OXFORD

Edited by Michael Drummond and Alistair McGuire

economic evaluation in health care

merging theory with practice

OHE

Economic Evaluation in Health Care

Merging theory with practice

Michael Drummond
University of York

Alistair McGuire
City University, London

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford OX2 6DP

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in

Oxford New York

Athens Auckland Bangkok Bogotá Buenos Aires
Cape Town Chennai Dar es Salaam Delhi Florence Hong Kong Istanbul
Karachi Kolkata Kuala Lumpur Madrid Melbourne Mexico City Mumbai
Nairobi Paris São Paulo Shanghai Singapore Taipei Tokyo Toronto Warsaw

and associated companies in Berlin Ibadan

Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries

Published in the United States
by Oxford University Press Inc., New York

First published 2001

A catalogue record for this title is available from

the British Library

Library of Congress Cataloguing in Publication Data

Drummond, M. F.
Economic evaluation in health care: merging theory with practice/
Michael Drummond, Alistair McGuire.
Includes bibliographical references and index.
1. Medical economics. I. McGuire, Alistair. II. Title.

RA410.5 .D768 2001 338.4'33621–dc21 2001033912

ISBN 0 19 263177 2 (Hbk) ISBN 0 19 263176 4 (Pbk)

Typeset in Minion
by Florence Production Ltd, Stoodleigh, Devon
Printed in Great Britain
on acid-free paper by
Biddles Ltd., Guildford & King's Lynn

Foreword

Economic evaluation is playing an increasing role in the allocation of health care resources. The National Institute for Clinical Excellence (NICE) in the UK is, perhaps, the best known of a number of recent European initiatives. These follow in the path of the Pharmaceutical Benefits Advisory Committee in Australia, the Canadian Co-ordinating Centre for Health Technology Assessment, and Provincial Formulary Committees in Canada, in seeking to make formal assessments of the value of health technologies. Health Maintenance Organisations in the USA are also showing greater interest in the cost-effectiveness of treatments, particularly new pharmaceuticals, as shown by the Association of Managed Care Pharmacists' development of guidelines for economic submissions. At local level in many health care systems, budget holders in hospitals and doctors' offices are seeking to identify whether more expensive but more effective treatments represent value for money.

This partner book to the Drummond *et al. Methods for the economic evaluation of health care programmes* is designed to give those involved in planning, undertaking, and reviewing economic evaluations of health technologies a clear understanding of the theoretical underpinning of the science of economic evaluation and of how to apply theory to practice. In particular it examines:

- the value judgements and theoretical assumptions that underpin different outcome measures and which need to be understood by users of these tools;
- options for measuring resource use and for pricing such use, including the treatment of learning curve effects, informal medical care, unrelated medical costs, and productivity effects;
- choice of decision analytic modelling technique, and the handling and reporting of uncertainty;
- theoretical and practical aspects of collecting and analysing resource use data in clinical trials;
- theory and evidence underlying the different approaches to choosing discount rates to value health benefits and costs over time;
- transferability of economic evaluation results from one setting to another and the implications for designing clinical trials and other studies.

This collection of papers written by the leading health economists in the field provides an invaluable guide to the current state of economic theory as applied to health technology assessment, and to the practical implications of this theory for the conduct, reporting and understanding of economic evaluations in health care.

Adrian Towse
Director of the Office of Health Economics
August 2001

Contents

Chapter 1

Theoretical concepts in the economic evaluation of health care

Alistair McGuire

1.1 Introduction

The traditional theoretical base of economic evaluation rests on welfare analysis. Individuals are said to maximize utility and societal welfare is defined as an aggregation of utility across all individuals. Governments, in taking societal decisions, do so in a completely benevolent manner; they have no additional objective other than to maximize welfare. Under idealized conditions certain institutions, including perfectly competitive markets, will allocate resources in a manner consistent with this aim. Where markets do not exist the technique of cost-benefit analysis can be used as means of mimicking market allocations to ensure welfare maximization.

While there is a long history in economic analysis of attempting to value health care benefits through the use of monetary measures, economic evaluation based on cost-effectiveness remains the most popular approach. It is not difficult to understand why this is the case: the growth in evidence-based medicine has provided a number of useful measures of effectiveness and the extension to cost-effectiveness is relatively straightforward. Moreover, cost-effectiveness has become a pragmatic tool given increasing technologies competing for limited budgets. At the same time individuals and decision-makers generally find it difficult to apply monetary measures to the benefits of health care for a variety of reasons; it is intrinsically difficult although not impossible.

Attempts to value health states face a multitude of difficulties. Individuals, even if they have adequate information find it problematic attaching value to something as fundamental as health. Even if this is overcome, there is a potential preference revelation problem when individuals do not finance their health care directly. Indirect methods, such as contingent valuation methods, rely on an appropriate specification of the characteristics to be valued. Given the range of potential characteristics which may play a role, aggregation techniques must be used to reduce the dimension of the issue to a manageable scale. Nonetheless economic evaluations of health care interventions continue to play a role in resource allocation decisions. The aim of this book is to provide some justification for that role and an assessment of the techniques and processes that may be used when undertaking an economic evaluation of health care.

This is a fairly limited objective, but even so is a difficult one to meet. The field is progressing rapidly, intellectually as well as on a pragmatic level. As an emerging tool, disagreements over its precise specification and utilization have emerged.

In this opening chapter the background to the general conceptual basis of cost-benefit and cost-effectiveness analysis is outlined. Within this chapter we consider the theoretical foundation arguing that at least two defences may be mounted for the use of such analysis. There is then a brief discussion of what is meant by cost-effectiveness analysis in the health care sector.

1.2 The conceptual basis of economic evaluation

There are many ways in which resources can be re-allocated including robbery, theft and fraud. Generally few would condone these particular methods. A number of other methods, fortunately, also exist and a related normative question arises as to which allocation or resources is best (optimal). The answer to this question obviously depends on the objective being pursued. Amongst these, economists have shown that, if there is a concern with efficiency in re-allocation, a competitive general equilibrium can be described which satisfies this concern. This equilibrium, moreover, is attained under a relatively weak set of conditions. The formal presentation of these conditions was outlined by Arrow–Debreu (1954). This Arrow-Debreu economy presents the basis of the welfare conditions that justify the claims made over efficient outcomes and their tie to competitive markets.

The Arrow–Debreu economy is concerned with exchange primarily between consumers and producers. Consumers have initial endowments of commodities while producers own inputs which produce commodities under known technologies. Each consumer attempts to maximize utility and has a utility function that represents their preferences for commodities. For these preferences to aid establishment of the equilibrium set of prices they must be reflexive, transitive, complete and continuous. They are assumed to be convex. Such restrictions impose useful functional form. Even adopting such well behaved preference functions, a major unresolved problem, to which we return below, is that of inter-personal comparison. This problem essentially recognizes that strength of preference, as well as equity concerns, may impact on outcome.

Producers maximize profits all of which are distributed to shareholders. The production set, which determines the supply of commodities, is assumed to be convex. Trade takes place simultaneously and is only undertaken at the established equilibrium prices. There are a finite number of commodities. Prices are taken as given by consumers and producers. That is no individual can affect the observed prices; they are said to be parametric. This assumption ensures no distortion through monopoly power.

Equilibrium is established when market demands equal market supply. Alternatively equilibrium can be defined as the position attained when excess demands are less than or equal to zero. This is related to as Walras' law, which states that the aggregate value of excess demands must be less than or equal to zero. The less than zero component encompasses situations where the economy may have endowments of commodities not demanded by any consumer, in which case the price will be zero. This condition of equilibrium is established by any set of prices that are consistent with excess demands being non-positive. Indeed, as may be suspected, there are many equilibria which may emerge depending on the initial endowments and the clearing prices.

With this remarkably small number of weak assumptions the Arrow–Debreu equilibrium is established. The beauty of which is that a complete economic system is specified which attains market clearing, that is equilibrium of demand and supply. Moreover, it is equilibrium attained through the individual actions of consumers and producers and, therefore, consistent with a long heritage of economic thought going back at least to Adam Smith. The equilibrium can be embellished to include uncertainty and can also be brought about through coalitions between the individual households; such outturns are referred to as being in the core of the economy.

1.2.1 Pareto optimality and welfare economics

Consistent with this equilibrium is the notion of Pareto optimality. Pareto optimality encompasses at least two concepts: Pareto improvement and Pareto efficiency. A Pareto improvement occurs if a re-allocation of resources increases the utilities of all individuals in an economy. A weaker version states merely that some individuals must gain from the re-allocation. If there are some gainers and some losers then it is not possible to rank the re-allocation outcomes with reference to the Pareto improvement criteria; the states are said to be Pareto non-comparable. This is shown in Figure 1.1 where the hypothetical utility levels of two individuals are plotted against each other. Consider the starting point as position e. A movement from e to y is a Pareto improvement; a move from e to w is a Pareto deterioration. Whereas it is impossible to state whether movements from e to either x or z are improvements or deteriorations. All points in quadrant B are said to be Pareto superior to e, while all those in quadrant C are Pareto inferior. All the states in quadrants A and D are non-comparable as in both cases one individual gains while the other loses utility. As utility cannot be compared directly across individuals, all states A and D are said to be Pareto non-comparable. This is merely a re-instatement of the problem of interpersonal comparability which will be returned to below.

Pareto efficiency is a more powerful concept. It defines the central notion of efficiency as follows: a re-allocation of resources is said to be efficient if at least one individual in the economy is made better off and no other individual is made worse off. This definition turns out to be generally rather useful. It is weak, however, in the sense that many changes in economic circumstances in real life involve the suffering of some and, if this is the case, we require some means of inter-personal comparison to adjudicate over the optimality of the re-allocation. Again discussion of this issue is delayed until later. The Pareto efficiency conditions comprise three main elements: attainment of optimality between inputs; optimality between outputs; and optimality between inputs and outputs.

Optimality between inputs leads to an analysis of producers. Assume that more output, or conversely less use of any input, is deemed a good thing. If the organization

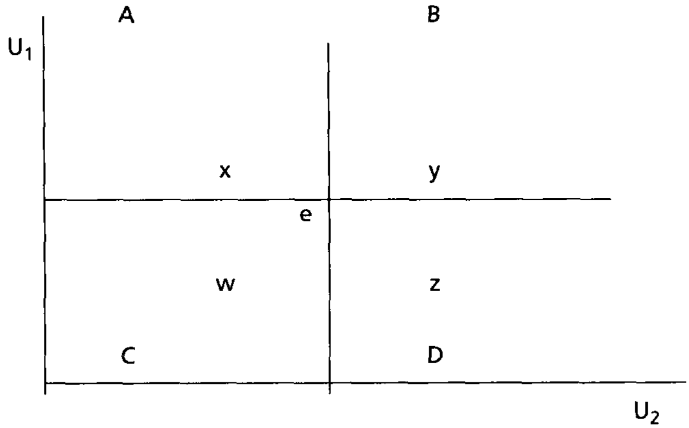

Fig. 1.1 Pareto comparable and Pareto incomparable states

of production is altered such that no further re-allocation would increase the output of one commodity without decreasing the output of another, holding inputs into the production process constant or, given the stock of commodities, decrease the use of one input without increasing the use of another input, such a situation is deemed Pareto optimal. The Appendix to this chapter outlines the technical conditions for this situation to be attained, showing it to be compatible with the marginal rates of transformation (MRT) between production inputs being equalized across all firms. Essentially this relies on re-allocating factor inputs across firms until at least one firm finds it is producing less than its maximum possible output.

Optimality between outputs can be thought of as the efficient distribution of commodities across consumers. It has already been noted a number of times that interpersonal comparisons of utility are, generally, unmanageable. It might appear, therefore, that we can say little about the distribution of commodities across individuals. In fact, Pareto optimality allows some statement of the distribution of commodities to be made. The Pareto criterion sanctions any change that makes one individual better off without making another worse off. So, if from any given distribution of resources such a re-allocation is possible, the given distribution must be sub-optimal. Again the Appendix gives the formal conditions that show that Pareto optimality is attained when the marginal rates of substitution (MRS) for commodities are equalized across consumers.

Optimality between inputs and outputs establishes efficiency between the consumption and production uses of resources. In this case concern is with the use of resources across consumers and producers. So if re-allocation is possible across the consumption and production sectors, which increases the utility of one individual without leaving another worse off this movement of resources, ought to take place.

The Arrow–Debreu economy and the definition of Pareto optimality gives rise to the First and Second Theorems of Welfare Economics. The first states that if a perfectly competitive equilibrium is attained, under the Arrow–Debreu assumptions, then this is also a Pareto equilibrium. The second is the corollary of the first and states that if a

Pareto optimum exists, under the assumptions consistent with Arrow–Debreu, then this is also a competitive equilibrium. These theorems lie at the basis of much of welfare economics. They are remarkable largely because they give rise to a definition of efficiency, Pareto efficiency, and of equilibrium states under relatively few assumptions and the seemingly innocuous ruling that the decision-maker is merely moving resources until no one individual can be made better off without someone being made worse off.

At least two damaging criticisms can be made of Pareto optimality, however. First the Pareto criteria are compatible with a large number of potential allocations some of which represent highly inequitable resource allocations. Pareto optimality is optimal only in one dimension; it is indifferent to the distribution of utilities across individuals. Second, note that we have already stated that Pareto optimality cannot rank all states; the Pareto improvement criteria are inconclusive across resources allocations where some gain and others lose. To adjudicate over these states would entail inter-personal comparison.

Hicks (1939) proposed an ingenious but ultimately flawed attempt to rank states which were Pareto non-comparable. Assume $u_i(y_a, z_a)$ is the utility of individual i under the circumstances described as 'a', with y_a being the individual's income and z_a being a vector of characteristics associated with the state 'a' which individual i values. For example state 'a' could be a state of ill-health and some of the z's could be associated with the individual's ability to work, rest and play in that state of ill-health. Let us introduce another state of the world 'a_1' where the individual's ill health is improved. Then we can define CV as the willingness to pay for an improvement in health:

$$u_i(y_a - CV, z_{a1}) = u_i(y_a, z_a)$$

This amount is referred to as the compensating variation (CV) and represents the maximum, permanent amount of income that may be taken from the individual leaving them as well off in utility terms as they were prior to the improvement in health. In the current case CV is a positive value which is subtracted from income. If the individual was made worse off by the change in the state of the world, CV would be negative and therefore would be added to income (two negative signs are positive remember) as the amount of money the individual would be willing to accept in compensation for the deterioration in health. Adding the compensating variation across individuals as we move from state of the world 'a' to 'a_1', we can see that $\sum CV > 0$ if the gainers place a higher monetary value on the move than the losers.

This is the so-called Hicks-Kaldor criterion that was proposed to overcome the Pareto non-comparability limitation. Looking back to Figure 1.1 it provides a means to state whether moves from e to x and e to z are 'good'. It appears to circumvent the need for inter-personal comparisons as individuals themselves place monetary value on the movement across different states, that is across the different re-allocations of resources. Aggregating these money values across individuals then easily attains the net gain or loss in consumer welfare. In fact four measures were proposed, of which we have outlined two, and the technique can be applied to situations where the relative prices of commodities change or, much more common in the public sector, where the relative quantities of the commodities supplied change (see Willig 1976 and

Randall and Stoll 1980 for further detail). These measures, all based on individual income gains or losses were initially assumed to give a very generalizable solution to the problem of comparing utilities across individuals. The criterion was often referred to as the potential Pareto criterion on the basis that the compensation measures were hypothetical; there was no true re-allocation of incomes associated with the resource re-allocation. It was a mind game.

This raises, at least, two flaws. First for the hypothetical compensation test to operate while avoiding direct inter-personal comparisons implies that all individuals have the same marginal utility of income; all individuals must place exactly the same value on a $1 income regardless of their initial level of income. Second, it is possible, although it must be said unlikely, for the Kaldor-Hicks criterion to suggest a given re-allocation increases social welfare but also, after the change, that a movement back to the original allocation also increases welfare (see Johannson 1991 for details). This leads to a third criterion, the so-called Scitovsky reversal criterion which states that a re-allocation is a potential Pareto improvement if, and only if, after the re-allocation a movement back to the original allocation is not supported by the Kaldor-Hicks criterion. The reversal criterion does not overcome the first flaw.

A number of studies have adopted this general approach to value the health benefits gained from health care interventions. Early studies relied on direct questionnaire approaches. These were fraught with difficulties, including the framing of the questions, the reliance on appropriate specification of the health benefit to be gained and the true revelation of preferences defined over health states by individuals. Johansson (1991) gives a full description of the difficulties faced. Latterly indirect methods have been relied on to value health benefits. For the most part these rely on methods that attempt to proxy such values through statistical models. Contingent valuation models, where the health benefits are valued contingent on a given experimental design, and the related conjoint analysis, where differing aspects which characterize the health state to be valued are considered jointly are the most common of such methodologies.

A number of methodological and practical issues accompany conjoint analysis. First there is the issue of what ought to be measured. In a number of studies the end-point addressed lacks justification. At least three alternatives are possible; valuation of the actual health outcome, valuation of an uncertain health outcome or valuation of a new health care technology where both the treatment benefit and the health outcome are uncertain. The valuation itself may be based on willingness to pay (WTP) or willingness to accept (WTA) premises. There is no theoretical reason to accept one measure over the other although income effects will mean that the estimates will differ. Shogren *et al.* (1994) show that there is convergence of WTP and WTA for market based goods but for non-market based goods (such as changes in health outcome) the estimated values based on WTP and WTA do differ in a persistent manner. As Deiner *et al.* (1998) note contingent valuation studies are, in any case, seldom clear as to whether their concern is with WTP or WTA and, in a related manner, whether compensating or equivalent variation is being measured. It is likely that the WTP measure is more practical as it is bounded by the income constraint. The valuation instrument also varies across studies with choice being largely defined by the alternatives of open ended and closed questionnaires versus bidding rounds. Johansson

(1995) outlines a number of practical concerns surrounding all forms of contingent valuation, including the framing reference for the questions, the incentive to misrepresent preferences and the potential to misrepresent the scenario being tested. Such issues raise concerns over the way in which such studies are designed. At almost every stage there is a potential for bias to be imparted.

It is important to recognize that, regardless of the method employed, the theoretical basis of cost-benefit analysis relies directly on the notion of the potential Pareto improvement and the Kaldor-Hicks criterion. That is, the case for cost-benefit analysis relies completely on a buy-in to the Pareto optimality concept. Individual utility is what is being maximized and to overcome the problem of inter-personal comparison monetary weights are placed on competing resource allocations. Consumers reveal their true preferences through their willingness to pay. There is no concern with distributional notions; compensation need not and, in general, will not be paid.

There have been two reactions to these limitations of the welfare economics approach. First some have advocated the formalization of interpersonal comparability within the welfare criterion (see Ng 1985). This has not occurred to any significant degree. More importantly there was an increasing reliance on the use of explicit welfare functions to aggregate individual preferences and to explicitly address the issues of resource distribution across individuals. The Bergson-Samuelson welfare function was one of the first to address this issue (Bergson 1938; Samuelson 1947). While welfare functions can, and generally should, encompass more than information confined purely to utilities, this form of welfare function tends to be specified such that it is no better than neutral towards all non-utility characteristics of the social preferences it is ordering and is explicitly derived solely from the utility levels of individuals. A particular re-allocation of resources is said to be beneficial if societal welfare, defined across the utility levels of individuals, increases. Explicit weighting of particular individuals' utilities may overcome the potential compensation requirement; essentially decision-makers reveal their trade-offs between efficiency and equity under the social welfare approach. In other words, comparisons across individuals with regard to resource allocation have to be made.

This approach has admittedly been difficult to implement. Arrow (1951) showed some time ago that under a weak set of assumptions it is impossible to aggregate individual preferences in a coherent fashion. Arrow's assumptions included Pareto efficiency; non-dictatorship; independence of preferences such that any social choice concerning a set of social outcomes only depended on the preferences relating to those outcomes; and complete and transitive preference ordering. Each of these assumptions has been closely examined to try to break the Arrow Impossibility Theorem. The conclusion being, from a long literature, that only if interpersonal comparisons are accepted, and convexity is imposed, does the Bergson-Samuelson social welfare function remain a valid criterion for the acceptance or rejection of re-allocation patterns. It is also normally assumed, although this is not necessary, that this form of welfare function satisfies the Pareto criterion such that if a utility ordering of states of the world places state 'a' above state 'a_1' then the social welfare function will also place the social welfare, which is purely defined across utilities, associated with state of the world 'a' above state 'a_1'.

Thus the highest criterion for welfare assessment is a welfare function that is defined purely across utilities and adopts, at least implicitly, some form of interpersonal comparability across individual utility. In other words, the ultimate decision-maker has to make some distributional judgement over which utilities should count higher than others. While other forms of social welfare function can be constructed again defined purely over individual utilities, all are based on a notion of interpersonal comparison (see Boadway and Bruce 1984).

1.2.2 An extension beyond welfarism

A number of investigators, most notably, Sen (1999) have argued that interpersonal comparisons are central to the breaking of the Arrow Impossibility Theorem. In his writings Sen has shown the conditions under which a social welfare function can remain a useful criterion for social decision-making. The conclusions are rather interesting. As he states 'It can be shown that admitting cardinality of utilities *without* interpersonal comparisons does not change Arrow's impossibility theorem at all ... In contrast even ordinal interpersonal comparison is adequate to break the exact impossibility... But it turns out that even weaker forms of comparability would still permit making consistent social welfare judgements, satisfying all of Arrow's requirements, in addition to being sensitive to distributional concerns' (Sen 1999, p. 357).

The weaker forms of comparability referred to enter the decision-making process when the social welfare function is defined across a wider range of arguments than mere utilities. Sen, amongst others (see Culyer 1990 and Wagstaff 1991 for similar arguments as applied to health care), argues that other types of interpersonal argument are as important as the consideration of interpersonal comparison of utilities. Indeed they may be more important than utility measures. If these other forms of interpersonal comparison can be separated from utility measures then they may even be considered in isolation from the interpersonal comparison restricted to utility measures alone.

Building on these insights it has been argued by some that health in particular is an important independent argument in the welfare function (Culyer 1990). An obvious measure for interpersonal comparison in the health care sector would be based on a quantifiable, commensurate measure of health benefit. In other words, some notion of effectiveness that is quantifiable and amenable to comparison across individuals would be a suitable starting point. An immediate candidate is the QALY measure largely as it may be used as a commensurate instrument that may be applied to any health care intervention. This raises an important question what is the QALY measuring? Is a QALY to be regarded as a measure of health benefit alone (i.e. health outcome) or is it, as some would argue, a measure of preference or utility associated with a given health outcome? Historically, the QALY has been used as a measure of preference, but the implication of doing so is that it adheres to the underlying axioms of expected utility theory and there is a literature defending this use (Pilskin *et al.* 1980; Garber and Phelps 1997; Garber 2000).

First let us be clear that expected utility theory is an approximation to behaviour under uncertainty, has a number of rival theories which are held to be better descriptions of actual behaviour (see Machina 1987 or Schoemaker 1982), and stipulates that preferences be specified in such a way that they have desirable mathematical properties and meaningful economic interpretations. Within expected utility theory individuals are assumed to choose amongst different options (health states for example) in a mathematically convenient and consistent manner – in short they are logical. The axioms of the theory define the logic. The important axioms include the following: Completeness; individuals must be able to judge utility across different health states. Transitivity; individuals must choose in a consistent manner such that if outcome H_1 is preferred to outcome H_2, and H_2 is preferred to H_3, then H_1 must be preferred to H_3. Continuity; which implies that each preference set of the same utility level is bounded (i.e. indifference curves cannot cross). Strong independence or the 'sure-thing principle'; states that the utility an individual attaches to an action (e.g. treatment choice) which results in a given outcome (e.g. health state) is independent of the utility attached to any other act if the outcomes are mutually exclusive. If QALYs are measured across different health attributes or characteristics, as most are, then, as Pilskin *et al.* (1980) show, a further axiom is needed concerning the independence of utility across the attributes (see Drummond *et al.*, 1997). The last two axioms give rise to a particular structure of preferences, namely they are said to be additively separable implying that the utility attached to the quality-of-life attained by an individual in any given time period is independent of the quality-of-life realized by that individual in any other time period.

Two further assumptions are necessary before QALYs can form the basis of well-behaved preference structures. First, individuals must adhere to constant proportional time trade-offs. That is, an individual must be willing to sacrifice a constant proportion of their remaining years of life to achieve a given improvement in their health status. Second, individuals must also exhibit risk constancy such that their risk preference is independent of their health state and their utility function is characterized by constant proportionality with respect to risk. If all such assumptions were adhered to then the QALY would be a true representation of individual preferences over health states, holding everything else constant. They would be a cardinal measure of utility in a rather limiting sense; they would not allow direct comparison across individuals but any individual utility function could be linearly transformed. Such utility functions demonstrate that cardinality, in itself, does not imply interpersonal comparison. Bleichrodt (1997), drawing on the work of Sen (1977) shows that only under certain further conditions, essentially that the QALY reflects a cardinal ratio scale, would it become a measure useful for interpersonal comparison. Although the conditions under which a cardinal ratio scale exist are extreme this would make the QALY capable of comparison of levels and gains and losses arising from health care resource allocations.

As noted, there is extensive empirical evidence that suggests that expected utility theory is not an adequate description of how individuals actually behave under conditions of uncertainty. While this is a harsh criticizm in itself it is even more devastating when applied to QALYs. The quality-of-life weights in the QALY are normally gained through instrument measures based on simulated or experienced circumstances. If individual respondents to these instruments do not base their responses on the logical structure (i.e. the assumptions) demanded by expected utility theory then the use of the resulting QALY measures as representations of preferences, and therefore their use as measures of utility within a Paretian framework is violated.

Indeed for many the axioms surrounding expected utility theory and Pareto optimality are too strong, and violated too often in reality, for them to accept this as a conceptual basis for using economic evaluation based on a cost per QALY calculation.

This leads back to the alternative interpretation of the QALY as a measure of health benefit unrelated to individual preferences over health states. The most thorough assessment of QALYs under this interpretation is given in Broome (1993). There Broome sets out examples of how QALYs fail generally to meet the axioms of expected utility theory and gives an interpretation of QALYs as a measure of health benefit. As he, and Culyer (1990) point out this interpretation is not de-limiting. Defining QALYs in this way leaves open the possibility that QALYs are cardinal measures of health benefit defined in some quantitative manner, or expressions of the value individuals attach to this health benefit or, even again invoking the assumptions of expected utility theory, expressions of the utility attached to this health benefit.

Ironically, accepting the QALY in terms of the lowest degree of preference revelation, which is simply as a cardinal measure of health benefit, gives the greatest power in terms of interpersonal comparison which is unrestricted by the axioms of expected utility theory. It is, therefore, consistent with the social welfare basis of resource allocation, outlined above, advocated initially by Sen based on arguments other than utility entering the social welfare function.

Adoption of this approach entails a move away from the Paretian definition of efficiency to some other notion of efficiency, the most obvious being that of, assuming separability in the welfare function, maximizing health or maximizing QALYs. This has been discussed by Wagstaff (1991) who points out that pure maximization of QALYs gives rise to a social welfare function which resembles a utilitarian welfare function (i.e. $\sum_i$ QALY = summation of QALYs across individuals). While this meets the requirement of inter-personal comparability, as he points out it seems to lead to definitions of efficiency which ironically turn out to be more restrictive than the Pareto criteria. If QALYs are gained more by the young, merely because they have more life to lead than the elderly, an efficiency criteria which states maximize QALYs would redistribute health care resources from the elderly to the young.

Of course, this raises distributional concerns. A different conceptual version of the social welfare function would retain the QALY as a health benefit measure but also incorporate explicit weights to be attached to individual QALYs to reflect distributional concerns. In fact the introduction of explicit weights gives rise in theory to a wide range of possible welfare functions (see Williams and Cookson 2000). In practice it may mean maximizing the distributional weights assigned by policy makers. In this respect note that the role of the QALY becomes akin to the health benefit measures used in the Oregon experiment.

The extensive use of cost-effectiveness, and the calculation of cost-per-QALY is included in this definition, can therefore be justified on at least two grounds. First, traditionally, but with difficulty, through recourse to Paretian welfare economics. Second, through adoption of the QALY as a measure of health benefit, and the pursuit of some extended form of social welfare function. Within the Paretian framework Mishan (1971) has stated that 'To be rather rude about it, the analysis of cost-effectiveness can be described as a truncated form of cost-benefit analysis'. As he details cost-effectiveness is concerned with resource allocation issues *after* the overall budget has been allocated to a particular activity, while cost-benefit analysis provides information necessary to determine the overall budget consistent with efficient resource allocation. In this sense it is implied that cost-effectiveness is piecemeal and operates on efficiency grounds within constraint. Such piecemeal approaches are more consistent with the second, more general justification. Acceptance of a more general justification is also consistent with cost-effectiveness analysis being complementary to cost-benefit analysis. Either way cost-effectiveness analysis remains concerned with the process of allocating resources efficiently; all that is being discussed is what definition of efficiency is appropriate. Yet if the definition is not explicitly detailed then the cost-effectiveness of an intervention cannot be anything other than flawed.

1.3 **Cost-effectiveness analysis**

Assuming that the definition of efficiency has been agreed upon, difficulties remain in implementation. The application of cost-effectiveness analysis follows directly from the attribution of relevant costs to relevant effects. A large component of this book is concerned with the definition of what is deemed relevant. In this section an overall introduction to the practical use of cost-effectiveness is given.

Generally cost-effectiveness is pursued to test the null hypothesis that the mean cost-effectiveness of one health care intervention is different from the mean cost-effectiveness of some competing intervention. It is calculated as a ratio:

$$R = \frac{C_a - C_b}{E_a - E_b} = \frac{\Delta C}{\Delta E}$$

defining the incremental cost per unit of additional outcome. This calculation can be undertaken in a number of ways reflecting both the available data and the preferred methodological approach.

Ideally data should be available from a true random sample of the patient population under study. If this is the case, estimation of the incremental cost-effectiveness ratio (ICER) follows the notion of statistical inference based on sample means. In other words the ICER becomes:

$$R = \frac{(\mu_{\mathrm{CA}} - \mu_{\mathrm{CB}})}{(\mu_{\mathrm{EA}} - \mu_{\mathrm{EB}})}$$

where μ is the expected value of the true population cost of treatment A (μ_{CA}), the expected value of the cost of treatment B (μ_{CB}), etc. O'Brien *et al.* (1994) noted that four states of the world could be observed based on the relative magnitude of the incremental costs and effects:

1. $\mu_{\mathrm{CA}} - \mu_{\mathrm{CB}} < 0$; $\mu_{\mathrm{EA}} - \mu_{\mathrm{EB}} > 0$; which is defined as a dominant outcome for intervention A as it is both less inexpensive to implement and more effective than treatment B.

2. $\mu_{CA} - \mu_{CB} > 0$; $\mu_{EA} - \mu_{EB} < 0$; which is defined as a dominant outcome for intervention B as it is both less inexpensive to implement and more effective than treatment A.
3. $\mu_{CA} - \mu_{CB} > 0$; $\mu_{EA} - \mu_{EB} > 0$; which defines a trade-off in which the higher cost of treatment A must be compared in terms of its higher effectiveness.
4. $\mu_{CA} - \mu_{CB} < 0$; $\mu_{EA} - \mu_{EB} < 0$; which defines a trade-off in which the lower cost of treatment A must be compared in terms of its lower effectiveness.

Note some important points here. First, these comparisons apply to where the decision-maker is already faced with a budget and mutually exclusive projects. Under these circumstances the objective is to maximize the level of health effects or the preferences attached to these health effects, depending on your persuasion, given the budget. ICERs are ranked with the least cost per unit of health effect (least cost per QALY produced) being the most desirable and the decision-maker then works down the list until the budget is exhausted. If there are really only two interventions to be compared then given dominance and a desire to maximize health the fourth situation is automatically redundant. If there are other interventions against which these two interventions are to be compared, in as much as they are competing for the same budget, as Weinstein and Zeckhauser (1973) note linear programming approaches may be adopted to find the optimal ranking of projects. Even in these wider circumstances if the desire is to maximize health effects, unless the true resource use of readily transferred resources from one treatment intervention to another can be established, it is unlikely that the fourth situation warrants much attention.

Dominance and trade-offs may not be clear-cut, however. This can be seen by concentrating on the third situation above ($\mu_{CA} - \mu_{CB} > 0$; $\mu_{EA} - \mu_{EB} > 0$), which is commonly the one of greatest interest. Figure 1.2 is illustrative. Note first that any cost-effectiveness ratio lying on a straight-line drawn from the origin will be the same; that is the value of the cost-effectiveness ratio for treatment A is equal to the cost-effectiveness value of treatment C; although obviously the budgetary implications differ.

Dominance exists when an option is both more costly and less effective than an alternative. Thus, starting at A, any point to the north-east (in the area labelled D) is dominated by A. However, extended dominance exists when an option is less effective and more costly than a linear combination of two other strategies with which it is mutually exclusive. Assuming constant returns to scale and perfect divisibility of projects, any combination of treatments A and C along the heavily shaded portion of the line joining them will be more cost-effective than Option B, as this will be provided at lower cost with more effect. However, an implication of this is that some proportion of the treated population will be receiving the less effective treatment (A). In practice ICERs should compare each intervention to the next most effective option, after eliminating options that are dominated.

Given that the expectations which generate the ICER (the means, μ are drawn from a random sample of the true population of the individual costs and effects they will have other distribution defining moments associated with them, such as variances, levels of skewness and so forth. This allows the possibility of hypothesis testing and most obviously a null can be formulated as $R = 0$, to be tested against an alternative