effective programme.

For 4B the incremental cost-effectiveness ratio is equal to the costs of 4B minus the costs of 4A (2,000-500=1,500) divided by the effects of 4B minus the effects of 4A (0.2-0.1=0.1), which gives an incremental cost-effectiveness ratio of $15,000 (1,500/0.1). For 4C the incremental cost-effectiveness ratio is equal to the costs of 4C minus the costs of 4B (9,000-2,000=7,000) divided by the effects of 4C minus the effects of 4B (0.3-0.2=0.1), which gives an incremental cost-effectiveness ratio of $70,000 (7,000/0.1).

For 4D the incremental cost-effectiveness ratio is equal to the costs of 4D minus the costs of 4C (10,000-9,000=1,000) divided by the effects of 4D minus the effects of 4C (0.4-0.3=0.1), which gives an incremental cost-effectiveness ratio of $10,000 (1,000/0.1).

The next step is to exclude dominated alternatives. A programme is dominated if the incremental cost-effectiveness ratio decreases for the next programme with higher effectiveness. The reason for excluding the dominated alternative is that irrespective of the size of the budget, more units of effectiveness are always gained by doing the more effective programme (provided that the more effective programme exhibits constant returns to scale; see below).

In this example programme 4C is dominated, since the incremental cost-effectiveness ratio for 4D decreases. Therefore 4C should be excluded from further consideration. After 4C has been excluded, the incremental cost-effectiveness ratios should be re-calculated. The incremental cost-effectiveness ratios for 4A and 4B will be the same as before, but the incremental cost-effectiveness ratio for 4D will change. For 4D the incremental cost-effectiveness ratio is now equal to the costs of 4D minus the costs of 4B (10,000-2,000=8,000) divided by the effects of 4D minus the effects of 4B (0.4-0.2=0.2), which gives an incremental cost-effectiveness ratio of $40,000 (8,000/0.2).

The incremental cost-effectiveness ratios show the implied price per life-year gained from implementing the different programmes. The ratios cannot, however, be used to rank the programmes, since without knowing the size of the budget or the price per life-year gained that society is willing to pay, it is impossible to know which of the four mutually exclusive programmes should be implemented.

If the fixed budget is used as the decision rule, we start by considering the programme with the lowest incremental cost-effectiveness ratio, and then switch patients to continuously more effective programmes until the budget is exhausted. In Table 5 the programme that should be implemented is shown for different sizes of the programme. In this case it is important to note that since the programmes are mutually exclusive, each patient can only receive one of the programmes (4A-4D). This means that as the budget increases we will switch from a less effective to a more effective programme, and the difference compared to the case with independent programmes is that we are now switching from one programme to another rather than adding more programmes as for the case with independent programmes.

For a budget of 50,000 programme 1 will be implemented. This can be seen by starting by considering programme 4A; if all patients are treated with programme 4A all the budget of $50,000 will be exhausted, so it is not possible to switch any patients to more effective treatments. If the budget is above $50,000 we can start to switch patients from treatment 4A to 4B, and if the budget is $200,000 we will be able to switch all the patients to 4B.

If the budget exceeds $200,000 we can continue, and start switching patients from 4B to 4D, and if the budget is $1,000,000 or greater all the patients will receive the most effective treatment, 4D. The different budgets also implicitly yield the price we are willing to pay for extra life-years gained, e.g. for a budget of $200,000 the implied price per life-year gained on the margin is $15,000 (i.e. the incremental cost-effectiveness ratio for programme 4B).

Note that in Table 5 the budgets which lead to all patients receiving the same treatment are shown. It should be stressed that it is also possible to have budgets so that all patients do not receive the same treatment. If the budget in the example is below $1,000,000 and is not $200,000 or $50,000, all the patients will not receive the

Table 5. The choice of mutually exclusive programme for different sizes of the budget

Budget	Choice of Programme
50000	4A
200000	4B
1000000	4D

same treatment, i.e. a so-called mixed solution will result.

If for instance the budget is $520,000, then 60 patients will receive treatment 4B and 40 patients will receive treatment 4D. This can be estimated by dividing the money left in the budget, after giving all patients 4B (520,000-200,000=320,000), by the cost difference between 4D and 4B (10,000-2,000=8,000) so as to get the number of patients who will receive 4D (320,000/8,000=40 patients), and then the remaining patients will receive 4B (100-40=60 patients).

The rationale for excluding dominated alternatives can also be illustrated by using different budgets. When the budget increases we switch patients to more effective treatments. In the example, once the budget exceeds $200,000 so that we can start switching patients to a more effective treatment than 4B, we should start switching patients directly to 4D rather than to the dominated alternative 4C. If the budget for instance is $312,000 we could treat 16 patients with 4C ([312,000-200,000]/[9,000-2,000] and 84 patients with 4B, yielding 21.6 life-years (16*0.3+84*0.2). For the same budget, however, we could treat 14 patients with 4D ([312,000-200,000]/[10,000-2,000] and 86 patients with 4B, yielding 22.8 life-years (14*0.4+86*0.2). Irrespective of the size of the budget, 4C should never be given, since when the budget exceeds 200,000 the extra money will always yield more life-years if spent on 4D rather than 4C (note however that this is based on the assumption of constant returns to scale, see below).

We can also use the price per life-year gained as the decision rule. This case is depicted in Table 6. If we know the price per gained life-year that we are willing to pay, we implement the programme with the highest incremental cost-effectiveness ratio that is at or below this price. In the example, if the price is below $5,000 no programme will be implemented, with a price between $5,000 and $14,999 programme 4A will be implemented, with a price between $15,000 and $39,999 programme 4B will be implemented, and with a price of $40,000 or higher programme 4D will be implemented. Note also that as in the independent case, the different prices will implicitly yield a budget, e.g. if a price of $5,000 is used this will yield a budget of $50,000 (i.e. only programme 4A will be implemented with this price and the total cost of programme 4A is $50,000). The difference compared to the

case with the independent programmes is that now only one of the programmes can be implemented; thus as the price increases we will switch the patients to a more effective programme.

The case with the price per gained life-year also shows the importance of using the incremental cost-effectiveness ratios. The incremental cost-effectiveness ratios show the incremental cost of adopting a more effective programme; in order for it to be beneficial to implement this programme, the price that we are willing to pay for extra life-years on the margin has to exceed this cost. For instance, programme 4D has an average cost-effectiveness ratio of $25,000, and if the price per life-year gained was $30,000 then the use of the average cost-effectiveness ratio would imply that programme 4D should be implemented. However, the incremental cost-effectiveness ratio is $40,000, and this correctly shows that the incremental cost of implementing 4D exceeds the incremental benefits. In a sense then, we should implement increasingly more effective programmes as long as the incremental benefits (expressed as the price per gained life-year that we are willing to pay) exceed the incremental costs (expressed as the incremental cost-effectiveness ratio).

Table 6. The choice of mutually exclusive programme for different prices

Price/cut-off	**Choice of Programme**
<5000	NO PROGRAMME
5000-14999	4A
15000-39999	4B
40000-	4D

The decision rules for mutually exclusive programmes are also based on the assumption of constant returns to scale; this can here be interpreted as meaning that every dollar increase in the cost of a more effective programme, compared to the next most effective programme, should give the same increase in the effectiveness of the more effective programme compared to the next most effective programme (i.e. the incremental cost-effectiveness ratio is independent of the size of the programme). This assumption guarantees that a dominated alternative will never be chosen, irrespective of the size of the budget.

For instance, if we return to the example above when we showed that 4C should never be given in spite of the size of the budget, this conclusion could change if constant returns to scale do not hold. If for instance 4D has an investment cost of $400,000, it is not possible to switch any patients to 4D with a budget of $312,000; the combination of 4B and the dominated alternative 4C in the example above, yielding 21.6 gained life-years, would therefore maximize the effectiveness for this budget (given that the dominated alternative 4C exhibits constant returns to scale).

As was noted above, if constant returns to scale do not hold, different scales of a programme should in principle be defined as mutually exclusive programmes and the incremental cost-effectiveness should be calculated for the different scales of the programme. The constant returns to scale assumption then only has to hold for each increment in costs that is analysed. But even if this is done, the assumption of constant returns to scale within each increment will be violated if there are increasing returns to scale, since in that case the full scale of a programme will dominate all the other scales of that programme (i.e. the scales that are less than the full scale of a programme will be excluded as dominated). Thus the assumption of constant returns to scale within the increment in costs due to such a programme will be violated.

9.4 Independent And Mutually Exclusive Programmes

In the general case we will have a combination of independent and mutually exclusive programmes available. In the most realistic case we would have a number of mutually exclusive programmes within each patient group, and the different mutually exclusive programmes within one patient group are then independent with respect to the mutually exclusive programmes in another patient group.

With both types of programme the incremental cost-effectiveness ratios should first be estimated within each patient group, and dominated alternatives should be excluded until a number of programmes with increasing incremental cost-effectiveness ratios within each patient group is achieved. The programmes should then be ordered in terms of their incremental cost-effectiveness ratios; note that if there is only one independent programme in one patient group, the average and incremental cost-effectiveness ratios will coincide.

With a fixed budget as the decision rule, we start by considering the programme with the lowest cost-effectiveness ratio and then switch patients to increasingly more effective mutually exclusive programmes and add independent programmes as the budget increases. If the price is used as the decision rule, the mutually exclusive programme within each patient group with the highest incremental cost-effectiveness ratio at or below the price should be implemented (if there is only one programme in a patient group, this should then be implemented if the cost-effectiveness ratio is at or below the price).

To illustrate this we can combine the programmes used in the examples above for independent and mutually exclusive programmes. The incremental cost-effectiveness ratios of these programmes are shown in Table 7. In the table there are programmes from four different patient groups, patient group 1, patient group 2, patient group 3 and patient group 4 and there are assumed to be 100 patients in each group. For patient group 4 there are also the three mutually exclusive programmes 4A, 4B, and 4D (the dominated programme 4C has been excluded). We refer to all the cost-effectiveness ratios in Table 7 as the incremental cost-effectiveness ratios,

although since there is only one programme each in patient groups 1-3, the incremental cost-effectiveness ratios for these programmes will coincide with the average cost-effectiveness ratios of the programmes.

Table 7. Example of both independent and mutually exclusive programmes

Programme	Incremental C/E
4A	5000
1	10000
4B	15000
2	20000
3	30000
4D	40000

In Table 7 the six programmes have been ordered in terms of their incremental cost-effectiveness ratios. As for the case with only mutually exclusive programmes, the ordering of the incremental cost-effectiveness ratios in Table 7 cannot be interpreted as a ranking. Which of the mutually exclusive programmes should be implemented depends on the budget or the price that we are willing to pay per life-year gained.

Programmes that are independent can be ranked, e.g. programme 1 should always be given priority over programme 2. However, as before it is impossible to rank the mutually exclusive programmes. Which of the mutually exclusive programmes should be implemented depends on the budget or the price that we are willing to pay per life-year gained. This has to be borne in mind when interpreting so-called "league tables" with cost-effectiveness ratios (Williams 1985; Schulman et al 1991).

These "league tables" do not provide a ranking of different programmes in terms of priority unless they consist of only independent programmes. If they consist of only independent programmes they may not be very useful anyway, since this would imply that there is only one treatment choice for each patient group (or that one treatment choice dominates all the other available treatment alternatives for a patient group). To interpret cost-effectiveness ratios we need information about either the budget or the price for the effectiveness unit.

According to the decision rule of the fixed budget, we start by implementing programme 4A and if the budget is $50,000 all patients in patient group 4 will receive treatment 4A. If the budget exceeds $50,000 we start implementing programme 1 also, and if the budget is $150,000 all patients in group 1 will receive treatment. If the budget exceeds $150,000 we start switching patients from treatment 4A to treatment 4B and if the budget is $300,000 all patients in patient group 4 will receive treatment 4B.

If the budget exceeds $300,000 we also start treating patients in group 2, and if the budget is $700,000 all the patients in group 2 will receive treatment. If the budget exceeds $700,000 we can also start treating patients in group 3, and if the budget is $1,600,000 all the patients in group 3 will receive treatment. If the budget exceeds $1,600,000 we start switching patients from treatment 4B to treatment 4D and if the budget is $2,400,000 or greater all the patients in patient group 4 will receive treatment 4D.

Table 8 shows the budgets that lead to all patients in each patient group receiving the same treatment. Note that if the budget is below $2,400,000 but not equal to any of the amounts in Table 8, not all the patients in each patient group will receive the same treatment. As before, the different budgets also imply different prices per gained life-year on the margin, e.g. a budget of $300,000 implies that we are willing to pay $15,000 per gained life-year on the margin (i.e. the incremental cost-effectiveness ratio of programme 4B, which is the programme with the highest incremental cost-effectiveness ratio that is implemented with a budget of $300,000).

Table 8. The choice of programmes for different sizes of the budget

Budget	Choice of Programmes
50000	4A
150000	4A+1
300000	4B+1
700000	4B+1+2
1600000	4B+1+2+3
2400000	4D+1+2+3

If we instead use the price per gained life-year as the decision rule, we should implement the programme with the highest incremental cost-effectiveness ratio which is at or below that price in each patient group (i.e. each "independent" group). In the example, if the price is below $5,000 no programme will be implemented, with a price between $5,000 and $9,999 programme 4A will be implemented, with a price between $10,000 and $14,999 programme 4A and programme 1 will be implemented, with a price between $15,000 and $19,999 programme 4B and programme 1 will be implemented, with a price between $20,000 and $29,999 programme 4B, programme 1 and programme 2 will be implemented, with a price between $30,000 and $39,999 programme 4B, programme 1, programme 2 and programme 3 will be implemented, and with a price of $40,000 or higher programme 4D, programme 1, programme 2 and programme 3 will be implemented.

Note also that the different prices will implicitly yield a budget, e.g. if a price of $15,000 is used this will yield a budget of $300,000 (i.e. programme 4B and programme 1 will be implemented with this price, and the total cost of programme 4B

is $200,000 and the total cost of programme 1 is $100,000, yielding a total cost for both programmes of $300,000).

Table 9. The choice of programmes for different prices

Price/cut-off	Choice of Programmes
<5000	NO PROGRAMME
5000- 9999	4A
10000-14999	4A+1
15000-19999	4B+1
20000-29999	4B+1+2
30000-39999	4B+1+2+3
40000-	4D+1+2+3

As in the previous examples, the use of the budget to maximise life-years gained is based on the assumption of constant returns to scale within the cost increase of each incremental programme. Using the price per gained life-year is not based on this assumption in the same way. If the price per gained life-year is used, programmes will never be divided, i.e. either the whole programme or none of the programme is carried out. This means that the health effects will always be maximised for the amount of money that is spent on the health programmes, given that all programmes have to be implemented fully or not at all. It is possible, however, that if the programmes are defined differently, e.g. divided into different scales, it would be possible to get more health effects for the same amount of money spent (i.e. the same budget). This could be the case if the last programme implemented (i.e. the programme with the highest incremental cost-effectiveness ratio) exhibits decreasing returns to scale, but then the different scales should in principle be defined as different mutually exclusive programmes. This means that the price can be used as the decision rule to maximise the health effects for a given amount of resources, without any assumptions about constant returns to scale. This applies as long as the scales of programmes where constant returns to scale do not hold are defined as mutually exclusive programmes.

The most important lesson from this section, about the decision rules of cost-effectiveness analysis, is the importance of estimating incremental cost-effectiveness ratios of mutually exclusive alternatives. It also has to be clear that these incremental cost-effectiveness ratios do not rank the different mutually exclusive programmes, and without information about the size of the budget or the price per effectiveness unit, the ratios as such provides no guidance about which of a number of mutually exclusive programmes should be chosen.

The fallacy of not estimating cost-effectiveness ratios correctly can be illustrated by a cost-effectiveness analysis done by Sintonen & Allander (1990). In this study, alternative drug treatments for the treatment of ulcers in patients with a confirmed

diagnosis was compared in terms of cost-effectiveness. A decision-tree model was used to estimate the cost-effectiveness. The model was based on a 6-month time period and the key probabilities in the model were the probability of healing with different treatments and the probability of having a relapse after discontinuation of treatment. Logistic regression analysis of the results of different clinical trials was used to estimate the probability of healing as a function of treatment time and the probability of relapse as a function of time after discontinuation of treatment.

The effectiveness unit used in the cost-effectiveness analysis was the number of healthy days during the 6-month period (defined as the number of ulcer-free days). The costs that were included in the study were the health care costs of physician visits, drugs and surgery. The cost of transportation was also included, as well as the loss of leisure time and working time for travelling, physician visits and surgery.

The study analysed 5 different treatment alternatives for ulcers, which were compared with a "base-case" alternative. The five treatment alternatives were omeprazole, ranitidine I and II, and sucralfate I and II (two alternatives were included for ranitidine and sucralfate depending on what treatment was used in resistant cases). The base-case alternative that these five treatments were compared to was defined as no healthy days and no costs for surgery, drugs, physician visits or travelling. The base-case alternative is one of the problems with the study, since no alternative exists which gives no healthy days and no costs. It would have been better to compare the drugs with surgery, for instance. The costs and effects as estimated in the study for the different alternatives are shown in Table 10.

Table 10. Costs and effects of five alternative treatments for ulcers

Treatment alternative	C	E	C/E
Sucralfate II	627.8	86.0	7.3
Ranitidine II	887.9	87.2	10.2
Ranitidine I	891.3	88.2	10.1
Sucralfate I	602.8	89.6	6.7
Omeprazole	785.8	121.5	6.5

Source: Sintonen & Allander (1990).

The costs and effects shown in Table 10 are the costs and effects compared to the base-line alternative and the average cost-effectiveness ratios. Although it would have been appropriate since the alternatives are mutually exclusive, no incremental cost-effectiveness ratios were estimated in the study. Instead the authors used the results presented in Table 10 to conclude that the three alternatives (sucralfate I and II and omeprazole) were equally cost-effective, since the average cost-effectiveness ratios were on the same level.

The conclusion drawn is not appropriate, however, as can be seen if the incremental cost-effectivenesses of the alternatives are estimated. Inspection of Table 10 shows that sucralfate II, ranitidine I and ranitidine II can be excluded from further consideration straight away, since they lead to both higher costs and less effects than sucralfate I. Table 11 shows the incremental cost-effectiveness ratios of the two remaining alternatives.

Table 11. Incremental cost-effectiveness ratios of sucralfate I and omeprazole

Treatment alternative	C	E	Incr. C/E
Sucralfate I	602.8	89.6	6.7
Omeprazole	785.8	121.5	5.7

First it is important to note that the choice between the two alternatives is not trivial for the patients, since the effectiveness of omeprazole is much greater than that of sucralfate I. The incremental cost-effectiveness ratio of sucralfate I compared to the base-case alternative is FIM 6.7 and the incremental cost-effectiveness of omeprazole compared to sucralfate I is FIM 5.7. Since the incremental cost-effectiveness ratio is decreasing for omeprazole, this means that sucralfate I should be excluded since it is dominated. This would mean that the incremental cost-effectiveness ratio of omeprazole should be re-calculated, and this gives FIM 6.5 (omeprazole compared to the base-case alternative, since this is the only alternative left after excluding the dominated alternatives).

Since the base-case alternative used was invalid, however, it is impossible to know whether omeprazole dominates sucralfate I or not. The incremental cost-effectiveness of sucralfate I compared to the base-case alternative depends on the costs and effects of the base-case alternative. A valid base-case alternative would be the treatment of ulcers if ulcer drugs are not used, i.e. probably surgery.

In Table 12 we have carried out a re-calculation of the incremental cost-effectiveness ratios, assuming that the costs are FIM 500 and the effects are 50 of the base-case alternative. This re-calculation is purely hypothetical, and has been done to show how the conclusion about whether or not sucralfate I is dominated is affected by the costs and effects of the base-case alternative.

Table 12. Hypothetical re-calculation of the incremental cost-effectiveness ratios

Treatment alternative	C	E	Incr. C/E
Sucralfate I	102.8	39.6	2.6
Omeprazole	285.8	71.5	5.7

In the hypothetical re-calculation of the incremental cost-effectiveness ratios in Table 12, the incremental cost-effectiveness ratio of sucralfate I is FIM 2.6 and the incremental cost-effectiveness ratio of omeprazole is FIM 5.7. In this case sucralfate I is not dominated, which shows that the invalid base-case alternative used in the study makes it impossible to determine whether sucralfate I is dominated or not.

The conclusion that can be drawn from the study is then that omeprazole is the most cost-effective of the drug alternatives if the price that we are willing to pay for a healthy day (i.e. an ulcer-free day, the effectiveness unit used in the study) is equal to or exceeds FIM 5.7. In this case omeprazole is always to be preferred to sucralfate I and the three other alternatives can be ruled out, since they had less effectiveness and greater costs than sucralfate I. However, since the base-case alternative was invalid it is not possible to conclude that omeprazole is more cost-effective than the "base-line" alternative (i.e. probably surgery), even if the price exceeds or equals FIM 5.7. It is essential to estimate the incremental cost-effectiveness ratios in order to reach the only valid conclusion that can be drawn from the study, that omeprazole is the preferred drug choice of the five alternatives analysed if the price per healthy day exceeds or equals FIM 5.7.

9.5 The Choice Of Effectiveness Measure

One important issue in cost-effectiveness analysis is obviously the choice of effectiveness measure, i.e. the units that we try to maximise with a given budget. Some studies use intermediate effectiveness measures such as the mm Hg blood pressure reduction (Logan et al 1981). One problem with such measures is that it enables only very limited comparisons of cost-effectiveness, e.g. if the mm Hg blood pressure reduction is used it is only possible to compare alternative ways of lowering blood pressure.

It is also extremely difficult to interpret the results of a cost-effectiveness analysis using intermediate effectiveness measures. For example, what does a specific cost per mm Hg reduction in blood pressure mean? In order to use intermediate effectiveness measures, they must also be clearly related to some health outcome, e.g. the risk of heart attacks, since it seems to be a dubious goal to maximise for instance the blood pressure reduction per se. Furthermore, every unit of the intermediate effectiveness measure should be associated with the same change in the health outcome of interest, e.g. a percentage unit reduction in the risk of heart attacks. If such a relation is known, however, it seems to be better to use the actual health outcome as the effectiveness unit. Thus the use of intermediate effectiveness measures cannot be recommended.

The next type of effectiveness unit consists of different types of event, e.g. the number of heart attacks prevented or the number of ulcers prevented (Jönsson & Haglund 1992). With such an effectiveness measure it is possible to compare different

alternative ways of reducing different events, e.g. alternative ways of preventing heart attacks.

One type of event that is sometimes used as an effectiveness measure is lives saved (Morrall 1986). By using lives saved as an effectiveness measure it is possible to compare all the different activities that reduce the risk of mortality. It is not obvious, however, how the number of lives saved should be defined, since lives are not actually saved but merely extended. The most common definition would seem to be the change in the conditional probability of surviving one period, e.g. a year (the conditional survival probability of a specific year is the probability of surviving during that year, if alive at the beginning of the year). This would mean that a life could be saved several times if the conditional survival probability changed in more than one period.

At least in the medical field, it has been common to use the number of life-years gained rather than lives saved as the effectiveness measure for programmes that reduce the mortality risk (Edelson et al 1990). The advantage of this measure is that it takes into account the number of life-years at risk when a mortality risk is reduced. It also has a clear-cut definition as the change in the sum of the cumulative survival probabilities for different years (the cumulative survival probability is the probability of being alive in different future years, viewed from some starting year when the individual is alive).

One problem with life-years gained is that it does not incorporate quality of life. To enable comparisons of programmes that affect both the quality and the quantity of life, one single effectiveness measure is needed that incorporates changes in both the quantity and the quality of life. One such measure is quality-adjusted life-years (QALYs) gained (Boyle et al 1983). This type of effectiveness measure is sometimes referred to as a utility measure. When QALYs or some other utility measure are used as the effectiveness measure in cost-effectiveness analysis, the analysis is often referred to as cost-utility analysis. Cost-utility analysis is dealt with in Chapter 10 and will not be discussed further here.

9.6 Cost-Effectiveness Analysis VS Cost-Benefit Analysis

It is important to investigate the relationship between cost-effectiveness analysis and cost-benefit analysis. The argument behind cost-effectiveness analysis has often been a so-called decision-maker approach to economic evaluation. According to this approach, the aim of economic evaluation is to maximise whatever the decision-maker wants to maximise (Sugden & Williams 1978), and to only include the costs and benefits that the decision-maker finds relevant.

This approach can also be tied to the decision rules of cost-effectiveness analysis above, in the sense that only those costs that fall on the budget of a specific decision-maker will be included. In practice in the health care field, this has often led

to only health care costs being included in a cost-effectiveness analysis, with the argument that the health care budget should be used to maximise health (somehow defined) (Weinstein & Stason 1976; Williams 1985; Edelson et al 1990).

The decision-maker approach to economic evaluation and cost-effectiveness analysis can be criticized on many grounds. Firstly, the approach lacks theoretical foundation in welfare economics and seems to be ad hoc. Often the studies are also inconsistent, since there is no investigation into what any real decision-makers want to maximise or include in the analysis. It is usually merely assumed that the decision-maker wants to maximise life-years or QALYs or some other outcome measure using the health care costs as the budget.

As mentioned above, the most common approach to cost-effectiveness analysis is to include only health care costs, with the argument that only costs that fall on the health care budget should be included. However, there is usually more than one budget and more than one decision-maker even within the health care sector. Take the US for instance, where no single health care budget can be identified. In order to be consistent when the health care budget approach is used, this would mean that all the user charges paid by the patients should be deducted from the health care costs (e.g. with user charges, part of the health care costs will be paid by the patients and these costs should not be included, since they do not fall on the health care budget). To our knowledge this has not been done in any study based on this approach.

Furthermore, since budgets are usually set in annual terms, a strict adherence to the approach would mean that all future changes in health care costs should be ignored (i.e. those costs that do not appear during the year a programme is implemented). This is so since the current annual budget should be used to maximise "health", and the current budget is not affected by future changes in health care costs. However, this method is not usually pursued in the studies based on this approach either.

Using the budget approach would also mean using the prices faced by the budget holder (e.g. charges) rather than the appropriate opportunity costs. If a strict budget maximisation was adhered to, this would be likely to lead to major problems of suboptimisation. From a societal perspective there seems to be no reason to encourage such an approach to economic evaluation. We will therefore not pursue this approach any further here, but will just note that the approach may be of interest to a specific health care provider or an HMO but that this book is concerned with economic evaluation from the societal perspective.

Instead we feel that it is important to investigate the relationship between cost-effectiveness analysis and cost-benefit analysis, given that the aim of both types of analysis is to include all costs and health effects (benefits) irrespective of the person or persons to whom they accrue. Cost-effectiveness analysis can then be viewed as a subset of cost-benefit analysis, where the aim of cost-effectiveness analysis is to estimate the cost function for producing health effects. In order to reach a decision

based on cost-effectiveness analysis, information is then needed about the WTP for the health effects (i.e. the price for the health effects).

In the standard cost-effectiveness case this would mean that the difference between cost-benefit analysis and cost-effectiveness analysis is that in cost-effectiveness analysis, the WTP for a health change is assumed to be the same for all individuals and for all sizes of the change in health.

We will start by assuming that this assumption is valid, i.e. that the WTP for health changes is the same for all individuals and is constant for all sizes of the change in health. The implications of relaxing this assumption for the distinction between cost-effectiveness analysis and cost-benefit analysis will also be explored below.

In Figure 1 we show the marginal cost curve of producing gained life-years (which are assumed to be the measure of health effects in the remainder of this chapter), based on the programmes and the incremental cost-effectiveness ratios in Table 7 from the section about the decision rules. We assume that the six health care programmes in Table 7 are the health programmes available for the population (plus the programme 4C which was excluded from consideration since it was dominated).

It is also assumed that the costs of the programmes include all costs that are not incorporated in the WTP for the health effects. The marginal cost curve is drawn by adding new independent programmes in the same order as the incremental cost-effectiveness ratios, or replacing mutually exclusive programmes with more effective programmes. The function is stepwise because the adding or replacing of programmes leads to an increase in the marginal cost. The marginal cost of producing more life-years is also the same as the incremental cost-effectiveness ratios. For each new programme the marginal cost is constant, because of the assumption of constant returns to scale.

The marginal WTP curve for the population has also been added to the figure, and since we assumed that the marginal WTP is constant the marginal benefit curve is just a straight horizontal line. In the figure it is assumed that the marginal WTP per gained life-year is $30,000. The optimum for the authorities is the point where the marginal cost equals the marginal benefit, i.e. the authorities will continue to implement programmes until the marginal cost equals the marginal benefit.

In the figure this leads to the production of 80 life-years, i.e. programmes 4B+1+2+3 are implemented. In this case cost-effectiveness analysis will lead to the same result as cost-benefit analysis, since according to cost-benefit analysis programmes should be implemented as long as benefits exceed or equal costs. Furthermore, it is not necessary to assume constant returns to scale for this result.

If the programmes as defined in Figure 1 are evaluated using cost-effectiveness analysis with a price per gained life-year of $30,000, this will yield the same result as a

cost-benefit analysis of the same programmes. Using cost-effectiveness analysis in this case will also maximise the number of life-years gained for the resources that will be spent, given the available information on costs and effectiveness (and the price per gained life-year).

It is possible of course that if the condition of constant returns to scale does not hold, we could find an even better allocation of resources by varying the scales of the programmes. To find the optimal scale of a programme is, however, a problem that is identical for both cost-effectiveness and cost-benefit analysis, i.e. often discrete changes are considered. In principle it is possible to calculate the incremental cost-effectiveness ratios for different scales of the programmes, which could possibly lead to a smoother marginal cost curve in Figure 1.

Note also that in this case cost-benefit analysis will also yield the same result when it comes to excluding dominated alternatives. If the marginal (incremental) cost per gained life-year decreases for a more effective programme, it is obvious that this programme is always to be preferred to the next most effective programme if the price per gained life-year is constant. For instance, in the example in the section about the decision rules, the incremental cost-effectiveness for programme 4D compared to 4C was $10,000, and this ratio was less than the incremental cost-effectiveness ratio of 4C compared to 4B, which was $70,000 (4C was thus dominated by 4D).

If we are willing to pay $70,000 per gained life-year (which would be needed to implement programme 4C), it is obvious that we will also be willing to pay for the more effective programme with the decreasing incremental cost-effectiveness ratio (i.e. programme 4D). The dominated programme will thus never be implemented (programme 4C in this example) regardless of the price per gained life-year, since if the price is high enough to consider the dominated programme there is always another more effective mutually exclusive programme that leads to more net benefits (programme 4D in this example).

Thus if the WTP per health effect is constant and is the same for all individuals under all circumstances, cost-effectiveness and cost-benefit analysis will yield the same result, i.e. both methods will recommend the implementation of the same health care programmes, provided that the WTP per gained life-year is used as the price in cost-effectiveness analysis. This means, however, that for cost-effectiveness analysis to be a useful tool we need information about the WTP per unit of health effects, e.g. the WTP per gained life-year, in order to provide the method with a useful decision rule.

The assumption of constant WTP per gained life-year (or some other measure of health effects) does not seem very realistic, and it is therefore important to investigate the relationship between cost-effectiveness analysis and cost-benefit analysis if this assumption is relaxed. We will therefore consider the case where the WTP per gained life-year varies, but cost-effectiveness analysis is based on a constant WTP (which is implied by the maximization of health effects).

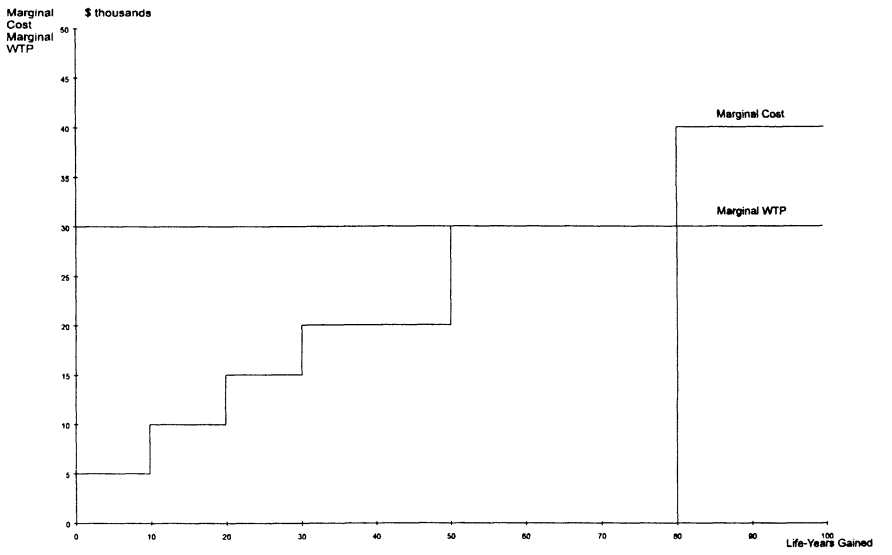

Figure 1. The marginal cost and marginal WTP of producing life-years gained for a society with a constant marginal WTP

We can start by considering a single individual. The marginal cost curve for producing gained life-years for the individual and the marginal WTP for gained life-years for the individual are shown in Figure 2. The marginal cost curve can be interpreted as a number of increasingly more effective mutually exclusive programmes for the individual with increasing incremental cost-effectiveness ratios The curve is drawn smoothly, on the assumption that every programme leads to a marginal cost increase.

The optimum for the individual is the point where the marginal WTP for gained life-years equals the marginal cost for gained life-years (denoted P* in the figure). The marginal cost curve shows the marginal (incremental) cost-effectiveness ratio of different health care programmes, and if we use P* as the price per gained life-year we could implement increasingly more effective health care programmes until the marginal cost-effectiveness ratio equalled this price. For the individual, cost-effectiveness analysis based on the price P* will thus yield the same result as using cost-benefit analysis (cost-benefit analysis here being defined as the point where the marginal cost and marginal WTP curves intersect, i.e. the optimum perceived by the individual on a market).

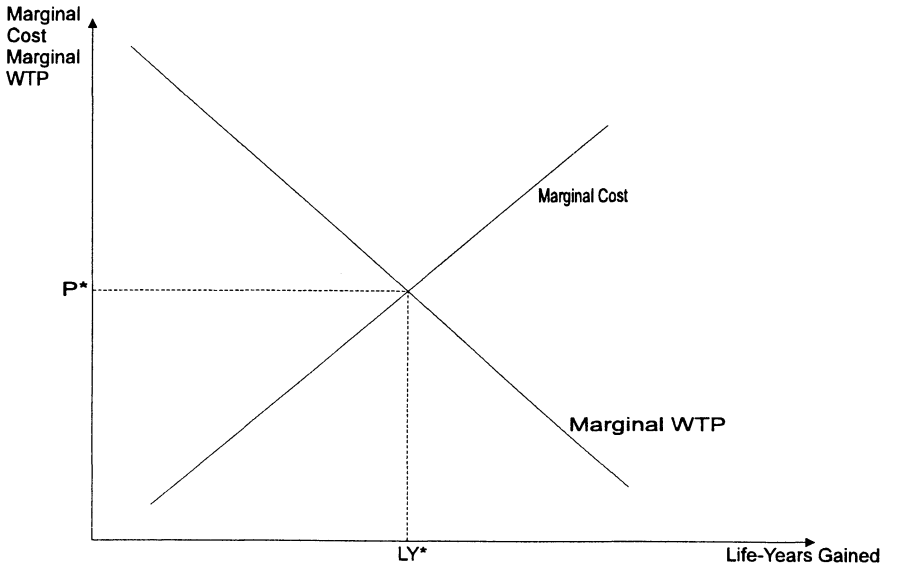

Figure 2, The marginal cost and marginal WTP of producing life-years gained for an individual

For non-marginal cost changes due to a programme, which would typically be the case, this would not necessarily be true, however. If for instance a programme does not lead to a marginal cost increase and the marginal cost curve is horizontal for the increase in cost, representing constant returns to scale (e.g. see Figure 1), and the marginal WTP curve intersects the marginal cost curve at such a segment, a cost-effectiveness analysis based on P* would recommend fully implementing the programme and a cost-benefit analysis would recommend implementing only part of the programme, i.e. implement the programme until the marginal WTP equals the marginal cost.

Similarly, it is no longer the case that a dominated alternative will not be chosen. One case of domination is when a new technology is introduced so that the same amount of life-years can be produced at a smaller cost. In such a case the dominated alternative would obviously never be chosen by the individual. However, if the domination is due to increasing returns to scale for a specific technology, so that it is not possible to produce the same amount of life-years at a smaller cost, this is not necessarily the case.

If the WTP per gained life-year exceeds the incremental cost of the dominated alternative per gained life-year, but the WTP per gained life-year falls short of the

incremental cost-effectiveness of the more effective programme compared to the dominated programme, the dominated programme will be the choice in a cost-benefit analysis.

This can be illustrated by an example. Assume that a drug can be given in two doses; the first dose (dose 1) produces 1 life-year at a cost of $50,000 and the second dose (dose 2) produces 2 life-years at a cost of $90,000. The incremental cost-effectiveness ratio of dose 1 is $50,000 (50,000/1) and the incremental cost-effectiveness ratio of dose 2 is $40,000 ([90,000-50,000]/[2-1]). Dose 1 is then dominated by dose 2, due to increasing returns to scale of the dose, and dose 1 would be excluded from further consideration in a cost-effectiveness analysis. If the individual is willing to pay $50,000 for 1 gained life-year and $80,000 for two gained life-years (i.e. the WTP decreases with the number of gained life-years), a cost-benefit analysis would show that the dominated programme (dose 1) should be carried out.

We can then consider the case with more than one individual; this is obviously where the assumption of a constant WTP per gained life-year in cost-effectiveness analysis produces most problems. It is now straightforward to show that cost-benefit analysis and cost-effectiveness analysis can produce dissimilar results. Even if we assume that we use the mean marginal WTP per gained life-year for further health improvements for each individual as the price in cost-effectiveness analysis, there is no reason to believe that the WTP will be the same for different individuals.

Say that we are considering implementing two different independent health programmes. One programme in one patient group costs $100,000 and yields 5 gained life-years, i.e. an incremental cost-effectiveness ratio of $20,000. Another programme in another patient group costs $100,000 and yields 4 gained life-years, i.e. an incremental cost-effectiveness ratio of $25,000. The price that we use in our cost-effectiveness analysis per gained life-year is $22,500.

According to the cost-effectiveness analysis the first programme should be implemented, but not the second. However, if for instance the marginal WTP per gained life-year is $15,000 in the first patient group and $30,000 in the second patient group, a cost-benefit analysis would give the result that the second programme should be implemented rather than the first.

The marginal WTP per gained life-year could vary in different patient groups, e.g. depending on age. Another important source of variation is that the marginal WTP per gained life-year can be assumed to vary, depending on the size of the health improvement. If cost-effectiveness analysis is used to maximise the health effects, no distinction will be made between the size of the health improvements for different groups, e.g. an increase of ten gained life-years for one individual will be valued the same as an increase in one life-year each for ten individuals. Cost-benefit analysis, on the other hand, could be expected to give less weight to an increase in 10 life-years for

one individual than one life-year each for 10 individuals, due to decreasing marginal WTP with the size of the health improvement.

As before, in this case with many individuals cost-benefit analysis could in principle give the result that a dominated alternative should be chosen. This could happen even if the more effective programme can be reduced in scale by reducing the number of patients, so that more health effects could be produced at a lower cost. This is so because the distribution of health effects between different individuals could differ in the two situations.

Consider for instance a case where two mutually exclusive treatment alternatives are available for a patient group of 100 individuals. Treatment 1A costs $3,000 per patient and yields 0.1 gained life-years per patient. Treatment 1B costs $5,000 per patient and yields 0.2 gained life-years per patient. The incremental cost-effectiveness ratio of treatment 1A is $30,000 (3,000/0.1) and the incremental cost-effectiveness ratio of treatment 1B is $20,000 ([5,000-3,000]/[0.2-0.1]). According to a cost-effectiveness analysis treatment 1A is dominated by treatment 1B and treatment 1A should be excluded from further consideration.

Assume further that treatment 1B exhibits constant returns to scale, so that the treatment can be reduced to the cost of treatment 1A and will yield more gained life-years by reducing the number of patients treated. With treatment 1A, 100 individuals would be treated at a total cost of $300,000 (3,000*100), yielding a total number of 10 gained life-years (0.1*100). If treatment 1B is reduced to this size, 60 individuals (300,000/5,000) can be treated at a total cost of $300,000, yielding 12 gained life-years (0.2*60). Assume now that the WTP per patient is $30,000 for one gained life-year and $45,000 for two gained life-years. According to a cost-benefit analysis, treatment 1A should then be carried out since the total benefits of $300,000 (100*30,000) equal the total cost of $300,000. For treatment 1B the total benefits for the same cost are only $270,000 (60*45,000) (for the full scale of treatment 1B the total cost would be $500,000 (5,000*100) and the total benefits would be $450,000 (4,500*100), thus also resulting in a loss).

Here the problem arises because even though treatment 1B can produce more gained life-years than treatment 1A at the same cost, the distribution of the gained life-years is more uneven between the individuals (note that if treatment 1B could be reduced so that each individual gained 0.12 life-years at a cost of $300,000, this would have been preferred).

It is thus clear that cost-benefit analysis and cost-effectiveness analysis based on a constant WTP per gained life-year can produce different results. How important these differences would be in practice is an empirical issue. In our view, cost-effectiveness analysis may best be interpreted as a subset of a cost-benefit analysis where the aim of the cost-effectiveness analysis is to estimate the cost function of producing health effects.

By combining the estimate of the cost function with the WTP per unit of health effects, a cost-benefit analysis can then be carried out. The WTP could then be allowed to vary depending on for instance the age of the recipients of the health care programme and the size of the health improvement for each individual (e.g. 0.01 gained life-years per person versus 1 gained life-years per person).

When estimating the cost function of producing health effects using the principles of cost-effectiveness analysis, the exclusion of dominated alternatives can be a potential problem. This is because it is possible, as shown above, that an alternative which is excluded as dominated in a cost-effectiveness analysis could be the preferred alternative in a cost-benefit analysis.

This problem is related to the problem of increasing returns to scale on a market that creates market failure. By excluding dominated alternatives, the cost function for producing health effects will always be horizontal or upward sloping and cannot have any downward sloping segments. The problem that can arise here is that a programme which dominates another one yields more health effects, and if the marginal WTP decreases with the size of the health change then the marginal WTP for these additional health effects will be lower than the WTP per health effect for the total change in health effects.

Thus if a dominated alternative is excluded and it has an incremental cost-effectiveness ratio close to the price cut-off point, it may be advisable to compare the incremental cost-effectiveness ratio of the more effective programme with the marginal WTP for that change in health effects, before the dominated alternative is excluded. In a sense an additional test should be added, to determine which mutually exclusive alternative should be carried out. For the alternative that should be carried out first, the WTP per unit of health effects for the incremental change in health effects, compared to the next most effective programme that has not been excluded as dominated, should exceed the incremental cost-effectiveness ratio. In addition, the WTP for the health change compared to an excluded dominated alternative should exceed the incremental cost-effectiveness ratio compared to the dominated alternative.

The cost-effectiveness approach also hinges upon the assumption that health effects can be estimated in some meaningful way, so that more of the effects are preferred to less (i.e. so that the health effects are consistent with individual preferences). This will be discussed further in the chapter below about cost-utility analysis. If this is not the case, the whole cost-effectiveness approach seems to be meaningless, and it would be better to do cost-benefit analysis directly on different health care programmes (i.e. to investigate the WTP for programmes as such rather than the WTP for health effects, to be used as a decision rule in cost-effectiveness analysis).

Revealed preference and expressed preference approaches could be used to try and estimate the WTP for different effectiveness units, to be used as the price in

cost-effectiveness estimations. One example of this is the common presentation of the results of WTP studies of mortality risk reductions in the terms of value per statistical life. The intention here is to compare the value per statistical life with the cost of saving statistical lives through different health programmes, i.e. the type of approach to cost-effectiveness analysis advocated in this book.

However, it should be realised that it may be appropriate for the value of statistical life to vary, depending on the population and the health programme. At least for studies in the health care sector, it seems more useful to express the results of WTP studies in terms of the value of gained life-years, since age is an important factor in health care programmes. More emprical studies are needed to determine whether the value per statistical life and gained life-year varies systematically, in order to take this variation into account in economic evaluations based on cost-effectiveness analysis.

The implication of this approach is also that all costs which are not included in the effectiveness unit should be included in a cost-effectiveness analysis. To some degree this depends on how the effectiveness units are defined and how the WTP for the effectiveness units is assessed. If for instance the number of hip fractures avoided is used as a measure of effectiveness, it is unclear whether or not this also includes the morbidity costs due to hip fractures. In turn this would depend on whether or not the WTP per avoided hip fracture that is used as the price includes these cost changes.

In a sense, the choice of costs to be included in the cost-effectiveness estimation determines the price with which it should be compared (e.g. the price per avoided hip fracture inclusive or exclusive of the changes in morbidity costs). The issues in the estimation of costs in a cost-effectiveness analysis are thus the same as for the cost concept of cost-benefit analysis, and the estimation of costs in cost-benefit analysis described above in chapters 4 and 7. The costs that should be included in a cost-effectiveness analysis will also be discussed further in the section below about cost-utility analysis.

It is also the case that if we do not follow the budget principle of only including the costs that fall on a specific budget and then try to maximise effectiveness given only this budget, we have to be careful in how we interpret the cost-effectiveness ratios. It is a well known problem in cost-benefit analysis that benefit-cost ratios cannot be used to rank projects; they can only be used to see whether or not benefits exceed costs (i.e. whether or not the ratio is higher than 1; see the section about cost-benefit analysis in Chapter 2).

The ratio is affected by whether for instance reduced costs are entered as a cost saving or as a benefit (for an example of this, see the section about cost-benefit analysis in Chapter 2). This issue becomes more relevant for the cost-utility case below, where there is more scope for whether items should be in the nominator or the denominator. This means that cost-effectiveness ratios should not be interpreted as a ranking of projects even for independent programmes, but they should only be used to determine

whether or not benefits exceed costs, i.e. whether the cost-effectiveness ratio is at or below the price per effectiveness unit.

This is also all that is needed if we do not face a budget constraint, since all independent programmes with positive net benefits should be carried out according to the decision rule (and the mutually exclusive programme with the highest net benefits should be carried out). If it is desirable to get an indication of the size of the net benefits for independent programmes, the net benefits of the whole programme can be estimated by using the price per effectiveness unit (in principle it then has to be assumed that the benefit per effectiveness unit is constant over the range of the programme; thus the price that should be used is in principle the average WTP per effectiveness unit for the change in effectiveness due to the programme).

It is only relevant to rank independent programmes with positive net benefits (or also programmes with negative net benefits) if there is a real budget constraint that has to be taken into account. If that is the case, however, then rather than suboptimising by only including the costs that fall into this budget in a cost-effectiveness analysis and dividing this by the health effects, it seems better to try and maximise the net benefits given this budget constraint. This means that all costs and benefits (including the health effects transformed to monetary terms using the WTP for health effects) which do not fall on this budget should be expressed as benefits in monetary terms. The cost-benefit ratios of different programmes can then be estimated and used to maximise the net benefits, in the same way as shown in the section about the decision rules of cost-effectiveness analysis (Baumol 1972).

Alternatively, if the assumption of constant returns to scale does not hold, non-linear programming techniques could be used to maximise net benefits (Winston 1991). The difference with this approach compared to the cost-effectiveness approach above is that it takes into account all costs and benefits and does not lead to suboptimisation.

In a sense, if a specific budget imposes a constraint, this can be viewed as meaning that the dollars from that budget have a higher cost than other dollars, and consequently that the shadow price of dollars from this budget should be set higher (or lower if it is a budget with few profitable alternative uses). It may therefore be better to try and estimate these shadow prices and incorporate them directly in the analysis, and then investigate whether or not benefits exceed costs. In practice it will probably also be difficult to use this kind of budget optimisation, since it implies knowing all the costs and benefits of the programmes of interest for a specific budget (and all their scale effects if this assumption is relaxed and non-linear programming employed).

9.7 Discounting In Cost-Effectiveness Analysis

One important issue in cost-effectiveness analysis is the discounting of costs and effects. The discounting of costs is usually carried out in the same way as for

cost-benefit analysis (see Chapter 8 for the discounting of monetary costs and benefits). It should be noted, however, that if the budget maximisation principle is strictly adhered to, costs never have to be discounted in a cost-effectiveness analysis, since only the costs in the annual budget for the year of the programme enter into the cost estimation. However, as stated above, this approach is not recommended.

The discounting of monetary costs is relatively uncontroversial in cost-effectiveness analysis, but the discounting of health effects such as gained life-years is more controversial. Since the health effects are not expressed in monetary terms it is not clear whether they should be discounted or not. It is not possible to apply the investment argument for discounting for health effects in the same way as for monetary costs and benefits, e.g. it is not possible to give up life-years now in order to gain more life-years in the future.

To some degree it is, however, possible to give up some health now in order to have more health in the future, e.g. to accept some side-effects of a treatment for a future health gain. It is unclear to what extent the time preference of individuals for health effects will coincide with the real discount rate. If the aim of cost-effectiveness analysis is to maximise the health gains, i.e. the price per health effect is set equal for everyone, then whether the effects should be discounted or not depends on whether discounted health effects or undiscounted health effects should be maximised, e.g. whether gained life-years or discounted gained life-years should be maximised.

From a cost-benefit viewpoint the WTP of the health gains should be discounted at the same rate as for costs. Discounting health effects at the same rate as for costs, which seems to be the most commom approach in cost-effectiveness applications, then implies that the WTP for health effects is the same in the future as at present. This is not necessarily the case, however, since the WTP for health effects for instance can depend on the age of the individual.

One common argument in the literature in favour of discounting health effects at the same rate as for costs is the so called postponement argument (Keeler & Cretin 1983). According to this argument, non-monetary effects should be discounted at the same rate as for costs; otherwise it will always be profitable to postpone a project indefinitely into the future. The reasoning behind the argument can be illustrated by a very simple example. Assume that a programme yields 10 gained life-years this year and costs $100,000 (incurred at the beginning of the year so that they do not have to be discounted), thus giving a cost-effectiveness ratio of $10,000 (100,000/10).

Assume that the identical programme can be carried out next year at the same cost and effectiveness. If the cost-effectiveness ratio of this programme is calculated in present value terms with a 5% discount rate for costs and no discounting of health effects, it will yield a cost-effectiveness ratio of $9,500 ([100,000*0.95]/10). It is thus profitable to postpone the project for one year, since the cost-effectiveness ratio is lower, and by the same argument it is profitable to postpone the project indefinitely into the future.

The postponement argument is, however, flawed. First note that if the cost-effectiveness analysis is carried out on a basis of strict budget maximisation, the postponement issue does not arise since the aim is to maximise the health effects with the current annual budget, and projects financed by next year's budget do not enter this maximisation problem.

In assessing this argument it is useful to make a distinction between health care programmes for the same cohort of individuals and health care programmes for other cohorts. Consider first programmes for the same cohort, e.g. a cohort of 50 year old men. In this case the argument does not hold because the effects cannot be assumed to be the same as this year in future years, for the same health programme. For instance, if we postpone the programme for 100 years then the effects of the programme will be zero, since all the members of the cohort will be dead. In the case with the same cohort the timing of health programmes is essential to achieve the maximum benefits, and in some cases it may be beneficial to postpone programmes until the cohort is older. An example of this is hypertension treatment, where the cost-effectiveness of hypertension treatment has been shown to improve with increasing age; this means that at some ages it may well be cost-effective to postpone the treatment (Johannesson 1994,1995).

We can then investigate the validity of the postponement argument in comparisons between different cohorts, e.g. a programme this year for 50 year old men versus the same programme next year for a new cohort of 50 year old men. In this case it may seem more reasonable that it would be possible to get the same effects of the programme next year as last year. If we use the cost-benefit perspective the WTP for the programme next year should be discounted, and if the income increases for new cohorts, as could be expected, the WTP for the health effects of the programme is likely to increase as well. This growth in WTP would then at least partially offset the discounting of WTP (Viscusi 1992). If health effects are used as the effectiveness unit it could then be argued that they should be discounted at a lower rate for new cohorts, or not discounted at all (i.e. the discount rate should reflect the net effect of the discount rate and the growth in WTP).

According to the postponement argument this would then imply that health care programmes should be postponed indefinitely to new cohorts since the discount rate for effects is lower than for costs. However, this is wrong since if the WTP grows over time, more resources will be devoted to health care programmes (if the technology stays constant). It is therefore not possible to postpone projects and get the same effects for future cohorts, since the marginal productivity of the resources next year, in terms of producing health effects, will not in general be the same as this year. In fact the amount of resources devoted to health care programmes increases over time as societies grow richer, since programmes that were not beneficial some years ago may be beneficial now when we have more resources to spend. In reality it is not just the income growth over time that will affect the WTP for health effects on the margin for new cohorts, since technological changes, the health status, total spending on health

programmes etc may also affect the WTP. The problem of discounting health effects of new cohorts is related to the problem of how to discount costs and benefits for future generations, which is not a trivial issue (Freeman 1993; Johansson 1993).

As shown above, the postponement argument does not provide a sound basis for discounting health effects, and it is not obvious that the health effects should be discounted at the same rate as for costs, since this method is based on an argument of constant WTP for health effects on the margin over time. For discounting health effects within the same cohort, it seems reasonable at the moment to use the same discount rate for costs and health effects, but to supplement this analysis with a sensitivity analysis where the discount rate is varied independently of the discount rate for costs, or at least to add a sensitivity analysis where the health effects are not discounted at all.

Discounting within a cohort will cover most cases that are considered in cost-effectiveness analysis of health care programmes, since in practice it does not seem very meaningful to compare doing a programme this year with doing the same programme next year for a new cohort. Whether the programme should be implemented next year should be based on the available information at the time of the decision next year. Different programmes will of course affect different cohorts, but this is probably best dealt with by discounting health effects the same way for each cohort, and then possibly using different prices per unit of health effects in different ages.

The exception to this, when discounting of health effects for new cohorts becomes important, is for programmes that if implemented now will improve the health or most likely reduce the health risks in the future for new cohorts (e.g. the storage of nuclear waste). If the reduction in risk is for cohorts currently alive (e.g the programme reduces the risk of mortality for persons who are now 20 years old when they reach the age of 50), the discounting can be carried out in the same way as above (i.e. discount the health effects with the same rate as for other cohorts, and then possibly use a different price per unit of health effects than for other cohorts). If the reduction in risk accrues to currently unborn generations, the issue has to do with the distribution of resources between present and future generations; in such cases it seems reasonable to include a case with zero discounting of health effects for future generations, at least as a sensitivity analysis.

In the empirical studies that have tried to estimate the trade-offs between life-saving now and life-saving for future generations using survey techniques (i.e. using hypothetical choices between X number of lives saved this year versus Y number of lives saved in the future), the trade-off between generations has been approximately consistent with the discount rates presently used in economic evaluations (Cropper et al 1991).

9.8 A Cost-Effectiveness Application

In this section we illustrate the use of cost-effectiveness analysis with an application. The application is an analysis of the cost-effectiveness of hypertension treatment in Sweden (Johannesson 1995). There are two fundamental policy issues pertinent to economic evaluations of hypertension treatment. The first issue concerns the criteria for intervention and involves comparisons between treatment and non-treatment in different patient groups. Such an analysis can guide decisions about the optimal cut-off point for treatment with respect to age, sex and other risk factors. The second issue concerns the choice of therapy and involves comparisons between alternative therapies to lower blood pressure. Such an analysis can guide decisions about cost-effective treatment strategies.

The aim of this study was to carry out an analysis of the cost-effectiveness of hypertension treatment in different patient groups in Sweden. This means that treatment was compared to no treatment in different patient groups and the decision that was analysed was thus whether patients should be treated or not.

A cost-effectiveness analysis was carried out with life-years gained as the measure of health effects. The cost-effectiveness ratios were calculated as the net costs (the treatment costs minus saved costs of reduced cardiovascular morbidity) divided by the number of life-years gained (the increase in life expectancy). No adjustment for quality of life was carried out, due to the lack of valid quality-of-life weights to apply. The analysis was carried out from a societal perspective, also including costs outside the health care system. All analyses were carried out with a treatment duration of one year to be valid at a specific age. Both costs and life-years were discounted by 5% to take into account the timing of costs and life-years.

All analyses were carried out for men and women separately in the age groups <45 years, 45-69 years and ≥70 years. The population was also divided into four different diastolic blood pressure (DBP) intervals: 90-94 mm Hg, 95-99 mm Hg, 100-104 mm Hg and ≥105 mm Hg. Costs were calculated in 1992 prices in Swedish Crowns (SEK; exchange rate 1992: about $1 = SEK 6).

A computer model for cost-effectiveness anlysis of cardiovascular disease prevention was used to carry out the cost-effectiveness analyses. The model was based on logistic risk functions for coronary heart disease (CHD) and stroke, taken from the Framingham heart study. The risk factors included in the risk functions were: diastolic blood pressure, serum cholesterol, glucose intolerance, smoking, and left ventricular hypertrophy (LVH). The values of these risk factors used in the cost-effectiveness estimations were based on a survey of hypertension patients in Sweden. Since reliable data about the prevalence of glucose intolerance were not available, the prevalence of diabetes was used instead.

In the model, CHD was divided into recognized myocardial infarction, unrecognized myocardial infarction, angina pectoris (uncomplicated), coronary insufficiency, and sudden death (within one hour from the onset of disease), according to the definitions used in the Framingham study. The survival after the disease events was also based on the Framingham study. The risk functions were used to determine the pre-treatment risks of CHD and stroke. The result of the most recent meta-analysis of clinical trials was used to determine the gain from treatment. In the above meta-analysis, the risk reduction was 38% for stroke and 16% for CHD. It was assumed that the risk reduction applied to all the different patient groups, since there was little evidence that the relative risk reduction varied systematically between the patient groups.

The costs of hypertension treatment were divided into the costs of drugs, the costs of physician consultations (including laboratory tests), and the costs of travel and time for the patients. Using previous estimations in Sweden, the annual treatment cost was estimated to be SEK 3,000 (SEK 1,600 for drugs, SEK 1,100 for consultations and SEK 300 for travel and time). This treatment cost was used for all the patient groups in the cost-effectiveness analysis.

In order to carry out the cost-effectiveness estimations, it is also necessary to obtain data about the increased costs due to morbidity in CHD and stroke so as to be able to estimate the reduced morbidity costs due to the reduced risk of stroke and CHD. These costs were taken from other Swedish studies, and both health care costs and production losses due to morbidity were included. The cost per gained life-year in the study in the different patient groups is shown in Table 13.

Table 13. Cost per life-year gained of hypertension treatment

DBP	Age					
	<45		45-69		≥70	
	M	W	M	W	M	W
90- 94	947	2506	68	215	25	21
95- 99	780	1894	34	133	14	7
100-104	636	1388	1	59	3	-
≥105	440	746	-	-	-	-

-=Cost saving

The cost per gained life-year decreased with age for both men and women. This cost was also much higher for younger women than for younger men, but the difference decreased with age and in the oldest age-group the results were similar for men and women. As expected, the cost per gained life-year also decreased with a higher pre-treatment blood pressure. In the lowest blood pressure range (90-94 mm Hg) the cost per gained life-year was SEK 947,000 and SEK 2,506,000 respectively for younger men and women, SEK 68,000 and SEK 215,000 respectively for middle-aged

men and women, and SEK 25,000 and SEK 21,000 respectively for older men and women.

A sensitivity analysis was also carried out for a number of variables such as the relative risk reduction, the absolute risk level, the morbidity costs, the survival after CHD and stroke, and the discount rate. The result of the sensitivity analysis showed that the cost per gained life-year was very stable in the oldest age group for both men and women, with a cost per gained life-year below SEK 100,000 in all analyses. The result was also relatively stable for middle-aged men and women, whereas the cost per gained life-year varied widely for younger men and women. The decrease with age for the cost per gained life-year was stable towards different assumptions for both men and women, but the size of the decrease was sensitive, especially towards the discount rate.

It was noted in the study that in order to determine whether or not a treatment is cost-effective, it is necessary to decide the price per gained life-year that society is willing to pay. As one comparison, the authors used the value per statistical life saved of SEK 11 million that is used in cost-benefit analysis of road investments in Sweden. This was divided by the average life-expectancy after a traffic accident in Sweden in order to obtain the implied price per discounted gained life-year of SEK 700,000 (at a 5% discount rate; about SEK 350,000 per undiscounted gained life-year). This price was used as one comparison with the cost-effectiveness ratios in the different patient groups.

It was also noted that it is common to compare the cost-effectiveness ratios of other alternative uses of health care resources, in order to implicitly decide how much society is willing to pay per gained life-year. It was stated that in most studies, treatments with cost-effectiveness ratios below SEK 100,000 are considered to be highly cost-effective. The value of SEK 100,000 per gained life-year was therefore used as a kind of lower bound for the price per gained life-year.

It was concluded in the study that the cost per gained life-year decreased with age for both men and women and that the difference was large, especially between the youngest age group and the other two age groups. It was also concluded that since the cost per gained life-year was below SEK 100,000 in all the sensitivity analyses, this indicates that it is cost-effective to treat elderly men and women who have a DBP $\geq$90 mm Hg.

For middle-aged men with a DBP of 90-94 mm Hg, the cost per gained life-year was SEK 68,000 in the base-case analysis and it varied between SEK 8,000 and SEK 425,000 in the sensitivity analyses. On the basis of these figures, it was concluded that the treatment also seems to be cost-effective for middle-aged men with a DBP $\geq$90 mm Hg, although the evidence was not as strong as for elderly men and women.

The cost per life-year gained in the base-case analysis was SEK 215,000 for middle-aged women with a DBP of 90-94 mm Hg, and it varied between SEK 62,000 and SEK 515,000 in the sensitivity analyses. It was concluded that the cost-effectiveness ratios for middle-aged women seem to be on a reasonable level, indicating that it is cost-effective to treat middle-aged women with a DBP ≥90 mm Hg, but that the conclusion was not obvious. However, the conclusion was reinforced by the observation that the drug cost of treating hypertension in Sweden is lower for women than for men, which was not taken into account in the base-case cost-effectiveness analysis.

For younger men and women it was concluded that the cost-effectiveness ratios seemed to be high (e.g. SEK 947,000 for men and SEK 2,506,000 for women with a DBP of 90-94 mm Hg in the base-case analysis), indicating that it is questionable whether it is cost-effective to treat younger men and women with mild hypertension, especially younger women.

Overall it was thus concluded that the results indicated that it is in general cost-effective to treat middle-aged and older men and women in Sweden who have a diastolic blood pressure of ≥90 mm Hg, but that it is questionable whether it is in general cost-effective to treat younger men and women with mild hypertension. It was stressed that these conclusions were general, since the cost-effectiveness within each patient group considered in the analysis may vary considerably. It may for instance be cost-effective to treat younger men with mild hypertension if a number of other risk factors are also present.

The study shows the importance of having some information about the willingness to pay per gained life-year. Without this information it is impossible to say anything about whether or not a treatment is cost-effective. The WTP per gained life-year is uncertain, which made it difficult to draw any definitive conclusions in the study. The study did not include quality of life either. It is possible to incorporate quality of life by using quality-adjusted life-years (QALYs) as an effectiveness measure (see the next chapter on cost-utility analysis).

Quality of life can affect the analysis in three different ways. Firstly, the gained life-years may not be in full health, which would reduce the number of quality-adjusted life-years gained. Secondly, the treatment reduces the risk of CHD and stroke morbidity, and including this quality of life gain would increase the number of quality-adjusted life-years gained. Finally, the treatment may cause subjective side-effects, leading to a loss in the quality of life following treatment.

Quality of life was not included in the base-case analysis because of the lack of valid quality weights to apply. However, a sensitivity analysis was used to test in what direction the results would change if the quality of life of CHD and stroke morbidity and the quality of life of treatment were included. According to the sensitivity analysis, this would further increase the differences between the patient groups, and the absolute

cost per quality-adjusted life-year gained would mainly be sensitive towards these assumptions among younger men and women with mild hypertension and middle-aged women with mild hypertension. It was therefore concluded that the inclusion of quality of life would probably not change the general conclusions, except possibly for middle-aged women.

If we look at the costs in the analysis, all the relevant programme costs seem to be included, i.e. health care costs, travel costs, and lost leisure time and working time for the patients. Even though some of the costs of leisure and working time are paid by the patients, it is appropriate to include these costs, since they are not included in the effectiveness measure used, i.e. life-years gained.

The morbidity costs included were the health care costs and the loss of production. Since both these costs are almost fully covered by the public insurance and tax systems, the estimation of the morbidity costs corresponds to the estimation of the external costs due to the change in morbidity. Even if they had not been covered by insurance it would have been appropriate to include these costs, since they are not included in the number of gained life-years used as effectiveness measure.

The change in leisure time was not included among the morbidity costs. The increased working time due to avoided CHD and stroke morbidity decreases the consumption of leisure, and this may also be offset to some extent by an increase in leisure due to decreased health production time. Even though the net effect on leisure was probably a loss, it seems safe in this case to assume that this was more than offset by the pure health gain due to decreased CHD and stroke morbidity.

No mortality costs were included. It could be argued that the external costs of increased length of life should have been included. This would probably have decreased the cost per gained life-year somewhat in the youngest age-group, and the effect on middle-aged men and women would probably have been small. For the oldest age group the cost per gained life-year would have increased, but since the cost per gained life-year in this age group was very low this would probably not have changed the conclusions.

9.9 Conclusions

This chapter was devoted to cost-effectiveness analysis, where the costs are measured in monetary units and the health effects in non-monetary units such as life-years gained. Cost-effectiveness analysis is based on the maximization of the health effects for a given budget. A fixed budget can be used to maximise the health effects, using information about the incremental cost-effectiveness ratios of different health care programmes; this will also yield a price per effectiveness unit on the margin (i.e. the incremental cost-effectiveness ratio of the programme with the highest incremental cost-effectiveness ratio that is implemented). Alternatively, a price per effectiveness

unit can be set and used as the decision rule, which will implicitly yield a budget (i.e. the total cost of all the health care programmes implemented).

In practice, in order to follow the budget maximization approach strictly it is necessary to identify a single budget that is used for maximising health effects. However, this approach would lead to suboptimisation since only costs which fall on that budget would be included.

In practice cost-effectiveness analysis in the health care field is often based on the net change in health care costs divided by the increase in health effects. This is often based on a decision-maker approach to economic evaluation, where the aim is viewed as maximising whatever the decision-maker wants to maximise. It is assumed that some decision-maker, who is usually not identified, wants to maximise the health effects using some kind of total health care budget.

This approach is not consistent with the strict budget maximisation approach, unless only costs that fall into a single annual budget are included; this would mean that any costs paid by the patients should be excluded, as well as any future cost changes and all costs that fall on other budgets. The prices that are used should also be the prices facing the budget holder, which need not have any relationship to opportunity costs. The approach of only using health care costs is thus inconsistent with the principles of budget maximisation of cost-effectiveness analysis and has no foundation in the theory underlying cost-benefit analysis.

Here it was proposed that cost-effectiveness analysis should instead be viewed as a subset of cost-benefit analysis. Cost-effectiveness analysis should then be interpreted as an estimation of the cost function to produce health effects, and all costs should be included irrespective of who pays. In order to decide whether or not a treatment is cost-effective, the estimation of the cost function has to be supplemented by information about the WTP per unit of health effects in order to decide whether or not benefits exceed costs.

According to the way cost-effectiveness analysis is usually used, this implies that the WTP per unit of health effects is the same for everyone and constant for different sizes of the health change. However, it would also be possible to use different prices per unit of health effects, depending for instance on the size of the health change, in order to make the analysis completely compatible with cost-benefit analysis.

REFERENCES

Baumol WJ. Economic theory and operations analysis. 3rd edition. London: Prentice Hall, 1972.

Boyle MH, Torrance GW, Sinclair JC, Horwood SP. Economic evaluation of neonatal intensive care of very-low-birth-weight infants. New England Journal of Medicine 1983;308:1330-1337.

Cropper M, Aydede SK, Portney PR. Discounting human lives. American Journal of Agricultural Economics 1991;73:1410-1415.

Edelson JT, Weinstein MC, Tosteson AN, Williams L, Lee TH, Goldman L. Long-term cost-effectiveness of various initial monotherapies for mild to moderate hypertension. Journal of the American Medical Association 1990;263:407-413.

Freeman AM. The measurement of environmental and resource values: theory and methods. Washington D.C.: Resources for the Future, 1993.

Johannesson M. The impact of age on the cost-effectiveness of hypertension treatment: an analysis of randomized drug trials. Medical Decision Making 1994;14:236-244.

Johannesson M. The cost-effectiveness of hypertension treatment in Sweden. PharmacoEconomics 1995;7:242-250.

Johannesson M, Weinstein MC. On the decision rules of cost-effectiveness analysis. Journal of Health Economics 1993;12:459-467.

Johansson P-O. Cost-benefit analysis of environmental change. Cambridge: Cambridge University Press, 1993.

Jönsson B, Haglund U. Cost-effectiveness of misoprostol in Sweden. International Journal of Technology Assessment in Health Care 1992;8:234-244.

Keeler EB, Cretin S. Discounting of life-saving and other nonmonetary effects. Management Science 1983;29:300-306.

Klarman HE, Francis JOS, Rosenthal G. Cost-effectiveness analysis applied to the treatment of chronic renal disease. Medical Care 1968;6:48-54.

Logan AG, Milne BJ, Achber C, Campbell WP, Haynes RB. Cost-effectiveness of a worksite hypertension treatment programme. Hypertension 1981;3:211-218.

Morrall JF III. A review of the record. Regulation 1986;10:25-34.

Schulman KA, Lynn LA, Glick HA, Eisenberg JM. Cost-effectiveness of low-dose zidovudine therapy for asymptomatic patients with human immunodeficiency virus (HIV) infection. Annals of Internal Medicine 1991;114:798-802.

Sintonen H, Allander V. Comparing the cost-effectiveness of drug regimens in the treatment of duodenal ulcers. Journal of Health Economics 1990;9:85-101.

Sugden R, Williams A. The principles of practical cost-benefit analysis. Oxford: Oxford University Press, 1978.

Viscusi WK. Fatal tradeoffs: public and private responsibilities for risk. New York: Oxford University Press, 1992.

Weinstein MC. Principles of cost-effective resource allocation in health care organisations. International Journal of Technology Assessment in Health Care 1990;6:93-103.

Weinstein MC, Stason WB. Hypertension: a policy perspective. Cambridge MA.: Harvard University Press, 1976.

Weinstein MC, Stason WB. Foundations of cost-effectiveness analysis for health and medical practices. New England Journal of Medicine 1977;296:716-721.

Weinstein MC, Zeckhauser R. Critical ratios and efficient allocation. Journal of Public Economics 1973;2:147-157.

Williams A. Economics of coronary artery bypass grafting. British Medical Journal 1985;291:326-329.

Winston WL. Operations research: applications and algorithms. Second edition. Boston: PWS-Kent Publishing Company, 1991.

10. COST-UTILITY ANALYSIS

In many cases it is difficult to apply cost-effectiveness analysis since the health effects are difficult to express in a single effectiveness unit. Apart from affecting survival, a treatment may for instance also affect the health status, which means that the effects on health status will not be included if the gained life-years are used as the effectiveness measure. A health care programme may also affect more than one type of event, making it difficult to use the number of avoided events as the effectiveness measure if the types of event are different, e.g. a treatment may reduce the risk of some disease but also increase the risk of some sort of side-effect.

For treatments that mainly affect the health status of individuals, it may also be difficult to find a suitable effectiveness measure, e.g. a drug that improves the quality of life of arthritis patients. If different effectiveness measures are used in different areas it also becomes difficult to compare the cost-effectiveness in different fields. Because of these problems with using some types of natural unit as the effectiveness measure, one of the main challenges of the cost-effectiveness approach has been to develop a single outcome measure that incorporates the effects on both the quality and the quantity of life.

Quality-adjusted life-years (QALYs), where the years of life are multiplied by a weight reflecting quality of life, were developed for this purpose (Klarman 1968; Bush et al 1973; Weinstein & Stason 1976; Boyle et al 1983; Williams 1985). When QALYs are used as the outcome measure in cost-effectiveness analysis, the method is frequently referred to as cost-utility analysis. We follow this distinction here, and use the term cost-utility analysis for the special case of cost-effectiveness analysis where a single effectiveness measure is used that incorporates effects on both the quality and the quantity of life. Since cost-utility analysis is a special case of cost-effectiveness analysis, the decision rules analysed in Chapter 9 apply also to cost-utility analysis and will not be discussed further here.

Although the QALY was the first outcome measure to be proposed in cost-utility analysis and is the most commonly used outcome measure in cost-utility analysis, it is also possible to use other outcome measures. Mehrez & Gafni (1989,1991) for instance argued that QALYs are unlikely to be consistent with individual preferences and proposed an alternative outcome measure, healthy years equivalents (HYEs). The relative advantages and disadvantages of QALYs and HYEs have received a lot of attention in the literature recently (Buckingham 1993; Culyer & Wagstaff 1993; Gafni et al 1993, Mehrez & Gafni 1993, Johannesson et al 1993).

In the first section in this chapter about cost-utility analysis we describe different methods of measuring the quality weights so as to construct QALYs. In the following two sections we analyse the relationship between QALYs, HYEs and individual preferences. The implications of discounting for QALYs and HYEs are analysed next.

We thereafter discuss the relationship between cost-utility analysis and cost-benefit analysis. After a cost-utility application we end the chapter with some conclusions.

10.1 QALYs And Their Measurement

QALYs are estimated by assigning every life-year a weight between 0 and 1 and summing the weights in the different years. A weight of zero reflects death or a health status that is equal to being dead, and 1 reflects full health. For example, the number of QALYs for a certain number of years with arthritis followed by death is then equal to the number of years multiplied by the quality weight, e.g. if the quality weight is 0.7 and the number of years is 10 the number of QALYs is equal to 7 (10*0.7).

There are two different lines of reasoning behind QALYs and the measurement of the quality weights. According to the first line of reasoning, represented by for instance Williams (1985), the quality weights should be determined by some kind of socio-political process or by the "decision-maker" and be politically acceptable. The quality need not have anything to do with individual preferences. This approach is rooted in the decision-maker approach to economic evaluation discussed in the last chapter. In our view it is a problematic approach.

Firstly, it is unclear who the decision-maker is, and no decision-maker is usually identified in studies using this approach. The approach also lacks theoretical foundation and makes it possible to determine the quality weights by more or less arbitrary procedures. The only way to try and give the approach a theoretical foundation would be to assume that we have a perfect political system where the decision-makers act as perfect agents for the population. Such an assumption does not seem very realistic (Buchanan 1969), and if the assumption was true the decision-makers would probably want to base the quality weights on the preferences of the population.

The second line of reasoning behind QALYs and the measurement of quality weights, represented for instance by Torrance (1986), is that QALYs should be based on individual preferences. This means that an individual should prefer a treatment that leads to more QALYs rather than a treatment that leads to fewer QALYs (assuming that all costs for the individual not included in the QALYs are the same for both treatments), if the number of QALYs is based on the quality weights of the individual. This is the approach we will mainly be concerned with here, in accordance with the individualistic foundation of welfare economics and cost-benefit analysis. In actual applications the distinction between the decision-maker approach and the individual preference approach to QALYs is often unclear, since in these studies it is seldom stated what the underlying rationale is for using QALYs.

There are also a number of different sources of the quality weights used to construct QALYs. Some studies are based on assumptions by the researchers about the quality

weights, which does not seem to be a very reliable approach. However, if it can be shown in a sensitivity analysis that the conclusions of a study are not sensitive towards a specific quality weight varied over a broad plausible range, it may be unnecessary to measure that weight. Expert judgements, for instance by physicians, are sometimes used to determine the weights. This seems to be close to the decision-maker approach to QALYs and cost-utility analysis and cannot be recommended. There seems to be no reason why guesses by experts should be a reliable indicator of the subjective tradeoff between quality and quantity of life of individuals. It is also possible to use quality weights from other studies. However, there is a tendency for this to turn guesses into "truths", i.e. an assumption about the quality weight is done in one study and then other studies use this assumption without reflecting over its validity.

The quality weights can also be obtained by direct measurements. There are three main methods of direct measurement: rating scale (also called visual-analog scale), standard gamble, and time-trade-off. These methods are described below.

In the rating scale method, the respondent locates the health state to be assessed on a straight line with dead and full health as end points. The scale is then normalised to 1 and the weight is equal to the score where the health state has been located. For instance, if we want to measure the quality weight for the health state arthritis and this health state is located at 7 CM on a 10 CM scale, the quality weight is equal to 0.7 for the arthritis health state.

The rating scale is shown in Figure 1. The advantage of a rating scale is that it is easy to use. However, it does also have a number of important drawbacks. Firstly, it involves no choice, so it is not possible to observe any tradeoff. This makes it very difficult to interpret the results of RS, and the method lacks a theoretical foundation and cannot be related to the underlying theory of QALYs (see the next section). It seems reasonable to assume that the method provides an ordinal ranking of health states, at least for the time horizon that the respondent uses to assess the health states, but in order to use the method to determine the quality weights we have to be able to interpret not only the ranking of the health states but also the differences between them. Thus from a theoretical point of view, it is problematic to use the rating scale method to determine quality weights.

The second method for measuring quality weights is the standard gamble method. With this method the quality weight of a health state is assesed by comparing a specific number of years in the health state to a gamble with a probability (p) of full health for the same number of years and a complementary probability (1-p) of immediate death. The probability of full health (p) is varied until the individual is indifferent between the alternatives, and the quality weight of the assessed health state is equal to p.

The second method for measuring quality weights is the standard gamble method. With this method the quality weight of a health state is assesed by comparing a

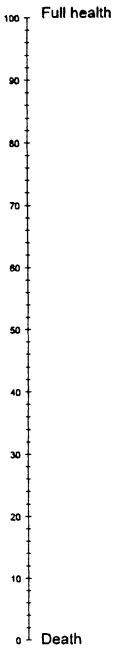

Figure 1. The rating scale method

specific number of years in the health state to a gamble with a probability (p) of full health for the same number of years and a complementary probability (1-p) of immediate death. The probability of full health (p) is varied until the individual is indifferent between the alternatives, and the quality weight of the assessed health state is equal to p.

For instance, if we are comparing 10 years in an arthritis health state with a gamble, with a probability (p) of full health for 10 years and a complementary probability (1-p) of immediate death, and the individual is indifferent at a probability of full health of 0.7, the quality weight is 0.7 for the arthritis health state. The standard gamble method is illustrated in Figure 2.

The standard gamble method is based on the von Neumann-Morgenstern expected utility theory (von Neumann & Morgenstern 1947). According to this theory, a valid cardinal utility function assigns numerical values to different outcomes in such a way that the expected utilities of gambles involving these outcomes result in a ranking of such gambles that agrees with the individual's preferences, i.e. the individual will rank the gambles according to their expected utility. The expected utility theory is based on a number of axioms that will not be described further here, but the reader is referred to the original work of von Neumann & Morgenstern (1947).

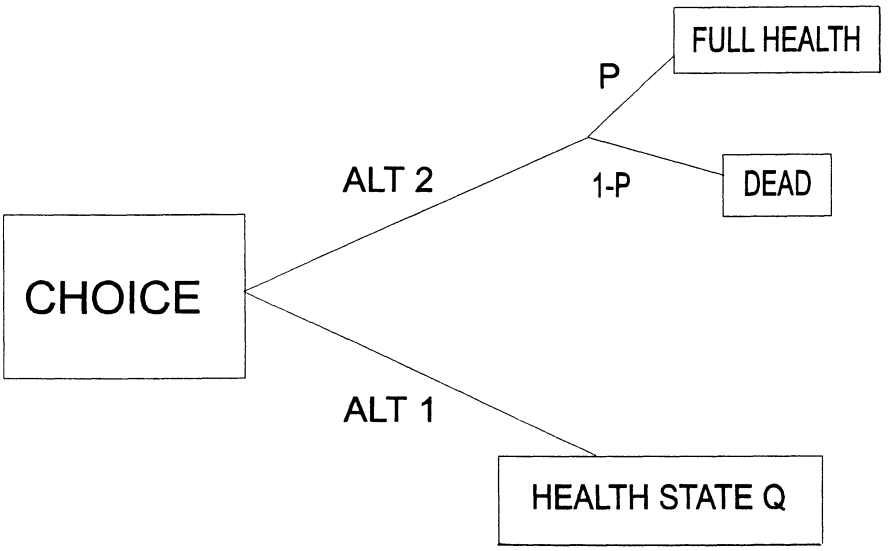

Figure 2. The standard gamble method

According to the expected utility theory, the standard gamble method can be used to estimate a cardinal utility function for different outcomes. This is done by arbitrarily determining the utilities of the best and the worst possible outcomes, and then the utilities of all other outcomes can be measured relative to these reference states using gambles, as in the standard gamble method. When the standard gamble method is used to measure quality weights, the reference states are full health for the same number of years as the number of years in the assessed health state, and immediate death.

Since the quality weights are measured on a scale between 0 and 1, full health is set to 1 and immediate death is set to 0. At the indifference probability the cardinal utility of the assessed health state is equal to the expected utility of the gamble, and the expected utility is equal to the probability of full health multiplied by the utility of full health plus the probability of immediate death multiplied by the utility of immediate death. Since the utility of immediate death is set to zero, the expected utility reduces to the probability of full health at indifference multiplied by the utility of full health.

In the above example, the expected utility is thus equal to 0.7 (0.7*1). This means that the standard gamble method measures the fraction of the utility of full health for the assessed health state as long as immediate death is set to 0. In the example the utility

of the arthritis health state for 10 years is 0.7 of the utility of full health for 10 years (this is the case irrespective of what utility we assign to full health).

The relationship between the standard gamble method and cardinal utility theory is illustrated in Figure 3. The figure shows the utility function with respect to time in full health and the utility function with respect to time in health state A, which is a health state with less than full health (e.g. arthritis). In the figure it is assumed that we want to measure the quality weight for health state A using a time horizon of 10 years. With the standard gamble method we then measure the utility of health state A for 10 years as a fraction of the utility of 10 years in full health. In the figure the utility of 10 years in health state A is 6 and the utility of full health is 10, and the fraction of the utility of full health is thus 0.6 (6/10).

The individual will be indifferent between 10 years in health state A and a gamble with a 60% probability of full health for 10 years and a 40% probability of immediate death. The expected utility of the gamble is 6 (0.6*10+0.4*0) and the expected utility of 10 years in health state A is 6 (1*6). The standard gamble method can thus be interpreted as measuring the fraction of the utility of full health for a health state, and this fraction is used as the quality weight to construct QALYs.

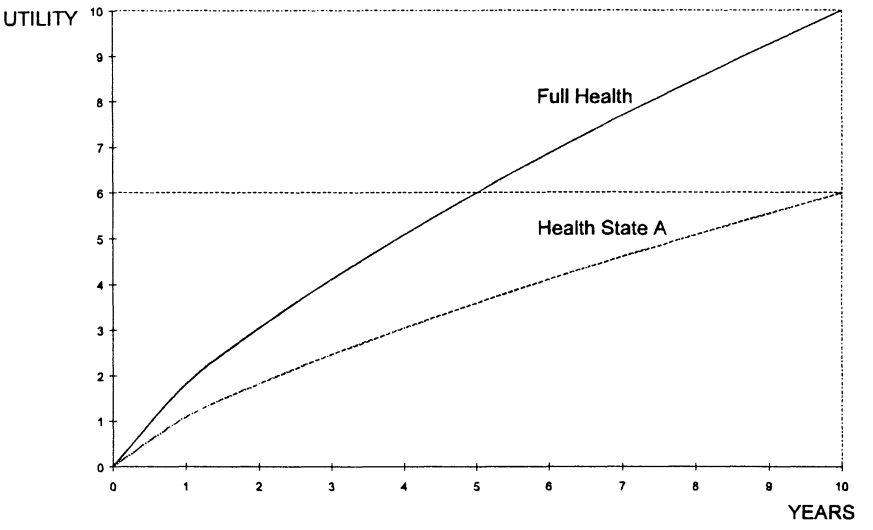

Figure 3. The standard gamble method and cardinal utility theory

The advantage of the standard gamble method is its theoretical foundation in expected utility theory, and in the theory of QALYs which is based on the expected utility theory (see below). The disadvantage of the standard gamble method is that it may be difficult to use because respondents may have difficulties in understanding probabilities. The hypothetical choice is also unrealistic since we seldom face choices between large improvements in health status and large mortality risks, and we never face outcomes where we know that we will live a specific number of years for sure (i.e. both the survival of the health state to be assessed and the full health state are stated in terms of some number of years for sure followed by death).

It is therefore difficult to validate the results of the standard gamble method to see if the hypothetical choices conform to the real choices. It may be possible to do some validation studies for situations for instance where individuals can undergo surgery that improves survival or quality of life, and the surgery is also associated with an immediate increased mortality risk. However, the assessed health state and the "full health alternative" would then be probability distributions over health state and length of life, rather than a specific number of life-years for sure.

The third method of measuring quality weights is the time-trade-off method, developed by Torrance et al (1972). With this method the quality weight of a health state is assesed by comparing T years in the health state to X years in full health. The number of years in full health (X) is varied until the individual is indifferent between the alternatives, and the quality weight of the health state is equal to X/T.

The elicitation of quality weights with the time-trade-off method is illustrated in Figure 4. Assume that we want to measure the quality weight for arthritis and that we carry out the measurement for 10 years in arthritis. If the individual is indifferent between 10 years with arthritis and 7 years in full health the quality weight is equal to 0.7 (7/10).

An advantage with the time-trade-off method is that it is based on a choice (as is the standard gamble method) and it is thus possible to observe a trade-off. In the time-trade-off method the individual is trading off survival for increased quality. Since we assess a point where the individual is indifferent between two alternatives, we also know that the (expected) utility is the same for these two alternatives (both cardinal and ordinal utility give the same utility number for two alternatives that the individual is indifferent between).

In the example, the utility is the same for 10 years with arthritis as for 7 years in full health. The quality weight with the time-trade-off method can thus be interpreted as the fraction of the number of healthy years that leads to the same utility as the assessed health state for the same number of years. In the example, 7 years is 0.7 of 10 years, i.e. the individual is indifferent between the utility of 10 years with arthritis and the utility of 70% of the same number of years in full health. The difference compared to the standard gamble method is thus that the standard gamble method measures the

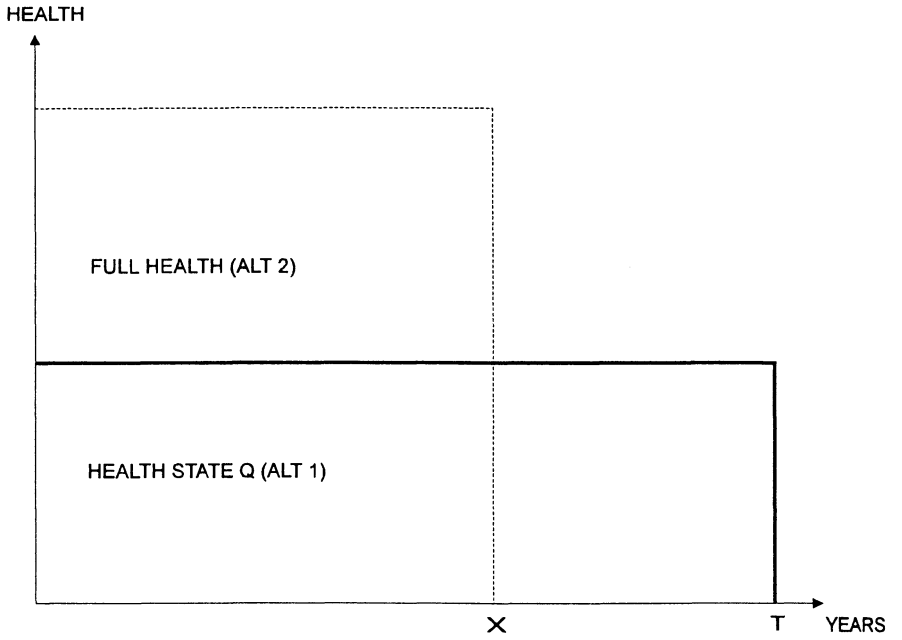

Figure 4. The time-trade-off method

quality weight as the fraction of the utility of healthy years and the time-trade-off method measures the quality weight as the fraction of the number of healthy years.

The time-trade-off method can also be illustrated using cardinal utility functions. This is done in Figure 5, where the same utility functions as in Figure 3 have been drawn, i.e. the utility functions with respect to life-years in full health and health state A. In the figure it is assumed that we are assessing the quality weight with the time-trade-off method for 10 years in health state A. To do this we find the number of years in full health that leads to the same utility as 10 years in health state A.

In the figure, 5 years in full health lead to the same utility as 10 years with health state A. The quality weight is derived by dividing the number of years in full health by the number of years in health state A, which gives a quality weight of 0.5 in the figure (5/10). This quality weight can be interpreted as showing that the utility is the same for 10 years in health state A as for 50% of 10 years in full health, i.e. the quality weight is measured as the fraction of the number of healthy years.

In the figure the quality weights differ between the standard gamble method and the time-trade-off method because the marginal utility of additional years is decreasing, due to the way the utility functions have been drawn. This means that 50% of the

number of healthy years leads to more utility than 50% of the utility of healthy years (in the figure 50% of the number of years in full health leads to 60% of the utility of the same number of years in full health). The shape of the utility function with respect to life-years in different health states will be further discussed below in the section about QALYs and individual preferences.

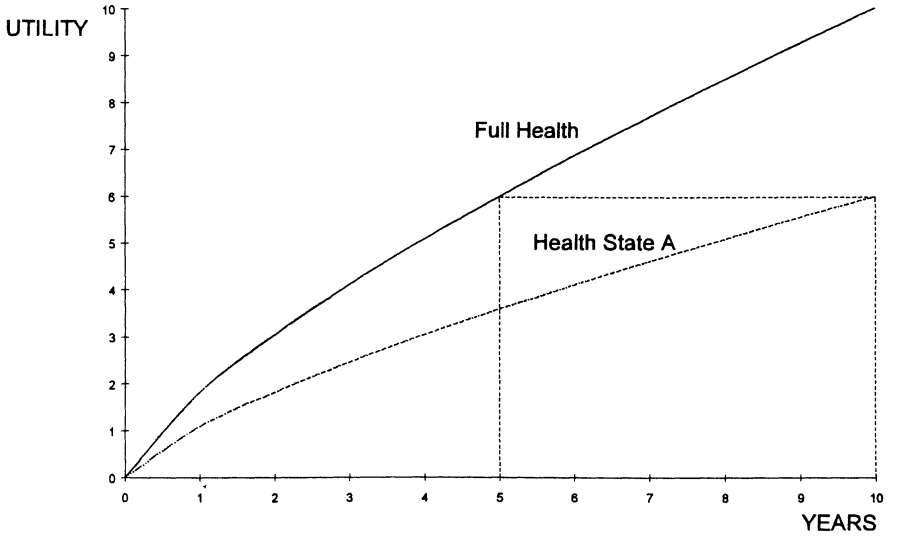

Figure 5. The time-trade-off method and cardinal utility theory

The advantage of the time-trade-off method is that the quality weights can be given a theoretical interpretation in the cardinal utility theory that underlies QALYs, i.e. it shows two combinations of life-years and quality with the same cardinal utility. The time-trade-off method may also be somewhat easier to use than the standard gamble method. The main disadvantage of the method is that the choice is unrealistic; in real life we never make choices between a number of years for sure in a health state with less than full health and less years for sure in full health. It thus seems impossible to validate the time-trade-off method against real choices as long as the alternatives are defined as a specific number of years in different health states with certainty.

One important issue is which of these three methods for measuring quality weights is to be preferred in cost-utility applications. In our view, based on the current state of knowledge, either the time-trade-off method or the standard gamble method is to be

preferred. Since the weights of the rating scale method are impossible to interpret in a meaningful way, this method cannot be recommended. If it could be established that the scores with the rating scale method can be converted to scores with the time-trade-off method or the standard gamble method in a reliable way, an argument could be made in favour of the rating scale, since it is easy to use (provided that the converted scores are used to construct QALYs).

Before one can recommend using the scores of the rating scale method to construct QALYs directly, it has to be shown in empirical studies that these scores lead to a QALY measure which represents individual preferences as well or better than QALYs based on the time-trade-off method or the standard gamble method. This could be tested by having individuals rank health profiles directly, i.e. using combinations of life-years and quality of life, and then comparing this ranking with the ranking of QALYs based on their own quality weights with the different methods. If the rating scale performes only slightly worse than the time-trade-off method and the standard gamble method, it may also be possible to argue that the method should be used because of its practical advantages.

It is more difficult to discriminate between the time-trade-off method and the standard gamble method, since if the underlying assumptions of QALYs are valid (see below) both methods should yield the same number of QALYs. To make any recommendation about the choice between the time-trade-off method and the standard gamble method, more empirical research on the ranking properties of QALYs based on the two different methods is needed.

One interesting point which should be noted, in relation to the section about the assessment of WTP with the contingent valuation method in Chapter 6, is that the bias issues discussed for contingent valuation also seem to be relevant for the assessment of quality weights. For instance, the standard gamble method and the time-trade-off method can be expected to be susceptible to starting point bias (i.e. the starting probability in the choice in the standard gamble method and the starting number of years in full health in the time-trade-off method). However, there has been surprisingly little research on these issues when it comes to the assessment of quality weights.

Another important issue in connection with the measurement of quality weights is what population to carry out the measurements on. The natural candidate seem to be the target population of a health care programme. This would mean people in the health states to be assessed for programmes that improve the quality of life in different chronic health states, and the population at risk for programmes that reduce the risk of different diseases.

An argument could also be made for always using individuals in the actual health states, since they have experience of the health states. However, this could also lead to incentives for answering the questions in a strategic way and exaggerate the seriousness of their own health states, so that health improvements will look more

attractive (the population at risk would also have the same incentives). Some people also argue that a general population sample should be used to assess the quality weights, since the general population represents potential patients. It is also argued that by using a general population, sample weights for more than one health state can be assessed from each respondent.

The main problem of using a general population sample is that the respondents lack experience of the health states. It is not obvious either that weights for more than one health state should be assessed from each respondent, since this can lead to problems with anchoring between the questions. However, to our knowledge this has never been tested in studies assessing quality weights.

There also exist other approaches to measuring quality weights, apart from the direct methods described above. Some instruments for measuring quality of life can be collapsed to a 0-1 scale and are sometimes used to construct quality weights. However, unless the weights with these instruments have been validated against either the time-trade-off method or the standard gamble method, this would seem to be a highly questionable strategy.

So-called multiattribute utility functions are also sometimes used to estimate the quality weights (Torrance et al 1982). This basically means that an attempt is made to derive a function for the utility of different health states depending on different attributes of the health states. By observing the level of the attributes of different health states, the quality weights can then be estimated by entering the level of these attributes in the function and deriving a quality weight. This also seems to be appropriate if the quality weights derived with this approach can be validated against the time-trade-off method or the standard gamble method.

10.2 QALYs And Individual Preferences

For QALYs to be a useful effectiveness measure in cost-utility analysis, they should reflect individual preferences. By individual preferences we mean that in the choice between two treatments the individual should prefer the treatment that gives most QALYs, provided that the quality weights are based on the weights of the individual, i.e. the QALYs of the individual should rank treatments according to individual preferences.

In this section we will analyse the relationship between QALYs and individual preferences. We will specifically analyse under what circumstances QALYs will be a valid cardinal utility function. This means that the assumptions of expected utility theory (von Neumann & Morgenstern 1947) are accepted despite the descriptive evidence that individual decisions often violate the theory (Schoemaker 1982).

This is because the only theoretical foundation that has been given for QALYs is in terms of expected utility. The conditions when QALYs are a valid cardinal utility function are important, since in situations with risks of different events the cardinal utility of an outcome is simply multiplied by the probability of that outcome to obtain the expected utility. If the axioms of expected utility are satisfied, the expected utility will then rank different "gambles" (i.e. probability distributions) of health profiles according to the preferences of the individual.

To illustrate the different assumptions for QALYs to be a valid cardinal utility function, we will consider a simplified model where the utility is assumed to depend on health status (Q) and the number of life-years (T). Q* denotes full health, and all health states are assumed to be preferred to death. Health states worse than death are not considered. Health states worse than death violate the mutual utility independence condition for QALYs (see below).

Lifetime health profiles are referred to as (Q,T), reflecting constant quality of life (Q) for T years followed by death. The health status Q is assumed to be constant over time, and health profiles with varying quality of life over time are not considered. However, the implications of varying quality of life over time for the assumptions of QALYs are discussed below. No discounting of QALYs is assumed either, since the implications of discounting outcome measures in cost-utility analysis are considered in a separate section below.

The model that is used to define QALYs and the underlying preference assumptions is the one used by Pliskin et al (1980) in their analysis of the theoretical properties of QALYs. It should also be noted that it is possible to derive QALYs from other theoretical models than the one used by Pliskin et al (1980), see for instance Broome (1993) and Bleichrodt (1995a). The strength of the Pliskin et al (1980) model is that it is directly related to the methods of measuring the QALY weights (i.e. the TTO method and the SG method).

In defining QALYs we follow the theoretical framework developed by Pliskin et al (1980) and a distinction is made between risk-neutral QALYs and the more general risk-adjusted form (Miyamoto & Eraker 1985,1989). It should be kept in mind, however, that the risk-neutral QALY model is the QALY model generally used, which is the main focus of our attention here (and when we refer to QALYs below in the comparison of different outcome measures we mean risk-neutral QALYs). A distinction is also made between whether the time-trade-off method or the standard gamble method is used to measure QALYs.

Pliskin et al (1980) identified three conditions for QALYs to be a valid cardinal utility function. These three conditions impose increasingly more restrictive assumptions on the shape of the utility function over life-years in different health states. The first condition is that life-years (T) and quality (Q) must be mutually utility independent. This means that the ranking of gambles over one of the two attributes (quality and

life-years) is independent of the other attribute (i.e. ranking of gambles of quality is independent of the number of life-years and ranking of gambles of life-years is independent of quality).

This means that if for example an individual is indifferent between 10 years in moderate pain and a gamble with a 50% chance of 10 years in severe pain and a 50% chance of 10 years in full health, then the individual should also be indifferent between these alternatives for all other numbers of life-years (i.e. the ranking of gambles of quality is independent of the number of life-years).

It also means that if for instance an individual is indifferent between 10 years in full health and a gamble with a 50% chance of 5 years in full health and a 50% chance of 20 years in full health, then the individual should also be indifferent between these alternatives for all health states other than full health (i.e. the ranking of gambles over life-years is independent of quality). As noted above, this assumption will be violated for health states worse than death since in those health states the utility decreases with more years and the ranking of gambles over life-years is thus not independent of the health status (i.e. a gamble will be ranked higher if the probability of more life-years increases in health states preferred to death, but a gamble will be ranked lower if the probability of more life-years increases in health states worse than death).

Mutual utility independence is illustrated in Figure 6, by considering full health and health state A for 10 years. The independence assumption guarantees that the shape of the utility function over life-years is the same for all health states (i.e. all Q). If utility independence holds, the fraction of the utility of full health for any health state is the same for all time horizons. In Figure 6 for instance the utility of health state A is 0.5 of the utility of full health for all time horizons. It should be noted also that the utility is the same for all health states if the length of life is 0 in the figure, i.e. all the utility functions start at 0. This property is assumed to hold throughout the analysis of QALYs.

The standard gamble method measures the utility of a health state as the fraction of the utility of healthy years, and if the independence assumption holds the quality weight with the standard gamble method will always be the same, irrespective of the number of years for which the measurement is carried out. In the example in Figure 6 the quality weight for health state A with the standard gamble method is equal to 0.5 for any time horizon.

It can be seen that mutual utility independence holds in the figure by noting that the individual will be indifferent between health state A and a gamble with a 50% probability of full health and a 50% probability of immediate death, for all time horizons in the figure, and the ranking of gambles over life-years will also be the same in both health states, since the shapes of the utility functions are identical in both health states. As long as health states worse than death are not considered, the fact that

ranking of gambles of quality is independent of life-years will also imply that ranking of gambles of life-years is independent of quality (and vice versa).

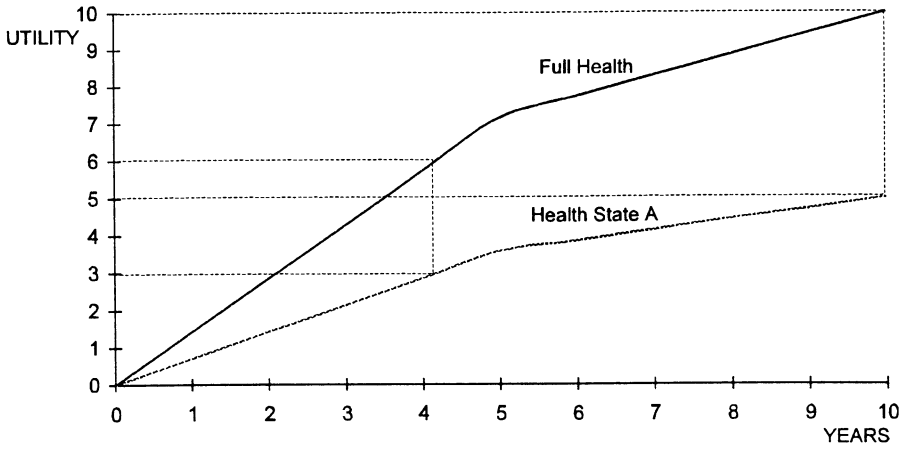

Figure 6. Mutual utility independence

The second preference assumption of QALYs is the constant proportional trade-off property, i.e. the proportion of remaining life that one would be willing to trade off for a specified quality improvement is independent of the amount of remaining life. This means that if for instance an individual is indifferent between 10 years with arthritis and 5 years in full health, then the individual should also be indifferent between 20 years with arthritis and 10 years in full health.

The constant proportional trade-off assumption is illustrated in Figure 7. In this figure the individual is willing to give up 50% of life-years in order to be in full health rather than in health state A. This is the case for any time horizon. The quality weight with the time-trade-off method is measured as the number of years in full health divided by the number of years in the assessed health state. The quality weight is thus measured as the fraction of the number of healthy years. If the constant proportional trade-off property holds, the quality weight with the time-trade-off method will always be the same irrespective of the time horizon for which the assessment is carried out; for

example, in the figure the quality weight with the time-trade-off method is 0.5 for any time horizon.

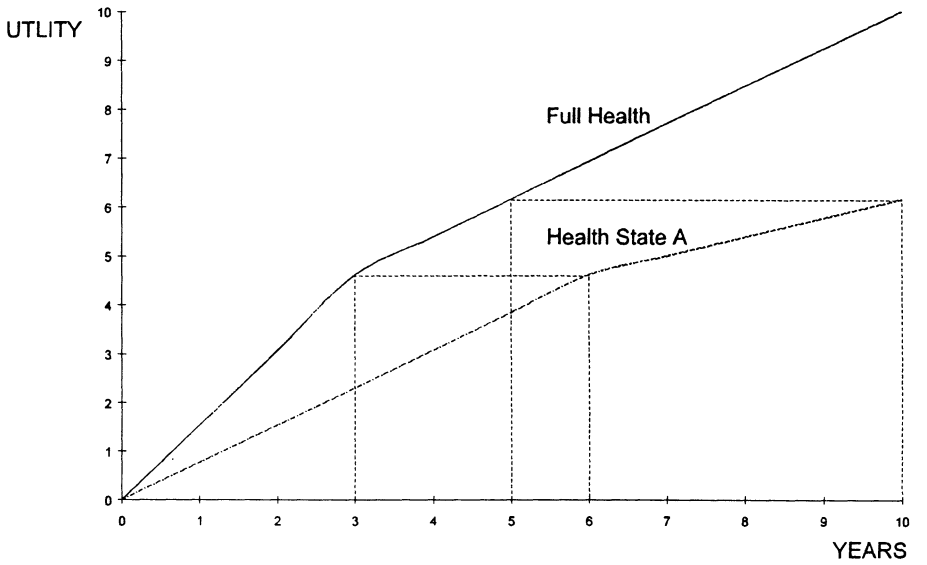

Figure 7. Constant proportional trade-off

As is evident from Figure 7, the constant proportional trade-off property does not imply that the independence assumption holds (the fraction of the utility of full health for health state A varies with the time horizon in Figure 7). Similarly, the independence assumption does not imply that the constant proportional trade-off property holds (the fraction of the number of healthy years for health state A varies with the time horizon in Figure 6).

If both the independence assumption and the constant proportional trade-off assumption hold, a specific shape of the utility function over life-years is imposed. This shape of the utility function, illustrated in Figure 8, is known as constant proportional risk posture over life-years or constant relative risk aversion (Pratt 1964). A utility function with constant proportional risk posture exists if $c=-T[U''/U']$ is constant, where U'' is the second-order derivative and U' is the first-order derivative of the utility function with respect to life-years (c is usually referred to as the Arrow-Pratt index of relative risk aversion) (Pratt 1964).

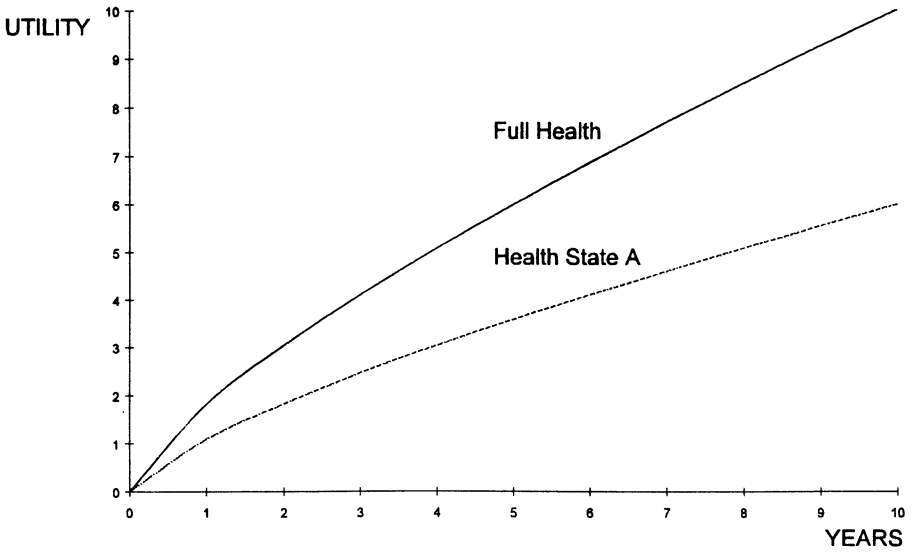

Figure 8. Constant proportional risk posture

If utility independence and constant proportional trade-off hold, risk-adjusted QALYs as defined by Pliskin et al (1980) are a valid cardinal utility function. Risk-adjusted QALYs are defined by the following equation:

$$\text{Risk-adjusted QALYs} = U(Q,T) = [V(Q) * T]^r, \qquad (1)$$

where r is a risk aversion parameter that is equal to 1-c in the definition of constant proportional risk posture above. When r=1, the subject is risk neutral. In equation (1), V(Q) is a value function of quality measured on a scale between 0 (death) and 1 (full health) (Pliskin et al 1980). The value function V(Q) measures the desirability of any health status (Q) as the fraction of X years of full health (Q*) that is the equivalent to X years in the health status under consideration. V(Q) thus corresponds to the quality weight measured by the time-trade-off method. Risk-adjusted QALYs can also be defined on the basis of the standard gamble method, which leads to the following definition:

$$\text{Risk-adjusted QALYs} = U(Q,T) = U(Q) * T^r \qquad (2)$$

In equation (2), U(Q) is a utility function of quality measured on a scale between 0 (death) and 1 (full health). The utility function U(Q) measures the desirability of any

health status (Q) as the fraction of the utility of X years of full health (Q*) that is the equivalent of the utility of X years in the health status under consideration. U(Q) corresponds to the quality weight measured by the standard gamble method.

Note that the difference when risk-adjusted QALYs are calculated on the basis of the time-trade-off method (equation (1)) or the standard gamble method (equation (2)) is that with the time-trade-off method the risk parameter is applied to both the quality weight and life-years, whereas with the standard gamble method the risk parameter is applied only to life-years. U(Q) is thus equal to $V(Q)^r$ and V(Q) is equal to $U(Q)^{1/r}$ if the assumptions of the risk-adjusted QALY model are satisfied. The difference between V(Q) and U(Q) is that V(Q) measures the fraction of the number of healthy years, whereas U(Q) measures the fraction of the utility of healthy years.

As can be seen from Figure 8, both the independence assumption (the fraction of the utility of healthy years for health state A is 0.6 for all time horizons) and the constant proportional trade-off assumption (the fraction of the number of healthy years for health state A is 0.5 for all time horizons) are satisfied. The quality weight with the standard gamble method is thus 0.6 and the quality weight with the time-trade-off method is 0.5 in Figure 8. However, both can be used to calculate the correct number of risk-adjusted QALYs by applying equation (1) or (2). The risk parameter in Figure 8 is 0.74 and $0.5^{0.74}=0.6$.

The assumptions of the risk-adjusted QALY model can thus be stated as follows: constant proportional risk posture over life-years has to hold in all health states and the risk parameter r has to be the same in all health states (i.e. the relative risk aversion has to be constant with life-years and the same in all health states). Note also that using the risk-adjusted QALY model would necessitate estimating the risk parameter r, which could be done by estimating the utility function with respect to life-years using the standard gamble method, with full health for the maximum years of survival and immediate death as reference states (see Miyamoto & Eraker (1985) for an estimation of the risk parameter). The risk-adjusted QALY model has seldom been used in actual applications. The model was derived for health profiles with constant quality over time and it is unclear how to apply the model if the quality varies over time.

For the commonly used risk-neutral QALYs, the shape of the utility function over life-years has to be restricted further and it has to be assumed that risk neutrality over life-years holds, i.e. the utility function over life-years is linear (Pratt 1964). Risk neutrality is illustrated in Figure 9.

Risk-neutral QALYs using the time-trade-off method are defined by the following equation:

$$\text{Risk-neutral QALYs} = U(Q,T) = V(Q) * T \qquad (3)$$

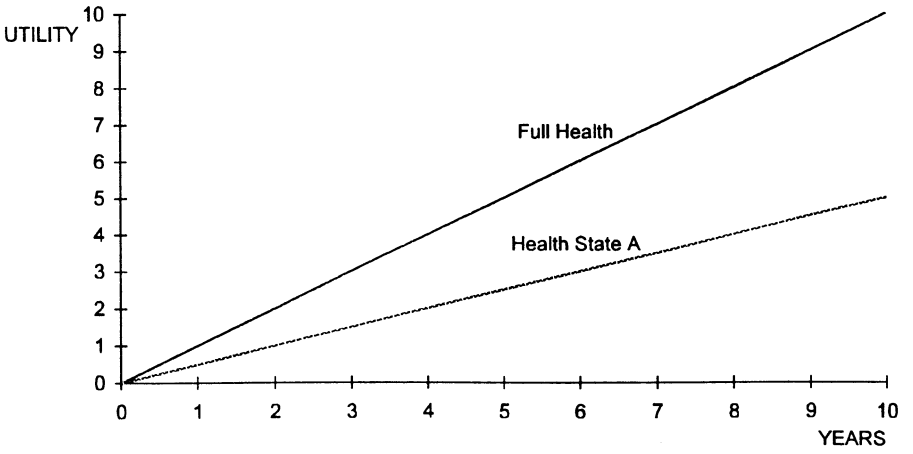

Figure 9. Risk-neutrality over life-years in all health states

Risk-neutral QALYs can also be defined using the standard gamble method:

$$\text{Risk-neutral QALYs} = U(Q,T) = U(Q) * T \qquad (4)$$

If risk neutrality holds (r=1), V(Q) is equal to U(Q) and both the quality weight and the number of QALYs will be the same for both the standard gamble method and the time-trade-off method. As can be seen from Figure 9, if risk neutrality holds in all health states then utility independence (the fraction of the utility of healthy years for health state A is 0.5 for all time horizons) and constant proportional trade-off (the fraction of the number of healthy years for health state A is 0.5 for all time horizons) will also hold. This means that the assumptions of risk-neutral QALYs can be stated as: risk neutrality over life-years has to hold in all health states, i.e. the utility function with respect to life-years has to be linear in all health states.

An important point to note after this section is that both the time-trade-off method and the standard gamble method have the same theoretical foundation in the theory of QALYs, since if either the risk-adjusted or the risk-neutral QALY model is valid both methods will in theory lead to the same number of QALYs. A common misconception in the literature is that only the standard gamble method has a theoretical foundation

and that the validity of the time-trade-off method depends on its ability to come up with the same weights as the standard gamble method.

The time-trade-off method even has a slight theoretical advantage over the standard gamble method. If we estimate risk-neutral QALYs, then in the case of certainty (i.e. no probabilities of different outcomes), the number of QALYs based on the time-trade-off method will always rank health profiles according to individual preferences as long as the constant proportional trade-off assumption holds. This is not true for the standard gamble method.

Assume for instance that we have the case depicted in Figure 8, where constant proportional tradeoff holds, but the individual is risk averse. If the quality weights based on Figure 8, of 0.5 for the time-trade-off method and 0.6 for the standard gamble method, are used to estimate the number of QALYs (i.e. risk-neutral QALYs), the number of QALYs for 10 years in health state A will be 5 according to the time-trade-off method (10*0.5) and 6 for the standard gamble method (10*0.6). Irrespective of the method used, the number of QALYs will be 5 for five years in full health.

QALYs based on the time-trade-off method will correctly rank full health for 5 years and 10 years in health state A equally (i.e. both give 5 QALYs), but QALYs based on the standard gamble method will rank 10 years in health state A higher than 5 years in full health, despite the fact that both health profiles lead to the same utility, as can be seen in Figure 8. Of course in this case none of the measures are a valid cardinal utility function, so if risk is introduced none of the measures will in general rank risky health profiles the same way as expected utility.

The above analysis was based on health profiles where the quality of life is constant over time. In realistic cases quality of life can vary over time, i.e. the individual has a probability distribution of different health states at each point in time and the number of QALYs for each year is estimated by summing the quality weights for the health states multiplied by the probabilities. For QALYs to be a valid cardinal utility function if the quality varies over time, additive utility independence between different periods has to be assumed, i.e. the utility of a health status in one period is independent of the health status in all other periods (Broome 1993).

Additive utility independence does not imply that risk neutrality necessarily holds. Consequently for risk-neutral QALYs to be a valid cardinal utility function when the quality varies over time, it has to be assumed that risk neutrality with respect to life-years in all health states holds and that additive utility independence holds.

The importance of the assumption of additive utility independence for the QALY model is that if the utility of a health status in a specific year is always the same, then this utility can be multiplied by the probability of being in that health status in that year to obtain its contribution to the total number of QALYs (utility). If the assumption

of additive independence does not hold this would not be possible, since the utility would depend on the health status in previous years.

If the health status is constant over time, the utility of a health status in a specific year would always be the same by definition, since the health status in previous periods would always be the same (i.e. there is only one way to get to that health status that year). If the quality varies over time this is no longer the case, and the utility of a health status in a specific year will depend on the health status in previous years unless additive utility independence holds.

The assumption that the quality weight for a specific health status is always the same is the essential feature of QALYs, since this is a practical assumption. It allows you to measure the quality weight of each health status only once and then use this same weight for this health state under all circumstances. This also means that for instance Markov models can be used to model a disease and to estimate the number of QALYs (i.e. Markov models are based on the assumption that all persons in a specific health state of the model are identical, which means that they will get the same additional utility irrespective of the way in which they arrived at that health state). See Sonnenberg & Beck (1993) for a description of Markov models and their use in disease modelling.

This means that in the general case of varying health status over time, QALYs have to be based on risk neutrality over life-years in all health states and additive utility independence. It would be possible to drop the assumption of risk neutrality and define a QALY measure based on the assumption that additive utility independence holds (this will imply that mutual utility indepence holds, since additive utility indepence is a stronger independence assumption than mutual utility independence). This would mean using the standard gamble method to measure the quality weights and adjust the utility of life-years for discounting and age as such. However, as yet no such QALY measure has appeared in the literature. If we drop the assumption of additive utility independence, we have to measure the health or utility of full health profiles directly instead of measuring quality weights. A health profile is a combination of life-years and health status levels each year until death.

This means that in the case of uncertainty, we have to identify every possible health profile and measure the utility of each health profile before it is multiplied by the probability of that health profile. This would mean a much larger assessment task than for QALYs for the realistic case. The idea of HYEs is based on this idea of assessing health profiles rather than health states. HYEs were proposed as an alternative to QALYs by Mehrez & Gafni (1989,1991), after arguing that the underlying preference assumptions of QALYs are too restrictive.

10.3 HYEs And Individual Preferences

In order to investigate the theoretical properties of HYEs, we will use the same simple model as in the last section, where there are a number of possible health profiles (Q,T) that are described by a specific number of years (T) in a constant health state (Q) followed by death. It should be noted, however, that the definition and theoretical properties of HYEs are the same as those stated above for the case where the quality of life varies over time. HYEs were defined by Mehrez & Gafni (1989) as the number of years such that:

$$U(Q^*,HYEs) = U(Q,T), \qquad (5)$$

where HYEs is the number of years in full health (Q*) that is considered equivalent to the lifetime health profile U(Q,T). The definition of HYEs is illustrated in Figure 10 for 10 years in health state A. As can be seen from Figure 10, 5 years in full health is considered equivalent to the utility of 10 years in health state A, and the number of HYEs is 5 in this example.

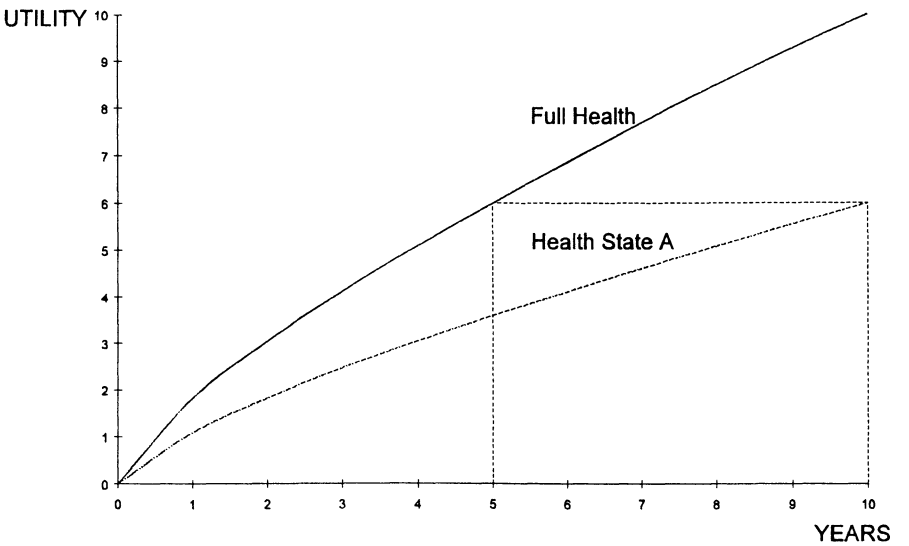

Figure 10. The definition of HYEs

The number of HYEs is the same as the number of healthy years in the time-trade-off method, and the time-trade-off method can thus be used to estimate the number of HYEs. The difference when using the time-trade-off method to construct HYEs is that the number of years in full health is used directly as the effectiveness measure, rather than using this number to construct the QALY weight.

The number of HYEs also has to be measured for every possible duration of time in a health state, whereas only one measurement of the QALY weight is carried out and this is assumed to be valid for any duration of time in the health state. If constant proportional trade-off holds, the number of QALYs based on the time-trade-off method and the number of HYEs will always be the same for health profiles with constant quality over time. This is easily seen in Figure 10 by estimating the number of HYEs for different time horizons in health state A. The number of HYEs increases by 0.5 for every year in health state A, and since the quality weight with the time-trade-off method is 0.5 the number of QALYs also increases by 0.5 for every year in the health state. If the quality varies over time constant proportional trade-off and additive value independence has to hold for QALYs based on the time-trade-off method and HYEs to be identical. Additive value independence means that the value of a health status in one period is independent of the health status in all other periods.

Mehrez & Gafni (1991) argued that it was not possible to use the time-trade-off method to measure HYEs, and instead proposed a two-stage procedure for measuring HYEs, with the argument that this would ensure that risk is incorporated in HYEs. It has been argued by Buckingham (1993), Culyer & Wagstaff (1993) and Johannesson et al (1993) that this two-stage procedure in theory yields results which will be identical to those obtained with a direct use of the time-trade-off method. The two-stage procedure for measuring HYEs is illustrated in Figures 11a and 11b for health state A during 10 years.

In the first stage, illustrated in Figure 11a, the standard gamble method is used to measure the utility of the health profile. As can be seen in the figure, the utility of the health profile is 6 on a scale between 0 (death) and 10 (full health), corresponding to a probability of 0.6 for 10 years in full health in the standard gamble method.

In the second stage of the two-stage procedure illustrated in Figure 11b, the utility of the health profile is set equal to the number of years in full health, which is 5 years. The number of HYEs for 10 years in health state A is thus 5, but this is exactly what is obtained by direct use of the time-trade-off method.

With the time-trade-off method, the number of years in full health that leads to the same utility as the assessed health profile is identified. In this example this is illustrated in Figure 10, where the number of HYEs is also 5. It is thus unnecessary to use the two-stage procedure to measure HYEs, since it only adds noise to the measurement.

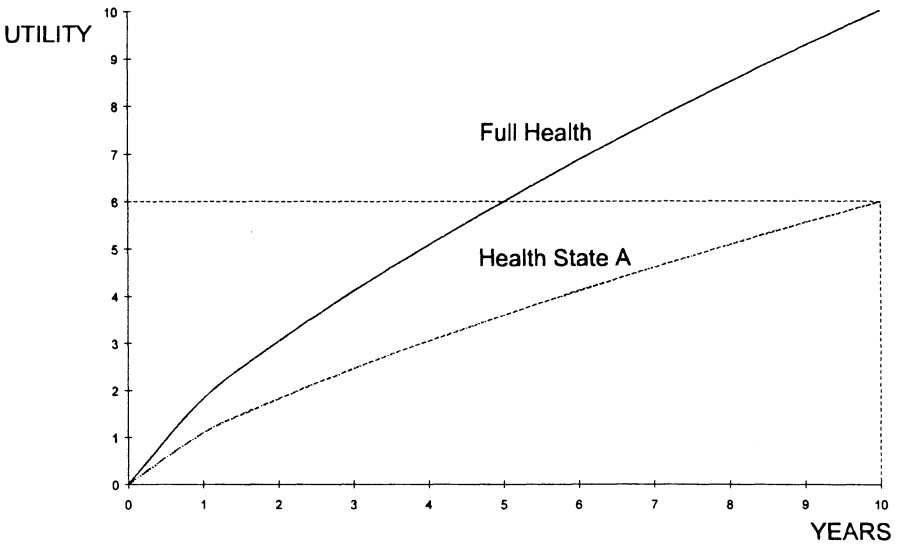

Figure 11a. Stage 1 in the two-stage procedure

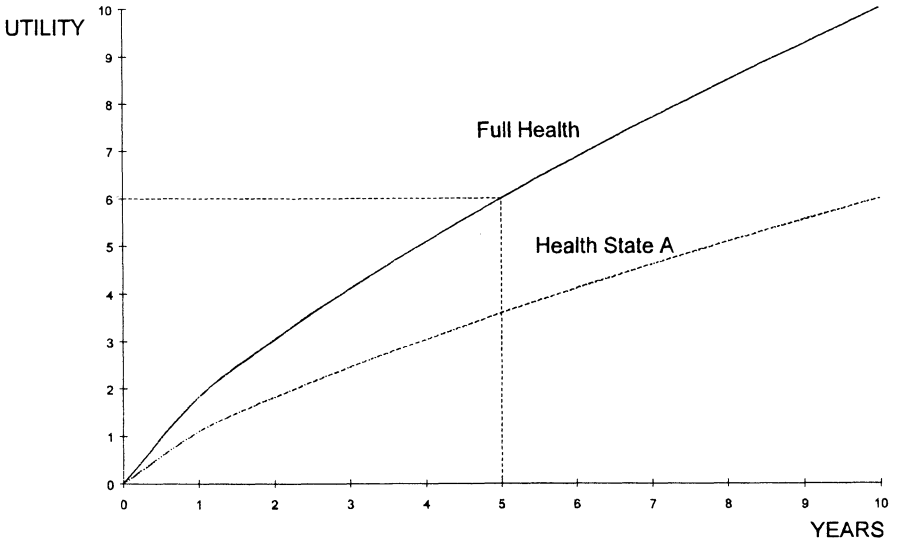

Figure 11b. Stage 2 in the two-stage procedure

The equivalence between the two-stage procedure and the time-trade-off method can also be shown more formally (Johannesson et al 1993). In the first stage of the two-stage procedure, the health profile (Q,T) is compared with a gamble involving a probability of immediate death (1-p) and a complementary probability (p) of full health (Q*) for the maximum possible remaining survival (T*). Denote this gamble SG(p;Q*,T*). The probability p is varied until the individual is indifferent between the two alternatives.

In the second stage, the gamble is at the indifference probability compared with some number of years in full health (Q*,HYEs). The number of years in full health is varied until the individual is indifferent. The two-stage procedure thus establishes that (Q,T)=SG(p;Q*,T*)=(Q*,HYEs).

The time trade-off-method compares the health profile (Q,T) with some number of years in full health (Q*,HYEs). The number of years in full health is varied until the individual is indifferent, i.e. it establishes that (Q,T)=(Q*,HYEs). If the number of HYEs were to differ between the two equalities, the basic axiom of transitivity would be violated. The only difference in theory between the measurement methods is that with the time-trade-off method, the indifference between (Q,T) and (Q*,HYEs) is established directly without an intervening gamble.

HYEs are defined as the number of years in full health, and for this to be a valid cardinal utility function, risk neutrality over healthy years has to be assumed (Johannesson et al 1993). Healthy years are not in general the same as utility, but if the utility function over healthy years is linear, the number of healthy years will be a valid cardinal utility function. This assumption is illustrated in Figure 12.

The difference in the assumptions for HYEs and QALYs is that for HYEs risk neutrality over healthy years is required, whereas for QALYs risk neutrality over life-years in all health states is required (plus additive utility independence with varying quality over time). As illustrated in Figure 12, the utility function over life-years in health states worse than full health can have any shape as long as the utility function over healthy years is linear.

This means that QALYs and HYEs will in theory coincide in two different cases. The first case is if additive utility independence and risk neutrality with respect to life-years in all health states holds and either the time-trade-off method or the standard gamble method is used to measure the quality weights. In this case both measures will be the same and they will both be valid cardinal utility functions.

The second case is if additive value independence and constant proportional tradeoff with respect to life-years holds and the time-trade-off method is used to measure the quality weights. In the second case neither QALYs or HYEs will be a valid cardinal utility function, but they will both be identical.

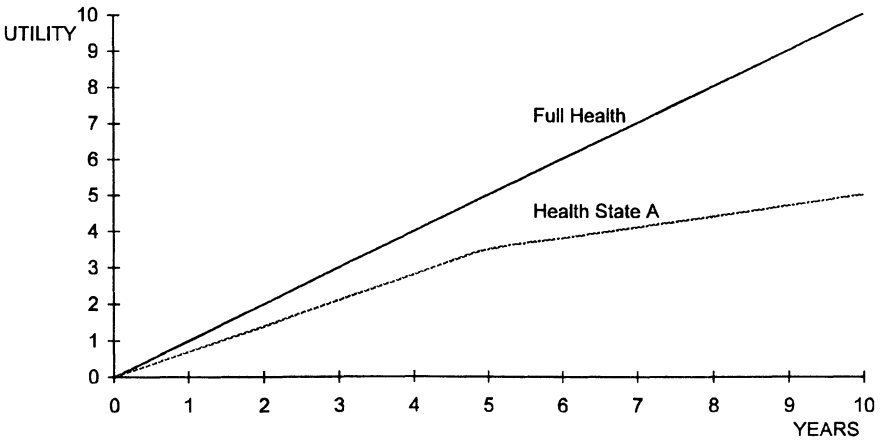

Figure 12. Risk neutrality over life-years in full health

The reason that it is important to investigate under what circumstances HYEs are a valid cardinal utility functions is that when risk is introduced (i.e. probabilities of different health states and length of life), the expected number of HYEs will rank probability distributions of health profiles according to individual preferences if the expected utility theory is valid. If HYEs are not a valid cardinal utility function, this property is not guaranteed. To estimate the expected number of HYEs is also the measure proposed by Mehrez & Gafni (1989,1991) in situations that involve risk (see for example the estimation in Gafni & Zylak (1990)). Johannesson et al (1993) showed, however, that it is possible to define an alternative measure, the certainty-equivalent HYEs, which will rank health profiles correctly in the case of uncertainty.

In order to show the difference between these measures it is necessary to introduce uncertainty, i.e. probabilities of different events. Assume that there are i possible health profiles (combinations of health status and length of life), each associated with a probability P_i (the probabilities sum to 1). The number of HYEs for each health profile is $HYEs_i$ and the utility level is U_i. The number of expected HYEs, then, is equal to:

$$\text{Expected HYEs} = \Sigma_i P_i * HYEs_i \qquad (6)$$

To get the number of expected HYEs for a probability distribution over health profiles, the probability of each possible health profile is multiplied by the number of HYEs for each health profile. The alternative is to directly assess the number of years of full health with certainty (the certainty-equivalent number of healthy years), which leads to the same utility as the probability distribution over health profiles. The certainty-equivalent HYEs (CE.HYEs) are defined by the following equality:

$$U(Q^*,CE.HYEs) = \Sigma_i P_i * U_i(Q,T) \qquad (7)$$

Note that the value of certainty-equivalent HYEs in equation (7) incorporates uncertainty with regard to length of life and health status, because the health profile is itself framed in terms of uncertainty. The risky health profile has to be framed as a probability distribution and compared with the equivalent number of years in full health with certainty. The distinction between certainty-equivalent HYEs and expected HYEs is similar to the distinction made between ex ante WTP and expected WTP in Chapter 3.

The difference between expected HYEs and certainty-equivalent HYEs is illustrated in Figure 13. In the figure it is assumed that an individual faces a 50% probability of 10 years in full health and a 50% probability of immediate death, and that we want to measure the number of HYEs for this risky health profile.

To get the number of expected HYEs, the probability of each outcome is multiplied by the number of HYEs for each outcome. Since the number of HYEs for 10 years in full health is 10 and the number of HYEs for immediate death is 0, the expected number of HYEs is 5 (0.5*10). The expected utility is equal to the utility of each outcome multiplied by the probability of each outcome, which is equal to 5 in this case (0.5*10+0.5*0).

The ranking problem with expected HYEs is easily seen if we assume that a treatment will give the individual 5 years in full health with certainty instead of the above gamble. 5 years in full health is also equal to 5 HYEs, but leads to a utility of 7. The number of expected HYEs will thus rank the "treatment" and "no treatment" options equally, in spite of the fact that the expected utility differs between the options. Note also that if the utility function over healthy years was linear in Figure 13, i.e. the assumption needed for HYEs to be a valid cardinal utility function, both options would lead to the same expected utility and the expected number of HYEs would correctly give both options the same ranking.

The number of certainty-equivalent HYEs in Figure 13 is defined as the number of years in full health that gives the same expected utility as a 50% chance of 10 years in full health. The expected utility is 5, leading to about 2.5 certainty-equivalent HYEs. The number of certainty-equivalent HYEs is the certainty-equivalent number of HYEs, and this measure will always rank options correctly as long as the utility

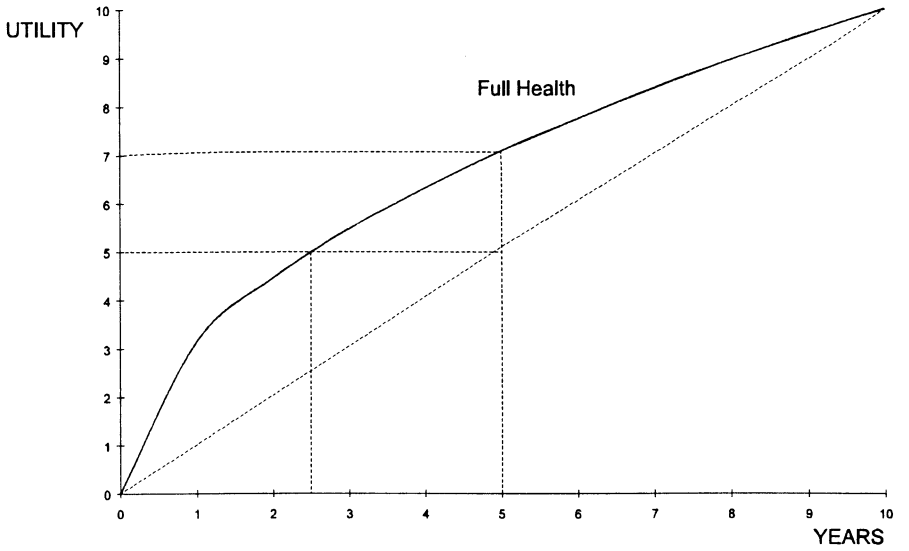

Figure 13. Certainty-equivalent HYEs VS expected HYEs.

increases with more years of life with certainty in full health. In the figure the "treatment" option is correctly ranked before the "no treatment" option, based on the number of certainty-equivalent HYEs.

It should be noted that certainty-equivalent HYEs are not a valid cardinal utility function, but this property is not needed since the measure is not used to estimate an expected value. There are two different approaches that in theory can be used to estimate the certainty-equivalent HYEs. The first approach is that in principle it should be possible to use the time-trade-off method, with the modification that the health profile to be assessed is framed in terms of uncertainty (i.e. a probability distribution). In terms of Equation (7) this means that the respondent compares the utility of the probability distribution over health profiles on the right-hand side of the equation with the utility of the number of years with certainty in full health on the left-hand side of the equation. The number of years with certainty in full health is then varied until the individual is indifferent between the left-hand side and the right-hand side of the equation, and this gives the certainty-equivalent HYEs.

Using certainty-equivalent HYEs measured directly by the time-trade-off method, in a cost-utility analysis of a treatment compared to no treatment, would mean that only two measurements have to be carried out, namely the certainty-equivalent HYEs with

and without the treatment, and the difference would give the gain in certainty-equivalent HYEs of the treatment. It is unclear, however, whether individuals can provide meaningful answers to this type of question.

An alternative approach for computing certainty-equivalent HYEs would be to estimate the utility (on a scale between death and maximal lifetime) of all health profiles using the standard gamble method (McNeil et al 1981). In terms of Figure 2, Alternative one would then be the health profile that the utility is measured for and Alternative 2 would be a gamble, with a probability P of being in full health for the maximal remaining lifetime and a complementary probability (1-P) of immediate death.

P is then varied until indifference is reached, and if the utility of immediate death is set equal to 0 the utility of the health profile is equal to P multiplied by the utility of full health for the maximal remaining lifetime (which can be set at any convenient number such as 1). Note that in this case the probability (P) will reflect both the reduced health status and the reduced life-years of the health profile compared to being in full health for the maximal remaining lifetime (thus the probability cannot be used as a quality weight in QALY calculations).

After the utility has been estimated for each possible health profile in this way, the expected utility of the probability distribution over health profiles can be estimated by multiplying the utility of each health profile by the probability of each health profile. The expected utility will rank probability distributions over health profiles according to individual preferences as long as the expected utility theory is valid. Note that the assumption of additive utility independence is not needed here, since the utilities of health profiles rather than health states are estimated.

The expected utility of a probability distribution over health profiles can then be converted to certainty-equivalent HYEs by estimating the utility function over healthy years. The utility function over healthy years can be estimated by using the standard gamble method in the same way as above, with maximal remaining lifetime and immediate death as the reference states (McNeil et al 1981). Alternative 1 in the standard gamble method will now be a different number of healthy years. By measuring the utility for the different number of healthy years between immediate death and maximal remaining lifetime, the utility function over healthy years can be estimated (the utility of different time horizons in full health may also have been estimated already, if these are part of the possible health profiles in the decision tree). By entering the expected utility of the probability distribution over health profiles in this function, the certainty-equivalent HYEs can then be inferred. It should be noted that this procedure differs from the method for estimating expected HYEs, since in the decision tree the expected utility is calculated instead of the expected number of HYEs. This approach suffers from the same practical problems as the method of estimating the expected number of HYEs in a decision tree, since a measurement of each possible health profile has to be carried out. If HYEs are used it may, however, be worthwhile

to go through the extra work of transforming expected utility to the certainty-equivalent HYEs in order to get a measure that is consistent with expected utility theory, rather than to estimate the expected number of HYEs in the decision tree as suggested by Mehrez & Gafni (1989,1991).

10.4 QALYs, HYEs And Discounting

In the above analysis no discounting of life-years was assumed for QALYs and HYEs. In most analyses using QALYs as the effectiveness measure, QALYs are discounted. If QALYs are discounted the underlying preference assumptions change. For QALYs to be a valid cardinal utility function, it now in addition to additive utility independence has to be assumed that risk neutrality holds with respect to discounted life-years in all health states rather than with respect to life-years.

In terms of Figure 9, which illustrates risk neutrality with respect to life-years in all health states, life-years on the horizontal axis should now be interpreted as discounted life-years (i.e. the utility function with respect to discounted life-years has to be linear in all health states).

If QALYs are discounted, the quality weight with the time-trade-off method cannot be estimated in the usual way by dividing the number of healthy years by the number of years in the assessed health state (Johannesson et al 1994). The quality weight has to be derived using the assumption of discounting. This means that the quality weight with the time-trade-off method has to be derived by dividing the number of discounted years in full health by the number of discounted years in the assessed health state (i.e. in terms of the section about the measurement of quality weights with the time-trade-off method, the weight is derived as X/T where both X and T are the discounted number of years).

For example, if the individual is indifferent between 10 years with arthritis and 7 years in full health, the quality weight without discounting would be estimated as 7/10, which is equal to 0.7. If life-years are discounted with a discount rate of 5%, the quality weight should be estimated as 5.7864/7.7212, where 5.7864 is 7 discounted years and 7.7212 is 10 discounted years (for the sake of simplicity it is assumed here that the first year is also discounted). The weight based on 5% discounting is equal to 0.75 in this example.

The quality weight with the standard gamble method does not have to be derived in a different way if life-years are discounted, since the time-horizon of the assessed health state and full health is the same. It has been noted in empirical studies of the quality weights with the time-trade-off method and the standard gamble method that the quality weight tends to be somewhat higher with the standard gamble method than with the time-trade-off method (see for example the cost-utility application below). This is usually attributed to risk aversion with respect to life-years, but an alternative

explanation could be discounting of life-years, since if the quality weight of the time-trade-off method is derived on the basis of discounting it increases, as seen in the example above.

If HYEs are discounted, the assumptions about when HYEs are a valid cardinal utility function change, so that risk neutrality with respect to discounted life-years in full health has to be assumed. In terms of Figure 12 the utility function with respect to life-years in full health now has to be linear, with discounted life-years on the horizontal axis of the figure.

If both QALYs and HYEs are discounted, the two measures will coincide in two different cases. The first case is if additive utility independence and risk-neutrality with respect to discounted life-years in all health states holds, and the quality weights are measured with the standard gamble method or the time-trade-off method (provided that the weights with the time-trade-off method are derived on the basis of discounting). In this case both QALYs and HYEs are valid cardinal utility functions.

The second case is if additive value independence and constant proportional tradeoff with respect to discounted life-years holds and the time-tradeoff method is used to measure the quality weights. In this case both measures are identical but none of the measures are a valid cardinal utility function.

10.5 Cost-Utility Analysis VS Cost-Benefit Analysis

Since cost-utility analysis is a special case of cost-effectiveness analysis, the same issues as were analysed in the chapter about cost-effectiveness analysis will also apply to cost-utility analysis, and the assumptions for the analysis are the same. As in cost-effectiveness analysis, the decision-maker approach to cost-utility analysis is commonly used. Since this approach is likely to lead to problems of suboptimisation, as discussed in the cost-effectiveness chapter, it will not be discussed further here.

Instead we feel that it is important to investigate the relationship between cost-utility analysis and cost-benefit analysis, given that the aim of both types of analysis is to include all costs and health effects (benefits) irrespective of to whom they accrue. Cost-utility analysis can then be viewed as a subset of cost-benefit analysis, where the aim of cost-utility analysis is to estimate the cost function for producing gained QALYs (we will use the term QALY in this section to reflect any type of outcome measure that incorporporates effects on both the quality and quantity of life, and we will also assume that it reflects individual preferences). In order to reach a decision based on cost-utility analysis, information is then needed about the WTP per gained QALY (i.e. the price per gained QALY).

In the standard cost-utility case this would mean that the difference between cost-benefit analysis and cost-utility analysis is that in cost-utility analysis, the WTP

per gained QALY is assumed to be the same for all individuals and for all sizes of the change in QALYs. As in the cost-effectiveness case, we will briefly investigate the relationship between cost-utility analysis and cost-benefit analysis when the WTP per QALY is assumed to be constant, and then look at the more realistic case where the WTP per gained QALY can vary between individuals and with the size of the change in QALYs.

We can then start with the case with a constant WTP per QALY. Figure 14 shows the marginal cost curve for producing gained QALYs for a particular population. It is assumed that all costs which are not included in the quality adjustment are included in the cost function. The marginal cost curve is constructed by adding new independent programmes in order of the incremental cost-effectiveness ratios, or by replacing mutually exclusive programmes with more effective ones. The marginal cost curve is drawn as a smooth function, even though a stepwise function as in the cost-effectiveness chapter may be more realistic. The slope of the cost function at different points represents the incremental cost-effectiveness ratios of different programmes.

The marginal WTP curve per gained QALY for the population has also been added to the figure and since we assumed that the marginal WTP is constant, the marginal benefit curve is just a straight horizontal line. The optimum for the population is the point where the marginal cost equals the marginal benefit, i.e. society will continue to implement programmes until the marginal cost per gained QALY equals the marginal benefit per gained QALY. In the figure this is the point where the marginal cost curve and the marginal WTP curve intersect, and the number of QALYS gained is given by the point QALYs* in the figure.

In this case, cost-utility analysis will lead to the same result as cost-benefit analysis, since according to cost-benefit analysis programmes should be implemented as long as benefits exceed costs (or equal costs). If the programmes as defined in Figure 14 are evaluated using cost-utility analysis, with a price per gained QALY that is equal to the constant marginal WTP per gained QALY in the figure, this will yield the same result as a cost-benefit analysis of the same programmes. Using cost-utility analysis in this case will also maximise the number of gained QALYs for the resources that will be spent, given the available information on costs and QALYs (and the price per gained QALY).

If the WTP per gained QALYs is constant and is the same for all individuals under all circumstances, cost-utility and cost-benefit analysis will thus yield the same result, i.e. both methods will recommend implementation of the same health programmes, provided that the WTP per gained QALY is used as the price in cost-utility analysis. This means, however, that for cost-utility analysis to be a useful tool we need information about the WTP per gained QALY to provide the method with a useful decision rule.

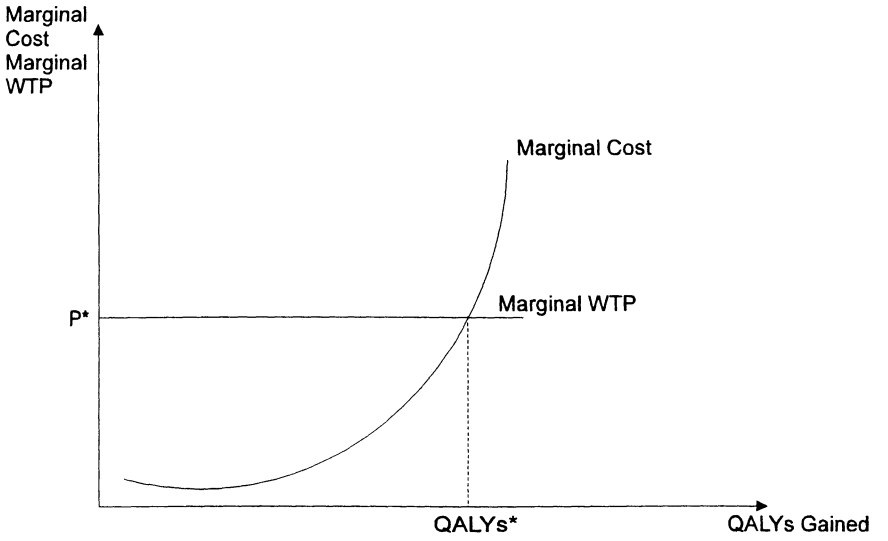

Figure 14. The marginal cost and marginal WTP of producing QALYs gained for a society with a constant marginal WTP

Of course the assumption of a constant WTP per gained QALY is not very realistic, and we will therefore also consider the case where the WTP per gained QALY varies, but cost-utility analysis is based on a constant WTP per gained QALY.

We can start by considering a single individual. The marginal cost curve for producing gained QALYs for the individual and the marginal WTP for gained QALYs for the individual are shown in Figure 15. The marginal cost curve can be interpreted as a number of increasingly more effective mutually exclusive programmes for the individual with increasing incremental cost-utility ratios. The curve is drawn smooth, on the basis of the assumption that every programme leads to a marginal cost increase.

The optimum for the individual is the point where the marginal WTP for gained QALYs equals the marginal cost for gained QALYs (denoted by P*, QALYs* in the figure). The marginal cost curve shows the marginal (incremental) cost-utility ratio of different health programmes, and if we use P* as the price per gained QALY we could implement increasingly more effective health programmes until the marginal cost-utility ratio equals this price.

Thus for the individual, cost-utility analysis based on the price P* will yield the same result as using cost-benefit analysis (cost-benefit analysis here being defined as the point where the marginal cost and marginal WTP curves intersect, i.e. the optimum as determined by the individual on a market). However, for non-marginal cost changes due to a programme, which would typically be the case, this would not necessarily apply, as was noted in the cost-effectiveness chapter (see Chapter 9 for details).

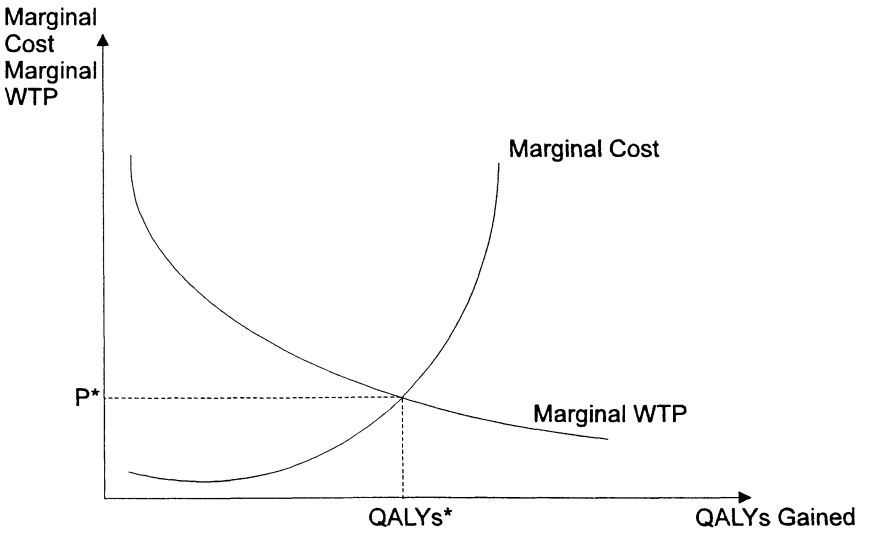

Figure 15. The marginal cost and marginal WTP of producing QALYs gained for an individual

If we then consider the realistic case with more than one individual and a varying WTP per gained QALY, it is straightforward to show that cost-utility analysis based on a constant price per gained QALY and cost-benefit analysis can yield different results. Even if we assume that we use the mean marginal WTP per gained QALY for further health improvements for each individual in society as the price in cost-utility analysis, there is no reason to believe that the WTP will be the same for different individuals.

We can consider the same example as in the cost-effectiveness chapter, but using gained QALYs as the effectiveness measure. Say that we are considering implementing two different independent health care programmes. One programme in

one patient group costs $100,000 and yields 5 gained QALYs, i.e. an incremental cost-utility ratio of $20,000. Another programme in another patient group costs $100,000 and yields 4 gained QALYs, i.e. an incremental cost-utility ratio of $25,000. The price that we use in our cost-utility analysis per gained QALY is $22,500.

According to the cost-utility analysis the first programme should be implemented, but not the second. If for instance the marginal WTP per gained QALY is $15,000 in the first patient group and $30,000 in the second patient group, a cost-benefit analysis would give the result that the second programme should be implemented rather than the first.

The WTP per gained QALY could vary in different patient groups. Another important source of variation is that the WTP per gained QALY can be assumed to vary, depending on the size of the gain in QALYs (i.e. the size of the health change). If cost-utility analysis is used to maximise QALYs, no distinction will be made between the size of the health improvements for different groups, e.g. an increase of 10 gained QALYs for one individual will be valued the same as an increase of one gained QALY each for ten individuals. Cost-benefit analysis, on the other hand, could be expected to give less weight to an increase in 10 QALYs for one individual than to one QALY each for 10 individuals, due to the decreasing marginal WTP with the size of the gain in QALYs.

It is thus clear that cost-benefit analysis and cost-utility analysis based on a constant WTP per gained QALY can produce different results. How important these differences would be in practice is an empirical issue. Cost-utility analysis may best be interpreted as a subset of a cost-benefit analysis, where the aim of the cost-utility analysis is to estimate the cost function of producing QALYs.

By combining the estimate of the cost function with the WTP per gained QALY, a cost-benefit analysis can then be carried out. The WTP per gained QALY could then be allowed to vary, depending on for instance the size of the change in QALYs for each individual (e.g. 0.01 gained QALYs per person versus 1 gained QALY per person).

However, from a normative viewpoint it could also be argued that it is appropriate to use cost-utility analysis based on a constant WTP per QALY. If QALYs are interpreted as a valid cardinal utility function then it could be argued that each additional QALY (i.e. utility unit) should be valued the same for everyone in order to maximise the gain in utility.

If QALYs are a von Neumann-Morgenstern utility function then this means that the function is unique up to a positive affine transformation, i.e. a+bQALYs (b>0) represent the same preferences as QALYs (cardinal utility functions will leave the rankings of absolute changes in utility unchanged by the transformation). In order to be able to aggregate QALYs, the constant b has to be fixed for each individual

(Bleichrodt 1995b). This assumption is sometimes referred to as cardinal unit comparability (Bleichrodt 1995b). Thus the possibility of aggregating QALYs in order to maximise utility for a population is based on the cardinal measurability of the utility functions of individuals, and cardinal unit comparability between the utility functions of individuals. It should be noted that the constant a does not have to be fixed for each individual, since if total utility is maximized the difference in utility between different social states is unaffected by the location of the utility function; this is because the location will cancel out in taking the difference in utility. The aggregation of total utilities is based on a utilitarian social welfare function (see the section about social welfare functions in Chapter 2). If, however, society is concerned not only about the total utility, but also about the distribution of utilities, then the constant a also has to be fixed for each individual and cardinal full comparability must be assumed (Bleichrodt 1995b). This would be the case if for instance a convex social welfare function is assumed (see the section about social welfare functions in Chapter 2).

In QALYs, the constant B is set to 1 for everyone so that one year in full health is given the same weight for everyone (or one discounted year in full health), implying that cardinal unit comparability is assumed for the QALYs of different individuals. It is not obvious that one year of (discounted) healthy life should be valued the same for everyone. Perhaps the utility of a life-year should vary with age for instance, or even differ between individuals since individuals seem to value health differently. Or perhaps b should be set to a value such that the remaining life is given the same utility for everyone.

In QALYs the constant a is also set to 0 for everyone, so that immediate death gives 0 QALYs (utility) for everyone. This means that it is assumed that immediate death is of the same utility for everyone. However, as noted above, this assumption does not have to be invoked in order to aggregate QALYs as long as a utilitarian social welfare function is used. The gain in QALYs of a programme will be unaffected by whether the QALYs for immediate death differ between individuals. However, if the gains in QALYs are weighted for different individuals, it is necessary to invoke this assumption.

It is clear that even if QALYs are valid cardinal utility functions, strong assumptions about the comparability of different individuals' QALYs have to be made in order to aggregate QALYs into a measure of social welfare (Bleichrodt 1995b; Boadway & Bruce 1984). Maximising QALYs means assuming a social welfare function that gives gained QALYs the same weight for everybody.

Even if QALYs could be aggregated into a measure of social welfare, it does not necessarily follow that QALYs should be maximised by a given amount of costs by setting a constant price on a QALY. To do this it also has to be assumed that the marginal utility of "costs" of health care programmes is constant across different health care programmes (i.e. that the marginal utility of income is the same for different financers of health care programmes). In a sense, costs should be converted

to utility or QALYs as well, in order to decide which health care programmes to carry out so as to maximise utility.

If QALYs is a valid cardinal utility function the WTP per additional QALY will decrease for the recipients of the health care programme, not because of decreasing marginal utility of QALYs but because of increasing marginal utility of income. Possibly this could be an argument for using the same WTP per gained QALY for different changes in the number of QALYs. These issues all have to do with the aggregation of preferences and social welfare functions, and it is unclear how QALYs should be aggregated and how they should be compared with costs. It may be possible to construct an argument that the WTP per gained QALY should be constant in cost-utility analysis, on the rationale that different WTP figures merely reflect different marginal utilities of income. If this approach is used, cost-benefit analysis and cost-utility analysis can yield different results.

At the moment it may be better to view cost-utility analysis as a subset of cost-benefit analysis, where the WTP per gained QALY has to be decided in order to reach a decision. If the WTP per gained QALY is shown to vary systematically under different circumstances, this could be incorporated in the analysis. Revealed preference and expressed preference approaches could be used to try and estimate the WTP per gained QALY to be used as the price in cost-utility analysis. Estimations of the value of risk reductions are often expressed as the value of a statistical life, but with information about the length and quality of life at risk it should be possible to convert these estimates into the value of a QALY (see chapter 5 for a conversion from the value of a statistical life to the value of a statistical life-year, on the basis of the results of the wage-risk studies). More empirical studies are needed to determine whether the value per gained QALY varies systematically, in order to take this variation into account in economic evaluations based on cost-utility analysis.

The implications of basing cost-utility analysis on a societal perspective and incorporating it within a cost-benefit framework are also that all costs which are not included in the adjustment for quality of life should be included in the costs of a health care programme. If it is assumed that all costs which are paid by the individuals who receive a health programme at a zero price for the programme are included in the quality adjustment, this leads to the conclusion that exactly the same costs should be included in a cost-utility analysis as in a cost-benefit analysis based on the WTP at a zero price for the programme (if the quality weight is measured for the treatment health state, it may also be argued that the part of the price that is paid by the individuals will be included in the quality adjustment).

The institutional factors in society would then also affect the cost concept in a cost-utility analysis, in the same way as for cost-benefit analysis. If for instance the health care costs and income losses of an individual are uninsured, it seems reasonable that an individual will take these differences into account when assessing the income available for consumption of non-health goods. For example, a move from the

assessed health state in a standard gamble or time-trade-off question to full health involves not only a move to a better health status, but also a move to an increased consumption of non-health goods.

If some costs are incorporated in the quality adjustment, it would then be double counting to include them also among the "costs". Whether some costs are included in the quality adjustment or not also depends on how the standard gamble or time-trade-off questions are framed; one example would be if it is stated that individuals should not include any differences in income between different health states in the questions. The issues are thus similar to the framing of a contingent valuation question and the cost concept in a cost-benefit analysis.

As far as we know, no empirical work has been carried out to test whether or not the quality weights using standard gamble or time-trade-off questions are affected by differences in income or health care costs between health states. The relationship between what is included in the quality adjustment and the cost concept of cost-utility analysis seems to be parallell to the relationship between what is included in the WTP and the cost concept of a cost-benefit analysis. Thus the section about the cost concept in cost-benefit analysis in Chapter 4 and the chapter about the estimation of costs in a cost-benefit analysis (Chapter 7) are also relevant for cost-utility analysis.

As in the cost-effectiveness case, we have to be careful in how we interpret cost-utility ratios if we do not follow the strict budget maximising approach, and only include those costs that fall on a specific budget. For instance, the cost-utility ratio can be affected by whether cost savings are included in the quality adjustment or as reduced costs.

Assume for example that an individual is willing to pay $20,000 for a treatment that improves his/her health status and reduces the health care costs paid for by the individual by $10,000. If the reduced health care costs are included in the quality adjustment, the gained QALYs are 1 for the individual, while if the reduced health care costs are not included in the quality adjustment the gained QALYs are 0.5.

This means that the individual is willing to pay $10,000 for the change in health status without taking into account the reduced health care costs (i.e. $10,000 for the 0.5 QALYs gained). Assume further that the treatment cost is $40,000. If the reduced health care costs are included in the quality adjustment, the cost per gained QALY becomes $40,000 (40,000/1; the treatment cost of $40,000 divided by the gained QALYs of 1).

If on the other hand the reduced health care costs are included as reduced costs, the cost per gained QALY becomes $60,000 ([40,000-10,000]/0.5; the net cost of $30,000 (i.e. the treatment cost minus the reduced health care costs) divided by the gain in QALYs of 0.5). Thus the cost per gained QALY differs, depending on whether the reduced health care costs are included as reduced costs or as increased QALYs.

Note that if the net benefits are estimated on the basis of the WTP per gained QALY and the cost per gained QALY in the two situations, the result will be the same. The WTP per gained QALY is $20,000 in the first case (20,000/1) and the cost per gained QALY is $40,000, leading to a net benefit per gained QALY of -20,000 (20,000-40,000), and since the gained QALYs are 1 the net benefits of the treatment are -$20,000. In the second case the WTP per gained QALY is $20,000 (10,000/0.5) and the cost per gained QALY is $60,000, leading to a net benefit of -$40,000 per gained QALY, and since the gained QALYs are 0.5 the net benefits of the treatment are -$20,000 (-40,000*0.5).

If an item, i.e. a cost change or a quality change, can be entered as either a cost change or a quality change, then the ratio between this item as a cost change and as a QALY change in absolute terms gives an implied WTP per QALY. In the above example this ratio was -$10,000/0.5 (i.e. the cost saving could be entered as either reduced costs of $10,000 or increased QALYs of 0.5). If this "WTP ratio" is below the cost-utility ratio, then the cost-utility ratio will be lowest if the item is entered so that the QALYs gained are as large as possible (since entering the item as a QALY change will have a larger relative effect on QALYs than the relative effect on costs will be if the item is entered as a cost change). If the "WTP ratio" is above the cost-utility ratio, then the cost utility ratio will be lowest if the item is entered so that the costs are as small as possible (since entering the item as a cost change will have a larger relative effect on costs than the relative effect on QALYs will be if the item is entered as a QALY change).

Due to the problems with using ratios, as mentioned in the cost-effectiveness chapter, the cost-utility ratios should not be interpreted as a ranking of projects even for independent programmes, but they should only be used to determine whether or not benefits exceed costs, i.e. whether the cost-utility ratio is at or below the price per QALY. This is all that is needed if we do not face a budget constraint, since all independent programmes with positive net benefits should be carried out according to the decision rule (and the mutually exclusive programme with the greatest net benefits should be carried out). If it is desirable to get an indication of the size of the net benefits, the net benefits of the whole programme can be estimated by using the price per QALY. See the cost-effectiveness chapter on how to deal with a real budget constraint.

10.6 A Cost-Utility Application

In this section we will describe in some detail a cost-utility analysis of hepatitis-B vaccination in Spain (Jönsson et al 1991). The alternatives compared in the study were screening and vaccination against hepatitis-B, versus no screening vaccination against hepatitis-B, for health care personnel. No separate analysis was carried out to determine whether or not screening should be carried out before vaccination, since screening is mandatory in public vaccination programmes for hepatitis-B in Spain.

For purposes of carrying out the cost-utility analysis, a decision-tree model was constructed. The starting point of the model was a cohort of health care personnel who could either be screened and vaccinated if marker negative (i.e. those individuals who did not carry hepatitis-B virus markers; this was assumed to cover 80% of the population of health care workers) or not screened and vaccinated. The susceptible population (i.e. the marker negative population) was exposed each year to an annual risk of hepatitis-B infection (referred to as the annual attack rate), and this annual risk of infection was assumed to be 1% in the base-case analysis without vaccination.

The annual attack rate was assumed to be reduced by 20% after the first dose of vaccine, by 50% after the second dose of vaccine, and by 90% after the third dose of vaccine. The compliance among the susceptible population for accepting vaccination was assumed to be 90% for the first dose, 85% for the second dose, and 80% for the third dose.

The hepatitis-B infections were divided into an acute infection stage and a chronic stage. In the acute stage the infections were divided into asymptomatic cases (assumed to be 50% of the infections), mild cases (assumed to be 30% of the infections), severe cases (assumed to be 19.8% of the infections) and fulminant cases (assumed to be 0.2% of the infections).

After the acute phase of the asymptomatic, mild and severe cases of infection it was assumed that 90% would return to full health and the remaining 10% would suffer from chronic impairments. For the fulminant cases, 70% were assumed to be fatal and of the remaining cases 90% were assumed to return to full health and 10% were assumed to suffer from chronic disease. The chronic cases were divided into mild chronic hepatitis (assumed to be 67% of the chronic cases) and severe chronic hepatitis (assumed to be 33% of the chronic cases).

The last stage of the decision tree was the survival rate for the different outcomes. The average age of infection was assumed to be 35 years for health care personnel and the average remaining life-expectancy was assumed to be 40 years for this group. For individuals who survived the acute infection with the infection not resulting in chronic hepatitis the life-expectancy was assumed to be 40 years (i.e. no loss of survival).

For 80% of the mild chronic cases the life-expectancy was assumed to be 40 years (i.e. no reduction) and for 20% of the cases the life-expectancy was assumed to be 30 years (i.e. a reduction of 10 years). For 50% of the severe chronic cases life-expectancy was assumed to be 40 years (i.e. no reduction) and for 50% of the cases life-expectancy was assumed to be 20 years (i.e. a reduction of 20 years). This led to a life-expectancy of 39.48 years after hepatitis-B infection, i.e. a reduction of 0.52 years on average for each infection. The clinical and epidemiological data in the decision tree were based on a variety of different sources and assumptions.

An adjustment for quality of life was also carried out in the model. No quality adjustment was carried out for the acute phase of the hepatitis-B infection. However, the life-years after mild chronic and severe chronic hepatitis were adjusted for quality of life. Two approaches were used to carry out the quality adjustment. In the first approach the quality weights were based on the Rosser/Williams quality of life matrix (Kind et al 1982). According to this approach, different health states are classified according to the two dimensions distress and disability. This gives a matrix, and health states in this matrix are given different weights where 1 is equal to full health and 0 is equal to death.

Mild and severe chronic hepatitis were placed in this matrix and mild chronic hepatitis was defined as no disability and mild distress, leading to a weight of 0.995; severe chronic hepatitis was defined as slight social disability and mild distress, leading to a weight of 0.986. These weights were then entered in the decision-tree, i.e. multiplied by the life-expectancy after mild and severe chronic hepatitis. With these weights the number of QALYs after hepatitis-B infection was estimated to be 39.45, which can be compared to 39.48 life-years without quality adjustment.

An alternative quality adjustment was also carried out. In this adjustment rating scale, time-trade-off and standard gamble was used to estimate the quality weights of mild and severe chronic hepatitis. These measurements were carried out on a small group of university students (11 Swedish students and 10 U.S. students). The results of these measurements are shown in Table 1.

Table 1. Quality weights of mild and severe chronic hepatitis

Method	**Health states**	
	Mild chronic hepatitis	**Severe chronic hepatitis**
Rating scale	0.60	0.30
Standard gamble	0.85	0.58
Time -trade-off	0.84	0.56
Mean of all methods	0.76	0.48

For mild chronic hepatitis the weight was 0.60 with rating scale, 0.85 with standard gamble, and 0.84 with time-trade-off. For severe chronic hepatitis the weight was 0.30 with rating scale, 0.58 with standard gamble, and 0.56 with time-trade-off. It is interesting to note that rating scale gives much lower weights than standard gamble and time-trade-off, which is also consistent with other studies (Read et al 1984; Hornberger et al 1992). The weights with standard gamble are somewhat higher than the weights with time-trade-off. Of course these weights are based on a very small sample, but the pattern of the weights with rating scale giving the lowest weight and standard gamble the highest weight is consistent with other studies (Read et al 1984; Hornberger et al 1992). It is also interesting to note the dramatic difference in the

weights based on the direct measurement and the weights of the Rosser/Williams matrix.

In this study the mean weight of the three methods was used to carry out the quality adjustment. This mean weight was 0.76 for mild chronic hepatitis and 0.48 for severe chronic hepatitis. The number of QALYs after hepatitis-B infection was estimated to be 38.35 on the basis of these weights (to be compared with the 39.45 QALYs based on the weights of the Rosser/Williams matrix and the 39.48 life-years without quality adjustment).

Using the above data, it is now possible to use the decision tree to estimate the number of gained QALYs for a reduction in the annual attack rate due to the screening and vaccination programme. This is done by estimating the expected number of QALYs in the decision tree for a cohort of health care personnel, before and after the reduction in the annual attack rate.

In order to carry out the cost-utility analysis, data about costs are also needed. The costs that were included in the analysis were the costs of screening, vaccination, hepatitis-B infection, and post-exposure prophylaxis. The health care costs of screening were estimated to be 950 pesetas per screened person, and the time cost for the screened personnel was estimated to be 300 pesetas (working time, since the screening was assumed to be carried out during working time), leading to a total cost of 1250 pesetas per person screened (it was not stated in the study in what year's prices the cost estimations were carried out).

The health care costs of the vaccination were estimated to be 1900 pesetas per dose, and the cost of lost time for the vaccinated personnel was estimated to be 300 pesetas per dose (working time, since the vaccination was assumed to be carried out during working time), leading to a total vaccination cost of 2200 pesetas per dose. To this was also added a cost of 50 pesetas per dose for the treatment of side-effects of the vaccination. The total vaccination cost for an individual taking 3 doses was thus estimated to be 6750 pesetas (including the costs of the side-effects).

The morbidity costs of hepatitis-B infection were divided into health care costs and lost production. The health care costs were estimated to be 34,000 pesetas per infection and the cost of lost production was estimated to be 153,000 pesetas per infection, leading to a total cost of 187,000 pesetas per infection (discounted at a 5% discount rate). These costs were used to estimate the expected morbidity costs without vaccination.

Since hepatitis-B infections which occur in spite of vaccination take a more benign course than those in non-vaccinated persons, the morbidity costs per infection after vaccination can be expected to be lower. The morbidity costs per infection after vaccination were estimated to be 77,000 pesetas, divided into health care costs of

12,000 pesetas and lost production costs of 65,000 pesetas (discounted at a 5% discount rate).

The cost of post-exposure prophylaxis was also estimated for unvaccinated persons. If unvaccinated persons are exposed to contaminated material (e.g. needles or blood), post-exposure prophylaxis is given to those persons. This cost was estimated to be 11,650 pesetas per person given post-exposure prophylaxis, and it was assumed that the risk of accidental exposure was 2% for unvaccinated persons for the time period of the analysis.

Finally, in order to carry out the cost-effectiveness analysis it was necessary to decide the number of years over which the vaccination would protect against infection. It was assumed that the vaccination would protect against infection for 5 years. The gained QALYs and the reduced costs of hepatitis-B infection due to the screening and vaccination programme based on the 1-year period in the decision tree (i.e. the decision tree was based on the annual attack rate and time was not explicitly included in the decision tree) was therefore multiplied by 5 to get the gained QALYs and the reduced costs due to the five years of protection against infection. Both costs and QALYs were discounted by 5%.

The results of the cost-utility analysis in terms of cost per gained QALY are shown in Table 2 for different annual attack rates (i.e. the annual risk of infection among the susceptible population). The cost per gained QALY obtained from the study is shown for both the quality adjustment based on the Rosser/Williams matrix and the quality adjustment based on the direct measurement using rating scale, standard gamble and time-trade-off.

Table 2. Cost per QALY gained of hepatitis-B vaccination

Attack rate %	**Quality adjustment**	
	Rosser/Williams matrix	Mean of RS, SG and TTO Measurement
0.1	3200000	1100000
0.2	1500000	500000
0.3	900000	200000
0.4	590000	200000
0.5	400000	130000
0.6	270000	91000
0.7	180000	63000
0.8	110000	38000
0.9	60000	20000
1.0	18000	6000

The first interesting observation from Table 2 is the big difference in the cost per gained QALY, depending on the approach used to determine the quality weights. The cost per gained QALY varies between 3,200,000 pesetas with an annual attack rate of 0.1% and 18,000 pesetas with an annual attack rate of 1.0% for the estimation based on the Rosser/Williams matrix. With the same attack rates, the cost per gained QALY varies between 1,100,000 pesetas and 6,000 pesetas for the estimation based on direct measurement of the quality weights.

In order to determine whether a cost per gained QALY was high or low, the gross domestic product (GDP) per capita in Spain of about 750,000 pesetas was used as a price per QALY. On the basis of this cut-off point it was concluded that health care personnel with an attack rate of 0.35% or higher should be screened and vaccinated if the Rosser/Williams quality adjustment was used, and that health care personnel with an attack rate of 0.15% or higher should be screened and vaccinated if the quality adjustment from direct measurements was used. Since the estimated average attack rate for health care personnel was 1.0%, it was concluded that the results indicated that screening and vaccination of health care personnel in Spain in general seemed to be cost-effective.

A sensitivity analysis was also carried out; this showed that in addition to the attack rate and the quality adjustment (as shown in Table 2), the result was also sensitive towards the duration of protection of the vaccine, the price of the vaccine and whether or not the lost production due to the hepatitis-B infection was included. If the lost production due to the hepatitis-B infection was not included, the attack rate where the cost per gained QALY equalled 750,000 pesetas changed to about 0.5% for the Rosser/Williams quality adjustment and to about 0.2% for the quality adjustment based on direct measurements.

The study is an interesting application of cost-utility analysis. The most interesting observation concerning the quality adjustment is perhaps the great difference between the weights based on the Rosser/Williams matrix and the weights based on the rating scale, standard gamble and time-trade-off measurements. This shows the danger of basing quality adjustments on arbitrary quality-of-life scales like the Rosser/Williams matrix.

The measurements of quality weights also show that the rating scale tends to produce much lower weights than the standard gamble method and the time-trade-off method. Thus, since the rating scale method lacks any theoretical foundation, it seems doubtful at the moment whether this method should be used to estimate the quality weights. It should also be noted, of course, that it is difficult to draw conclusions on the basis of direct measurements of a convenience sample of 21 students. However, the pattern of the results, in terms of the differences between the various direct measurement methods, is consistent with the results of other studies (Read et al 1984; Hornberger et al 1992).

The cost per QALY obtained in this study should be interpreted extremely carefully, however, since both the quality adjustment and many of the underlying assumptions and data are uncertain. For an appropriate measurement of the quality weights, a Spanish population should also have been used, since the possibility cannot be ruled out that the quality weights differ between countries due to cultural or institutional differences.

One issue that was not included in the study is also the likelihood that vaccination will reduce not only the probability of hepatitis-B infection for the vaccinated person, but also the probability of other persons being infected. Since this externality was not included, the cost-effectiveness of vaccination will be underestimated. The costs included in the study were the health care costs of the screening and vaccination programme and the time costs of the screened and vaccinated individuals. Thus all the relevant programme costs seem to have been included.

When calculating morbidity data, the health care costs and the lost production were included. The change in leisure time due to the morbidity was not included, but it seems reasonable to assume that this is included in the quality adjustment. An interesting issue is also whether some of the health care costs and lost production due to morbidity are also included in the quality adjustment. It was claimed by the authors that this was not the case for the direct measurements, since the students were asked to value the "pure" health effects and ignore earnings.

As mentioned above, the quality weights should be measured on a Spanish population. The question of whether or not some health care costs or lost production are included in the quality weights would then depend on how the questions are framed, and how health care costs and income losses are financed in Spain. No mortality costs were included and it could be argued that the external costs of increased survival should also have been included.

10.7 Conclusions

This chapter has considered cost-utility analysis of health programmes, which is a special form of cost-effectiveness analysis where effects on both the quantity and the quality of life are combined in one effectiveness measure. The most common such measure is QALYs. QALYs are constructed by assigning each health state a quality weight between 0 (dead) and 1 (full health), and this weight is assumed to be valid for the health state under all circumstances (i.e. independent of the number of years in the health state and the health states in other periods).

This is a practical assumption, but it also means that for QALYs to always represent individual preferences, very strong assumptions have to be made. It has to be assumed that risk neutrality with respect to (discounted) life-years holds in all health states and

that additive utility independence holds. In addition it also has to be assumed that the expected utility theory is valid.

In order to be able to relax the additive utility independence assumption of the QALY model, whole health profiles (scenarios) rather than health states have to be assessed. The assessment of the utility or value of all possible health profiles (i.e. all possible combinations of life-years and health states over time) independently is a major task compared to the QALY approach, since for realistic cases the number of possible scenarios is much greater than the number of possible health states.

The primary scenario-based measure proposed in the literature is HYEs, which is the number of years in full health that give the same utility as a health profile. However, HYEs are also based on relatively strong assumptions, since they assume risk neutrality with respect to (discounted) life-years in full health. It was shown that it was possible to define a measure, certainty-equivalent HYEs, which is consistent with individual preferences. One way of estimating this measure is to estimate the utility of each health profile and then estimate the expected utility of each treatment alternative in the decision tree. The expected utility can then be converted to certainty-equivalent HYEs by estimating the utility function over healthy years.

More empirical research is needed about the ranking properties of different measures (i.e. to what extent different measures rank probability distributions over health profiles according to individual preferences) and the practical feasibility of using different measures. At the moment, for practical applications of cost-utility analysis it seems most appropriate to use QALYs, where the quality weights are based on either the standard gamble method or the time-trade-off method.

In some applications it may be relevant to use scenario-based measures if the possible health profiles can be reduced to a small number. In such a case we would recommend estimating the utility of each possible health profile using the standard gamble method, and estimating the expected utility of the probability distribution over health profiles. The expected utility can then be converted to the certainty-equivalent number of HYEs. This appears to be better than estimating the HYEs of each health profile, since if the scenario approach is chosen it seems preferable to do the additional work of estimating the utility function over healthy years so as to get a consistent measure. However, in general it does not seem to be a feasible approch in practice.

The best way to view QALYs may be as an approximation of the expected number of HYEs, i.e. a measure of the expected health change, rather than as a utility. This is similar to using gained life-years as the outcome measure in cost-effectiveness analysis, since gained life-years can only be guaranteed to be consistent with individual preferences if individuals are risk neutral with respect to life-years and the quality is the same in all life-years (and the expected utility theory is valid).

The measures may then be some sort of health measures rather than utility measures, but for this to be useful the measures still have to be reasonably consistent with individual preferences (i.e. if individuals in general prefer treatments with fewer QALYs to treatments with more QALYs, the approach seem to be meaningless).

We also argued that cost-utility analysis, like cost-effectiveness analysis, can be viewed as a subset of cost-benefit analysis where the cost function of producing QALYs is estimated. In order to determine whether or not a health care programme should be implemented, the cost estimations have to be supplemented with information about the price per QALY that we are willing to pay. This also means that cost-utility analysis should be based on the same cost concept as cost-benefit analysis, and all costs that are not included in the quality adjustments should be included among the costs.

REFERENCES

Bleichrodt H. QALYs and HYEs: under what conditions are they equivalent? Journal of Health Economics 1995a;14:17-37.

Bleichrodt H. Health utility indices and equity considerations. Mimeo, 1995b.

Boadway RW, Bruce N. Welfare economics. Oxford: Blackwell, 1984.

Boyle MH, Torrance GW, Sinclair JC, Horwood SP. Economic evaluation of neonatal intensive care of very-low-birth-weight infants. New England Journal of Medicine 1983;308:1330-7.

Broome J. Qalys. Journal of Public Economics 1993;50:149-167.

Buchanan JM. Cost and choice: an inquiry in economic theory. Chicago: University of Chicago Press, 1969.

Buckingham K. A note on HYE (healthy years equivalent). Journal of Health Economics 1993;12:301-309.

Bush JW, Chen M, Patrick DL. Cost-effectiveness using a health status index: analysis of the New York State PKU screening program. In (Berg R, ed.) Health Status Indexes. Chicago: Hospital Research and Educational Trust, 1973.

Culyer AJ, Wagstaff A. QALYs versus HYEs. Journal of Health Economics 1993;12:311-323.

Gafni A, Birch S, Mehrez A. Economics, health and health economics: HYEs versus QALYs. Journal of Health Economics 1993;12:325-339.

Gafni A, Zylak CJ. Ionic versus nonionic contrast media: a burden or a bargain? Canadian Medical Association Journal 1990;140:475-478.

Hornberger JC, Redelmeier DA, Peterson J. Variability among methods to assess patients' well-being and consequent effect on a cost-effectiveness analysis. Journal of Clinical Epidemiology 1992;45:505-512.

Johannesson M, Pliskin JS, Weinstein MC. Are healthy-years equivalents an improvement over quality-adjusted life-years? Medical Decision Making 1993;13:281-286.

Johannesson M, Pliskin JS, Weinstein MC. A note on QALYs, time tradeoff, and discounting. Medical Decision Making 1994;14:188-193.

Jönsson B, Horisberger B, Bruguera M, Matter L. Cost-benefit analysis of hepatitis-B vaccination. International Journal of Technology Assessment in Health Care 1991;7:379-402.

Kind P, Rosser R, Williams A. Valuation of quality of life: some psychometric evidence. In Jones-Lee MW (Ed.). The value of life and safety. Amsterdam: Elsevier/North-Holland, 1982.

Klarman HE, Francis JOS, Rosenthal G. Cost-effectiveness analysis applied to the treatment of chronic renal disease. Medical Care 1968;6:48-54.

Loomes G, McKenzie L. The use of QALYs in health care decision making. Social Science and Medicine 1989;28:299-308.

McNeil BJ, Weichselbaum R, Pauker SG. Speech and survival: tradeoffs between quality and quantity of life in laryngeal cancer. New England Journal of Medicine 1981;305:982-987.

Mehrez A. Gafni A. Quality adjusted life years, utility theory, and healthy-years equivalents. Medical Decision Making 1989;9:142-149.

Mehrez A, Gafni A. The healthy-years equivalents: how to measure them using the standard gamble approach. Medical Decision Making 1991;11:140-146.

Mehrez A, Gafni A. Healthy-years equivalents versus quality-adjusted life years: in pursuit of progress. Medical Decision Making 1993;13:287-292.

Miyamoto JM, Eraker SA. Parametric models of the utility of survival duration: tests of axioms in a generic utility framework. Organizational Behavior and Human Decision Processes 1989;44:162-202.

Miyamoto JM, Eraker SA. Parameter estimates for a QALY utility model. Medical Decision Making 1985;5:191-213.

von Neumann J, Morgenstern O. Theory of Games and Economic Behavior. Princeton: Princeton University Press, 1947.

Pliskin JS, Shepard DS, Weinstein MC. Utility functions for life years and health status. Operations Research 1980;28:206-24.

Pratt JW. Risk aversion in the small and in the large. Econometrica 1964;32:122-136.

Read JL, Quinn RJ, Berwick DM, Fineberg HV, Weinstein MC. Preference for health outcomes: comparison of assessment methods. Medical Decision Making 1984;4:315-329.

Schoemaker PJH. The expected utility model: its variants, purposes, evidence, and limitations. Journal of Economic Literature 1982;20:529-563.

Sonnenberg FA, Beck RJ. Markov models in medical decision making: a practical guide. Medical Decision Making 1993;13:322-338.

Torrance GW. Measurement of health state utilities for economic appraisal: a review. Journal of Health Economics 1986;5:1-30.

Torrance GW, Boyle MH, Horwood SP. Application of multi-attribute utility theory to measure social preferences for health states. Operations Research 1982;30:1043-1069.

Torrance GW, Thomas WH, Sackett DL. A utility maximization model for evaluation of health care programmes. Health Services Research 1972;7:118-133.

Weinstein MC, Stason WB. Hypertension: a Policy Perspective. Cambridge, MA: Harvard University Press, 1976.

Williams A. Economics of coronary artery bypass grafting. British Medical Journal 1985;291:326-329.

11. Economic Evaluation and Policy Making

The aim of economic evaluations of health care programmes is to serve as an aid to decisions and to affect policy making. If economic evaluations of health care programmes are not going to have any impact on decisions about the allocation of resources to health care programmes, this is a meaningless activity. In this chapter we discuss different decision and policy situations where economic evaluations of health care programmes could potentially be used. We discuss economic evaluations as an aid to: the development of treatment guidelines, decisions within health care organizations, introduction of new medical technologies, reimbursement decisions, and pricing decisions. We end the chapter with some conclusions about the relationship between economic evaluations and policy making.

11.1 Treatment Guidelines

In the medical field it is common practice to issue treatment guidelines about how different diseases should be managed. These guidelines may be issued directly by the medical profession, e.g. a college of physicians, or by the public authorities, e.g. the department of health. Even if the guidelines are issued by a public authority, the medical profession would normally play some part in the development of the guidelines. The guidelines could also be developed on different levels, such as for a whole country or only for a specific hospital. The aim of these guidelines is to affect the treatment of different diseases; for example, they may involve recommendations about which drug to choose for a specific patient group.

An example of treatment guidelines is the guidelines issued in Sweden by the National Board of Health and Welfare for different diseases. For example, in the guidelines for the management of hypertension it is stated which classes of drugs should be used as a first-line treatment for hypertension. According to the current guidelines diuretics, beta-blockers, calcium-antagonists and ACE-inhibitors are recommended for first-line treatment. The choice of drugs for first-line treatment is based solely on medical considerations, but it would be possible to base the recommendations on economic evaluations instead (i.e. on both medical consequences and cost consequences).

Thus the incorporation of economic evaluations into the development of these guidelines could be a way of affecting decision-making about health care resources. This would mean that the guidelines would be based on economic evaluations rather than on the health effects alone. The advantage of basing treatment guidelines on both costs and health effects is of course that this could lead to a more efficient use of resources to treat a disease, i.e. we could get more health effects (measured as WTP or some measure of health effects) for the same amount of resources or the same amount of health effects for less resources.

The reluctance of the medical profession to incorporate economic considerations into treatment guidelines could be due to the fact that the concept of opportunity costs is not well understood; it is not recognised that resources devoted to a treatment programme have alternative uses. Another argument, which is sometimes advanced as a reason for not considering costs in treatment decisions, is that every treatment which improves health should be given, irrespective of the costs. However, the logical conclusion of this argument could be that we should devote all of society's resources to health care, which shows the unreasonableness of the argument.

Hitherto, economic evaluations do not seem to have played an important role in the development of treatment guidelines. One reason for this could be the reluctance of the medical profession to incorporate cost considerations into the development of guidelines. Another reason is that economic evaluation is a relatively new field, where the methods are still being developed; there may also have been a lack of high-quality economic evaluations to use as a basis for the formulation of treatment guidelines. It does also appear that there is increasing interest in the use of economic evaluations in the development of treatment guidelines.

The development of treatment guidelines appears to be a field where economic evaluations could potentially make an important contribution. For the treatment guidelines to affect actual decisions in the health care field, they have to have a high degree of acceptance within the medical profession and among other health care decision-makers. If the guidelines are developed in cooperation between medical experts and experts on economic evaluation, it may be possible to achieve a high degree of acceptance. An advantage of using economic evaluations in the development of treatment guidelines is also that economic evaluations can be introduced stepwise into the process, as the field matures and the quality of the evaluations hopefully improves. The focus on the economic evaluation part in the guidelines can then also vary from case to case, depending on the availability of studies in different fields and the quality of the evidence.

Treatment guidelines are a form of dissemination of information, where the aim is to affect voluntary decisions made by different actors by providing them with better and updated information. This is an important difference compared to some of the fields below, where the allocation of resources is affected directly through rules and regulations.

Since treatment guidelines are based on dissemination of information and the different actors can use their discretion as to whether or not they will follow the guidelines, a critical factor is of course the extent to which treatment guidelines have any impact on actual decisions. If treatment guidelines have no impact on health care decisions, the issue of whether or not economic evaluations are incorporated in the process becomes unimportant and the whole process becomes meaningless. Research about how treatment guidelines affect clinical practice is thus of great importance, in order to

determine the usefulness of treatment guidelines as a way affecting the allocation of resources to different treatments.

The impact of treatment guidelines on clinical practice can be assumed to depend on the credibility of the organization or authority that issues the guidelines, and the credibility of the people involved in the preparation of the guidelines. The impact of treatment guidelines on clinical practice can also be expected to depend on the structure of the health care system and the incentives embodied in that system; for example, if the guidelines are based on economic evaluations then their impact on clinical practice can be assumed to depend on the incentives present for taking into account economic evaluations in decisions affecting clinical practice (see below in the next section about decision making within health care organizations).

11.2 Decision Making Within Health Care Organizations

Another potential field for economic evaluations is as a basis for decision making within health care organizations. This is related to the development of treatment guidelines, since the aim of treatment guidelines is to affect the decisions made in health care organizations. However, economic evaluations could also be used directly by a health care organization such as a hospital or an HMO as an aid for decisions about alternative treatment strategies.

The structure of the health care system will affect the incentives for using economic evaluations within health care organizations. It will also affect the costs that will enter into the decisions. If for instance the costs of drugs are not part of the budget of a health care provider, e.g. the costs of drugs are reimbursed by the state, the health care provider has no incentives to consider the costs of drugs when chosing between different treatment strategies. Thus the fact that different costs fall on different budgets increases the risk of suboptimisation.

Both the incentives for the patients and the incentives for the health care provider seem to be important. If for instance different providers compete for insured patients, then if any increased costs can be passed on to a third-party payer such as an insurance company, an employer or the state there can be incentives to only compete with quality, and to ignore the differences in costs between alternative treatment strategies (except perhaps costs paid by the patients, see below). Thus the health care provider would have no incentives to consider economic evaluations.

An interesting case is the HMO type of model, where the HMO receives a sum of money for each individual and then contracts with different providers. If the patients can choose freely between different HMOs, the HMO has incentives for considering the effects on individuals' health and the costs that are paid by the members of the HMO such as leisure costs, user charges and travel costs. This is because otherwise they may lose members (of course if some members consume more health care paid

for by the HMO than the sum of money that is paid to the HMO, the HMO will have an incentive to get rid of these patients, but we will assume here that the HMO cannot discriminate between different members).

However, if the HMO can raise its fee and pass it on to a third-party payer, such as an employer or the state, there would be little incentive to consider cost differences between alternative treatment options. If the membership fee is paid by the members of the HMO, or if the HMO cannot raise the fee or cannot control the size of the fee, it would have incentives to take into account all the costs that are paid for by the HMO. In such a case the HMO would have incentives to use economic evaluations that incorporate all costs falling on the HMO and all costs and health effects that are paid by the members. An example of such a system would be one where the state pays and determines the size of the membership fees for the HMOs through taxes, and the patients can choose between different HMOs.

It also has to be assumed that the HMO will go bankrupt if it fails to break even, so that for instance the state does not guarantee the existence of the HMO. Unless the HMO or the patient pays for all the cost consequences of different treatments, however, there will be no incentives for taking all costs into account in economic evaluations. The typical case would probably be one where the HMO covers all health care costs (or expenditures if the prices used do not reflect costs), except those paid for by the patients. The biggest problem would then be that the HMO has no incentive to consider insured production changes for its members. However, it may be possible to integrate these costs also into the budget of the HMO, so as to provide incentives that reflect a societal perspective.

In a private insurance system where the individuals pay the insurance premiums directly, there would also be incentives for the insurance companies to use economic evaluations which take into account the costs paid by the insurance companies and the individuals with insurance policies. However, as in the HMO case above, there may be a risk of suboptimisations since there is no incentive to take into account production changes that are covered by insurance.

In all health care systems there is also an incentive for the ultimate payers of health care such as the employers or the state (i.e. in the end the consumers of products and the taxpayers) to consider both costs and health effects, but they may have a limited ability to affect or control the decisions of health care organizations. It can thus be concluded that economic evaluations can be a useful tool for decisions within health care organizations, but that the incentives for the organizations to use economic evaluations are crucial.

11.3 Introduction Of New Medical Technology

A possible role for economic evaluations is also as a basis for decisions concerning the introduction of new medical technology. For individual firms, the incentives to use economic evaluations will largely depend on the institutional arrangement of the potential market. For example, in health care systems where there are no incentives to consider costs in decisions about treatment strategies, the providers of new technologies have only incentives to try to develop more effective technologies, without being concerned about the costs (Weisbrod 1991). This issue is thus related to the discussion above about the incentives for health care organizations to use economic evaluations.

However, the state may also serve as a gatekeeper for new technology through some state authority. The prime example of this in the health field is probably the approval of the sales of new drugs. In most countries, new drugs have to be approved by some drug approval agency before they can be marketed and sold, e.g. in the U.S. new drugs have to be approved by the FDA (the Food and Drug Administration).

A possible use of economic evaluation would then be as a basis for decisions about the approval of new drugs (or other new technologies that have to be approved). It could for instance be a requirement that the incremental cost per QALY has to equal or fall short of a specific price per QALY in order to be approved. The drug companies could for instance be required to submit an economic evaluation when they submit applications for the approval of new drugs. The drug approval agency would then base its decision not only on safety and the health effects of the drug, but also on the costs associated with using the drug. Drugs that are not cost-effective would then not be approved and would not be available on the market. This would make it possible to ensure that drugs that are not cost-effective are not used on the market.

However, there are several weaknesses associated with the use of economic evaluations as a basis for deciding whether or not a drug should be approved. The cost-effectiveness of a new drug depends on the patient group in which it is used, and this means that a decision about whether or not to approve a new drug is an inefficient way to achieve a cost-effective use of drugs.

A new drug for the treatment of high blood pressure (hypertension), which leads to a slightly greater reduction in the risk of coronary heart disease than existing alternative drugs, can be used as an example to show this. This drug can be cost-effective in some patient groups at a very high initial risk of coronary heart disease, such as elderly men with severe hypertension, but not cost-effective in some patient groups with relatively low initial risk of coronary heart disease, such as young women with mild hypertension.

The problem here is to make sure that the drug is used in the patient group where it is cost-effective, e.g. elderly men with severe hypertension, and not in other patient

groups where it is not cost-effective, e.g. younger women with mild hypertension. However, this cannot be achieved by a general decision for or against approval. One possibility is that it should be sufficient to show that a drug is cost-effective in one patient group for approval. For drugs that are approved on this criterion, it then remains to target their use towards cost-effective patient groups. In the hypertension example above the drug would be approved, and the problem of making sure that it is only used in cost-effective patient groups would remain.

Another alternative decision rule is that the decision about approval should be based on the overall cost-effectiveness of the drug in all patient groups where it is expected to be used. However, such a decision rule would mean that some drugs would not be available for some patient groups where they are cost-effective. If it is assumed that overall in all patient groups the above hypertension drug is not cost-effective and therefore is not approved, this would mean that the drug would not be available for men with severe hypertension where it is cost-effective. If it is impossible to use any other mechanism to target the use of drugs to the cost-effective patient groups, this seems to be the appropriate decision rule, since the only decision is whether or not the drug should be used at all.

One possibility would of course be that drugs are only approved for specific indications, e.g. only for elderly men with severe hypertension. However, in clinical practice it may be difficult to ensure that the drug is only used for that indication. As new data become available, the estimation and the approval decision would also have to be reconsidered, which may be impractical. The requirement for economic evaluations before any approval decisions could also delay such decisions and thus also the time until a new drug becomes available for patients. However, the timing of an approval as such is an issue which may be suitable for economic evaluations. In general a longer approval time means that some patients will not receive some health benefits, but it also reduces the risk of side-effects of the drug.

The approval time is thus an issue that seems suitable for economic evaluations, where the potential health effects of a shorter approval time have to be weighed against the increased risk of side-effects and the expected cost consequences of the drug. It is probable that this trade-off differs between drugs for different types of disease. A new drug for a life-threatening disease like aids, where there is a lack of effective treatments, may motivate a shorter approval time than a new drug for the treatment of high blood pressure where effective alternative drugs already are available. If the latter type of economic trade-off was incorporated, this could increase the flexibility of the approval system. Thus it appears that the use of economic evaluations as a basis for deciding whether or not a new drug should be approved may not be the most useful role for economic evaluations; on the other hand, economic evaluations may be a useful way of analysing the optimal approval time for different types of drug.

11.4 Reimbursement Decisions

Decisions about public reimbursement for health care technologies is another field where economic evaluations could be used. One possibility would be to use economic evaluations to decide the optimal subsidy for goods that improve the health of individuals (e.g. healthy food), or the optimal tax for goods with hazardous health effects (e.g. smoking). The subsidy or tax should then equal the externalities that arise due to the consumption of the good. The externalities equal the external costs plus the altruistic externality as defined in Chapters 3 and 4. An example of a study which attempts to estimate the external costs is the one on optimal taxes on tobacco and alcohol (Manning et al 1991).

In some fields such as health care, the decisions regarding subsidisation tend to concentrate on whether or not some technology should be subsidised, rather than determining the size of the subsidy. In publicly funded health care systems, for instance, decisions have to be taken about whether or not treatments should be reimbursed (i.e. included in the public insurance system). In such cases it may be difficult to vary the subsidy between different technologies and between different patient groups according to the way in which the size of the externality varies (i.e. the size of the subsidy will usually vary, not according to the externality but according to the cost of the technology, since normally the user charge would tend to be constant between technologies). It may for instance be difficult to vary the subsidy for different types of drug, depending on the size of the externality.

The decision about whether or not a particular health care technology should be publicly reimbursed appears to be an issue where economic evaluations could be a useful aid. Economic evaluations could then be used to determine whether a particular treatment should be reimbursed. For example, it might be decided that only treatments with a cost per QALY at or below a specific price would be reimbursed.

Such a system, however, is similar to the decision about the approval of new drugs discussed above, and exhibits some of the same difficulties. The cost-effectiveness of a treatment can be expected to vary in different patient groups, and it is not clear which patient group should be used as a basis for the decision about reimbursement. If a treatment is cost-effective in a selective patient group, this can motivate reimbursement of the treatment. However, if the treatment is reimbursed for this group then it may also be used in other patient groups.

The problem is that the issue is seldom whether or not a treatment should be used at all, but rather in which patient groups it should be used. However, if it is made compulsory to use economic evaluations for decisions about either approval or reimbursement of treatments, this could be an instrument for sorting out treatments that are not cost-effective in any patient group. If it is impossible to use any other mechanism to target treatments towards cost-effective treatment groups, then it also appears that the decision about reimbursement should be based on the overall

cost-effectiveness of the treatment (i.e. the only decision that can be made is whether or not the treatment should be used within the public health care system).

A difference between the decision about approval and the decision about reimbursement is that if a treatment is not reimbursed, it will still be available for those who are willing to pay the full price. It could be argued that it is efficient when only those who are willing to pay use a treatment, but this is not the case if the treatment leads to externalities. Since alternative treatments may also be reimbursed, this also creates a distortion in the choices of individuals. For example, if treatment A costs $2,000 and treatment B costs $1,500 and treatment B is fully reimbursed whereas treatment A receives no subsidy, an individual has to be willing to pay $2,000 more to get treatment A rather than B, even though the cost difference is only $500, assuming that the consumption externality is also the same for both treatments.

There is a way to get around the problem that the cost-effectiveness varies in different patient groups, by making it possible for a treatment to be reimbursed in some patient groups but not in other patient groups. Such a system would be more flexible and would increase the possibility of targeting treatments towards cost-effective patient groups using economic evaluations. However, the administration of such a system may be difficult. In some cases it could also create perverse incentives, e.g. to increase one's blood pressure in order to receive a subsidy.

The area where economic evaluations have received most attention as a basis for reimbursement decisions is that of drugs. Since January 1993, Australia has required economic evaluations in applications for new drugs that are to be included in the national drug reimbursement scheme and thereby be subsidised by the state. Australia then became the first country to require economic evaluations in applications for drug reimbursements.

In Australia, in order to obtain reimbursement, prescription drugs must be included in the Pharmaceutical Benefits Scheme (PBS), operated by the federal government. For a drug to be listed on the PBS, an application must be made to the Pharmaceutical Benefits Advisory Committee (PBAC), which makes recommendations to the federal government about the listing of new drugs. The new requirement has its roots in a 1987 amendment to the National Health Act, which required the PBAC to consider both costs and effectiveness in recommendations about the listing of new drugs on the PBS.

After the legislative change, a group of experts was engaged to assist in implementation of the new policy. This resulted in draft guidelines, issued by the federal government, for the preparation of submissions to the PBAC (Department of Health, Housing, and Community Services Commonwealth of Australia 1990; Evans et al 1990; Henry 1992). The guidelines were revised once and the inclusion of an economic evaluation in all submissions to the PBAC was made mandatory in January 1993. Economic evaluations are required only for new drugs. If the PBAC

recommends listing it will also pass on satisfactory economic evaluations to a separate pricing authority (the Pharmaceutical Benefits Pricing Authority (PBPA)).

It is interesting to examine the Australian guidelines from a methodological perspective. It is stated in the guidelines that the reimbursement decision and thus the economic evaluations should be based on a societal perspective. However, in the original version it was stated that changes in market production of the patients should not be included among the costs. In the revised version this was changed to a statement that if production changes were included among the costs, it had to be demonstrated that these production changes were real; for instance, it has to be demonstrated that production changes were not compensated for by changes in unemployment. It was also stated that the result should be presented both with and without production changes included.

The level of evidence needed to include production changes is thus stricter than for other costs, and it is not clear why different standards are employed for different types of costs. Furthermore, no distinction is made between production losses due to morbidity and mortality, which would have been appropriate (see Chapter 4 about the definition of the resource consequences of health care programmes).

It is also stated in the guidelines that new drugs should be compared with the drugs they are likely to replace in practice. However, it is often not clear which drug a new drug will replace. If for instance a new drug for the treatment of high blood pressure is introduced, it is impossible to know which of the existing drugs this drug is likely to replace; there are a number of different drug classes such as diuretics, beta-blockers, ace-inhibitors and calcium-antagonists, and also several different drugs within each class that are currently in use, and it is not clear which of these drugs is likely to be replaced by a new drug (in reality it would probably replace a number of different drugs).

In the guidelines, it is recommended that cost-effectiveness analysis should be used to carry out the economic evaluations (unless the health effects are identical for two alternatives, which makes it sufficient to carry out a cost-minimisation analysis). They also give cautious support for the development and use of cost-utility analysis with QALYs as the outcome measure, but note that there is inadequate experience with cost-utility analysis and QALYs to recommend it as a standard in current analyses.

It is also recommended that cost-benefit analysis should not be used in the economic evaluations. However, cost-benefit analysis is defined as the human-capital approach in the guidelines, and thus no distinction is made between a proper cost-benefit analysis and an analysis based on the human-capital approach (see Chapter 4 for the difference between these approaches).

The use of intermediate effectiveness measures, such as the change in blood pressure, in cost-effectiveness analysis is also encouraged in the guidelines. However, as noted

in the chapter about cost-effectiveness analysis (Chapter 9), the use of intermediate effectiveness measures is problematic. Intermediate outcome measures enable very limited comparisons of cost-effectiveness, and it is also extremely difficult to interpret the results of a cost-effectiveness analysis using intermediate effectiveness measures. For example, what does a specific cost per mm Hg reduction in blood pressure mean?

In order to use intermediate effectiveness measures, they must also be clearly related to some health outcome, e.g. the risk of heart attacks, since it seems to be a dubious goal to maximise for instance the blood pressure reduction per se. Furthermore, every unit of the intermediate effectiveness measure should be associated with the same change in the health outcome of interest, e.g. a percentage unit reduction in the risk of heart attacks. However, if such a relation is known then it seems better to use the actual health outcome as the effectiveness unit. The use of intermediate effectiveness measures can therefore not be recommended.

As a decision rule in cost-effectiveness analysis it is recommended that PBS should develop "yardsticks" to compare the cost-effectiveness of different drugs. It is unclear exactly what "yardsticks" means in the guidelines, but if this is to be used as a decision rule it means that the WTP (or price) per unit of effectiveness has to be determined for every effectiveness unit used (i.e. both for every intermediate measure used in submissions to the PBAC and for all other effectiveness units such as life-years and gained QALYs used in submissions to the PBAC).

A system similar to that introduced in Australia has also been implemented in the province of Ontario in Canada (Ontario Ministry of Health 1994), and draft guidelines exist for the whole of Canada (Detsky 1993). In Canada each province has a drug benefit programme which subsidizes the use of drugs for its residents (Detsky 1993). The drugs that are reimbursed are listed on different "formularies" in each province. In Ontario, the Minister of Health makes decisions about which drugs should be included in the Ontario Drug Benefit Formulary, after advice from the DQTC, which considers effectiveness, safety and cost in its recommendations.

According to the Canadian guidelines, a societal perspective should be adopted for economic evalutions, but it is also stated that the results should be presented both with all costs included and with only the costs funded by the provinces included (Detsky 1993). The guidelines further state that new drugs should be compared with the cheapest alternative treatment, but that comparisons with other alternative treatments may also be appropriate. The guidelines propose using cost-utility analysis with QALYs as outcome measure, but they also encourage experimentation with other approaches, such as cost-benefit analysis where the benefits are based on the willingness to pay of individuals.

A system for assessing the quality of the economic evaluations is also included in the Canadian guidelines. A checklist of questions is provided against which the economic evaluations are to be assessed, addressing for instance the quality of the cost and

effectiveness data used in the economic evaluation. The guidelines also recognize that if the cost per QALY is measured in the economic evaluations, then it is necessary to determine the willingness to pay (i.e. the price) per QALY that society is willing to pay in order to reach a recommendation based on the economic evaluation.

Tentative thresholds for the price per QALY were included in the initial version of the Ontario guidelines, giving ranges for what was considered to be a high or a low cost per QALY (Ontario Ministry of Health 1991). Studies were divided into three categories based on the cost per QALY. It was stated that the evidence was strong for the adoption and appropriate utilization of treatments with a cost per QALY below $20,000 (Canadian dollars in 1990 prices), and that there was moderate evidence for the adoption and appropriate utilization of treatments with a cost per QALY between about $20,000 and $100,000 (Canadian dollars). Finally, it was stated that the evidence was weak for the adoption and appropriate utilization of treatments with a cost per QALY above $100,000 (Canadian dollars).

The ranges provided were based on a review of existing economic evaluations (Laupacis et al 1992), rather than on the willingness to pay of individuals. These ranges for the price per QALY were removed from the revision of the guidelines, with the argument that the decision-maker has to decide the appropriate price per QALY to be used (Detsky 1993).

There are some important methodological differences between the Australian and the Canadian guidelines. The Canadian guidelines for instance incorporate a scheme for assessing the quality of the economic evaluations, whereas in the Australian guidelines it is not clear how the quality assessment will be dealt with. The Australian guidelines also encourage the use of intermediate effectiveness measures, while the Canadian guidelines do not.

From a policy perspective the Australian and Canadian guidelines are similar. In both countries the registration, reimbursement and pricing of drugs are separated, and economic evaluations are primarily intended as an aid to decisions about reimbursement. In Canada, however, prices are set before the decision about reimbursement, whereas in Australia the reimbursement decision precedes the pricing decision. Furthermore, reimbursement decisions are made on the provincial level in Canada and on the national level in Australia.

In both the Australian and the Canadian guidelines, it is unclear how flexible the system is intended to be when it comes to differentiating between the reimbursement status of the same drug in different patient groups (see the discussion above about the importance of a flexible system for the decision about reimbursement, if it is to be an effective tool to target drugs towards cost-effective patient groups).

According to a recent survey of health economists in Europe, economic evaluations are currently used in a number of European countries as an input into decisions about

drug reimbursement (Drummond et al 1993). An example of this is the reimbursement for the two cholesterol-lowering drugs simvastatin and pravastatin in the Netherlands; this was based on an economic evaluation, and was restricted to individuals with high cholesterol levels who had one or more additional risk factors for cardiovascular disease (van Hout & Rutten 1993). This is also an example of a flexible reimbursement decision where the decision is not whether or not the drug should be reimbursed overall, but in what patient groups it should be reimbursed. However, formal requirements for economic evaluations as an aid in reimbursement decisions about drugs or other health care technologies do not yet exist in any European country.

There have also been discussions about the use of economic evaluations as an aid for decisions about the reimbursement of other health care technologies than drugs. An interesting example of this is the Health Care Financing Administration's (HCFA) proposed regulation to add cost-effectiveness as a criteria for medicare coverage decisions in the U.S. (Health Care Financing Administration 1989). These regulations were never finalized, but they still provide an interesting example (Neumann & Johannesson 1994).

HCFA argued that considerations of cost were relevant in deciding whether health care technologies should be covered by Medicare. HCFA further explained that cost-effectiveness would only be one of several potential factors to be considered in coverage decisions, and that HCFA would not necessarily consider cost-effectiveness in every coverage decision. HCFA also proposed the analytic steps that should be followed in order to carry out the analyses. The steps included: considering relevant alternative technologies to which current interventions would be compared; identifying all relevant costs expected from the intervention; and considering non-quantifiable factors. HCFA also stated that it would carry out further studies on how to collect and analyse primary and secondary data for cost-effectiveness (Health Care Financing Administration 1989).

In fact, the regulations proposed by the HCFA to add cost-effectiveness as a criteria for Medicare coverage were never finalised and implemented; this shows that introducing economic evaluations as a basis for reimbursement decisions can be sensitive, since it implies making the trade-off between improved health and reduced consumption of other goods explicit. For decision-makers it may be less sensitive to be as implicit as possible about this trade-off.

This section shows that economic evaluations may be useful as an aid to decisions about reimbursement of health care services, and that economic evaluations are already used for this purpose, with the prime example being the requirement for economic evaluations in applications for public drug reimbursement in Australia. However, since the cost-effectiveness of almost all health care technologies varies widely in different patient groups, it is crucial that the reimbursement system should be flexible, so that reimbursement for any given health care technology can be granted

for some patient groups but not for others. Otherwise the reimbursement decision can probably only serve as a gate-keeper, where health care technologies that are not cost-effective in any patient group are not reimbursed.

11.5 Pricing Decisions

One more possible area where economic evaluations may be used is that of pricing decisions, where a public payer or authority can affect the prices of health care technologies. The prime example here is probably also drugs, since the prices of drugs are regulated in many countries.

The reason for the regulation of drug markets is the specific characteristics of drugs and the market for drugs. New drugs (i.e. new chemical entities) are protected by patent for a number of years, which creates a temporary monopoly for the producer of a patent-protected drug. The patent protection is a way of providing incentives for drug companies to invest in research and development for new drugs.

Furthermore, the consumer of prescription drugs is often not very price sensitive, since the costs of prescription drugs are often covered by some form of insurance. To make sure that the prices of drugs are not set unreasonably high, the prices are often regulated or determined in negotiations between the producers and some public authority responsible for setting the prices of drugs. The prices of drugs may also be tied to public reimbursement, in the sense that the drug will only be reimbursed if the price is not set above a specific amount.

From an economic viewpoint, a distinction should be made between the societal optimal drug prices and the drug prices that should be used in an economic evaluation of drugs. The relevant price to use in an economic evaluation is the opportunity cost of producing and administering the drug, and the costs of research and development for the drug constitute sunk costs that should not be included.

However, this price is not the optimal drug price, because if drug prices were set equal to this price there would be no incentives for investing in research and development for new drugs (i.e. it would be the same as having no patent protection). The optimal prices of drugs should be set at such a level that the optimal investment in research and development of new drugs results. This is then a form of cost-benefit judgement, and the optimal level is the point where the marginal cost of research and development of new drugs equals the expected marginal benefit of research and development of new drugs (i.e. the marginal benefit is equal to the marginal willingness to pay for the new drug minus the marginal cost of producing and administering the drug). Thus economic evaluations may be used to analyse the optimal drug prices (or the optimal time of patent protection, which determines the time a company can expect to use a price that exceeds the marginal cost).

It is not, however, without its problems to use economic evaluations of current drugs as a basis for determining the prices of those drugs. As mentioned above, the appropriate price to use in an economic evaluation is the cost of producing and administering the drug (i.e. the opportunity cost of using the drug), and this cost is the same irrespective of the price of the drug. A higher price than the true cost then represents a transfer from the consumers of the drug to the owners of the drug companies. See also Chapter 7 in the section about programme costs for a discussion about the cost of drugs to be used in an economic evaluation.

On the other hand, if drugs are imported then it could be argued that the price represents the true opportunity cost for the country that imports the drug, since the import has to be paid by exports to the same amount. In such a case the optimal strategy for the country seems to be to set prices on imported drugs as low as possible; however, this may not be optimal if the price of imported drugs has an impact on the prices in other countries of drugs that were exported from the original country. Given that the price is assumed to reflect the true cost for a society, an economic evaluation can show the price at which a drug will be cost-effective in different patient groups (e.g. the price at which the cost per QALY will be equal to the willingness to pay per QALY). In a sense, this price shows the maximum price a company can charge for a drug to be cost-effective in a specific patient group. The cost-effectiveness is also likely to vary in different patient groups, and the issue then arises as to which patient group the price should be based on.

From the drug companies' point of view, economic evaluations can be a useful tool in decisions about pricing. Economic evaluations can be used to defend prices of patented drugs that may otherwise have been perceived as unreasonably high. Economic evaluations could also be used to set the price so that the drug becomes "cost-effective" compared to other competing drugs for the same patients. Economic evaluations which show that a drug is cost-effective can also be used in the marketing of the drug to increase the sales. These factors together with the use of economic evaluations in reimbursement decisions explain the greatly increased interest in economic evaluations shown by the drug companies.

To conclude this section: it seems as if there can also be a role for economic evaluations in pricing decisions, especially pricing decisions for new drugs. The issue is, however, complicated by the fact that the price may not be the relevant opportunity cost to use in the economic evaluation.

11.6 Conclusions

In this chapter we have examined a number of policy situations where economic evaluations could potentially be used. One potential role for economic evaluations of health care is in the formulation of treatment guidelines for different diseases. For this

to be useful the guidelines have to have an impact on actual decision-making, and the relationship between treatment guidelines and clinical practice is unclear.

Economic evaluations of health care could also potentially be used as a basis for decisions about the approval and reimbursement of medical technologies. Of these two decision situations, economic evaluations may be most useful as a basis for decisions concerning reimbursement. Reimbursement decisions also constitute the health care field where the use of economic evaluations as a decision tool has received most attention and seems to be most common. This is perhaps especially the case for drugs, where Australia has been the first country to require economic evaluations in applications for public drug reimbursement.

Even though the interest in incorporating both costs and health effects into public decisions concerning health and safety seems to be increasing, the impact of economic evaluations on actual decisions is still largely unclear. It also has to be realised that the structure of for instance the health care system will be of great importance for the incentives to carry out and use economic evaluations.

REFERENCES

Department of Health, Housing, and Community Services, Commonwealth of Australia. Draft guidelines for the pharmaceutical industry on preparation of submissions to the Pharmaceutical benefits committee, including submissions involving economic analyses. Canberra, Australia: Department of Health, Housing, and Community Services, 1990.

Detsky AS. Guidelines for economic analysis of pharmaceutical products: a draft document for Ontario and Canada. PharmacoEconomics 1993;3:354-361.

Drummond MF, Rutten F, Brenna A, et al. Economic evaluation of pharmaceuticals: a European perspective. PharmacoEconomics 1993;4:173-186.

Evans D, Freund D, Dittus R, et al. The use of economic analysis as a basis for inclusion of pharmaceutical products on the pharmaceutical benefits scheme. Canberra, Australia: Department of Health, Housing, and Community Services, 1990.

Health Care Financing Administration. Medicare Program: criteria and procedures for making medical services coverage decisions that relate to health care technology. Federal Register 1989;54:4302-4317.

Henry D. Economic analysis as an aid to subsidization decisions: the development of Australian guidelines for pharmaceuticals. PharmacoEconomics 1992;1:54-67.

van Hout B, Rutten F. Economic appraisal of health technology in the European Community. In Schubert F (ed.). Proceedings of a Canadian collaborative workshop on pharmacoeconomics. Princeton: Excerpta Medica Inc, 1993:8-13.

Laupacis A, Feeny D, Detsky AS, Tugwell PX. How attractive does a new technology have to be to warrant adoption and utilization? Tentative guidelines for using clinical and economic evaluations. Canadian Medical Association Journal 1992;146:473-481.

Manning WG, Keeler EB, Newhouse JP, Sloss EM, Wasserman J. The costs of poor health habits. Cambridge MA.: Harvard University Press, 1991.

Neumann PJ, Johannesson M. From principle to public policy: using cost-effectiveness analysis. Health Affairs 1994;13:206-214.

Ontario Ministry of Health. Guidelines for preparation of economic analysis to be included in submission to the Drug Programs Branch for listing in the Ontario Drug Benefit Formulary/Comparative Drug Index. Toronto: Ministry of Health, 1991.

Ontario Ministry of Health. Ontario guidelines for economic analysis of pharmaceutical products. Toronto: Ministry of Health, 1994.

Weisbrod B. The health care quadrilemma: an essay on technological change, insurance, quality of care, and cost containment. Journal of Economic Literature 1991;29:523-552.

12. CONCLUDING REMARKS

This book has presented the theory and methods of economic evaluations of health care programmes. When a new health improving technology becomes available, the relevant question is whether the increased costs of the health care technology are compensated for by increased health effects and cost savings elsewhere in the economy, so that benefits exceed costs.

The aim of economic evaluation is to answer that question. In theory the usefulness of economic evaluations is also quite obvious, since they enable us to use our limited resources so as to get as much health or welfare as possible. The need for economic evaluations then depends on two issues. Firstly, if it is possible to carry out economic evaluations in a valid and reliable way. Secondly, if it is possible to use the results of these economic evaluations to affect decision-making. These two issues are related to the methods of economic evaluation and the policy implications of economic evaluations. The main concern of this book has been the methods of economic evaluation.

The methods of economic evaluation are being continuously developed. What started as analyses based on the human-capital approach has now been developed to include cost-effectiveness analysis, cost-utility analysis and cost-benefit analysis. When two alternative health care programmes are associated with the same health effects the analysis is simple, since it is enough to estimate the costs to see which health care programme leads to the lowest costs.

However, in the standard case the health care programme with the best health effects will also be associated with the greatest costs, and in such a case the increased health effects have to be weighed against the increased costs. This is because the increased resources have an alternative use, and it is possible that the value of the benefits of that alternative use will exceed the value of the health care programme.

One approach to this trade-off is to carry out a cost-effectiveness analysis, and for instance estimate the cost per gained life-year of the health care programme. A more refined approach is to carry out a cost-utility analysis and calculate the cost per gained QALY. This will increase the number of possible comparisons. The cost per gained life-year or QALY essentially shows the different available investment possibilities to improve the health of the population. Using only the information about the cost per gained life-year or QALY for different health care programmes, it is impossible to determine which health care programmes should be implemented.

The rationale for cost-effectiveness/utility analysis is often said to be that the health effects should be maximised for a given budget. However, the use of a specific real-world budget to maximise the health effects would lead to suboptimisation, since costs outside the budget would be ignored, so that this is a dubious approach for a societal economic evaluation. Alternatively, however, the price that society is willing to pay

per gained life-year or QALY can be used as the decision rule, which will lead to the maximisation of the health effects for a given amount of resources (with resources being defined as the resources that are included among the costs in a cost-effectiveness/utility analysis).

Such an approach would make it possible to include all the costs of a health care programme. To use cost-effectiveness/utility analysis it is then necessary to have some information about how much society is willing to pay in order to gain a life-year or a QALY. This is the type of question that cost-benefit analysis deals with, where both the costs and the health effects are expressed in monetary terms. The health effects are measured in terms of willingness to pay, which reflects the amount of consumption of other goods and services that individuals are willing to forego. Cost-effectiveness analysis and cost-utility analysis based on a specified price per gained life-year or QALY can then be interpreted as a cost-benefit analysis, where the cost per gained life-year or QALY is assumed to be constant with the size of the health change, and the same for everyone.

Since the cost per gained life-year or QALY is not likely to be constant, this means that cost-benefit analysis and cost-effectiveness/utility analysis based on a constant price can yield different results. However, by varying the WTP per gained life-year or QALY under different circumstances the assumption of a constant price could be relaxed. This means that cost-effectiveness/utility analysis can be interpreted as an estimation of the cost function to produce life-years or QALYs, and in order to make a decision the price per life-year or QALY has to be determined. This would make cost-effectiveness/utility analysis a subset of cost-benefit analysis.

However, some people would also argue on normative grounds that the price per gained life-year or QALY should be constant and the same for everyone. It could for instance be argued that the variations in WTP per gained life-year or QALY reflect differences in the marginal utility of income, and if these differences are corrected for the utility of life-years or QALYs gained would be the same for everyone. These issues are related to the aggregation of individual preferences for social decisions and the social welfare function of a society (see Chapter 2). In practice it is unclear how much the WTP varies for gained life-years or QALYs, and what the importance would be of assuming a constant WTP versus a WTP that differs according to the circumstances.

One approach is of course to carry out a direct cost-benefit analysis as well, without first quantifying the health effects in terms of gained life-years or QALYs. Irrespective of whether this approach is followed, or whether we estimate the cost per gained life-year or QALY and combine this with the WTP per gained life-year or QALY, the way in which we obtain information about the willingness to pay for health effects becomes a crucial issue.

There are basically two approaches to estimating the WTP for health changes, namely revealed preference and expressed preference. Most of the work on revealed preference

has been carried out on job risks, estimating the wage-risk tradeoff. This has yielded important information about the value per statistical life, but the approach also has some problems and it is unclear to what extent the valuations of job risks can be extrapolated to other populations, other types of health risk, and other risk levels and risk changes. One important issue seems to be to find new areas and health risks where the revealed preference framwework may be applied, such as the purchase of healthy food or non-prescription drugs.

The alternative approach of expressed preference involves using the contingent valuation method to assess the WTP for health changes in surveys. The great advantage of this approach is its flexibility and the ability to directly get at the desired WTP for a health change. The crucial issue with respect to contingent valuation is of course the extent to which hypothetical choices mimic real behaviour, and this issue is still unsettled.

It is important to stress the experimental nature of many of the methods used for economic evaluations of health care programmes, such as the measurement of QALYs and the measurement of WTP. It is therefore important that the approach can continue to be developed without imposing too many restrictions on the precise way in which economic evaluations should be carried out or what type of economic evaluation should be used in decision-making. Thus it should not be expected that there will always be consensus about how to carry out an economic evaluation. It is also important to stress that the main problems of many economic evaluations are not the economics part of the analysis, but instead the weak epidemiological and medical data that are used to estimate the health effects.

The approach of taking into account both health effects and the costs of health care programmes has gained wider acceptance over time, and it seems clear that in spite of the uncertainties encountered in economic evaluation, this approach can already provide useful information about the costs and benefits of different health care programmes. This was shown here for instance, in some of the applications of economic evaluations that were included in order to illustrate the different methods.

The key issue then becomes the question of the extent to which economic evaluations have an impact on policy making and decisions concerning health care programmes. This would seem to depend upon the will of different real-world decision-makers to actually use economic evaluations, and the laws and regulations that are used to allocate resources to health care programmes. The will of the decision-makers is of course also a function of the organization and incentives embodied in the system of which the specific decision-maker is a part. For example, the way in which health care is organized and financed is likely to affect the incentives for different decision-makers within the health care system to initiate and use economic evaluations. An increased use of economic evaluations can then be achieved in different ways, e.g. by a law or regulation that requires the use of economic evaluations in reimbursement decisions,

or a change in the health care system that increases the incentives for taking into account both health effects and costs in health care decisions.

INDEX

Developments in Health Economics and Public Policy

1. P. Zweifel and H.E. Frech III (eds.): *Health Economics Worldwide.* 1992 ISBN 0-7923-1219-8
2. P. Zweifel: *Bonus Options in Health Insurance.* 1992 ISBN 0-7923-1722-X
3. J.R.G. Butler: *Hospital Cost Analysis.* 1995 ISBN 0-7923-3247-4
4. M. Johannesson: *Theory and Methods of Economic Evaluation of Health Care.* 1996 ISBN 0-7923-4037-X

KLUWER ACADEMIC PUBLISHERS – DORDRECHT / BOSTON / LONDON