THE NORMATIVE ECONOMICS OF HEALTH CARE FINANCE AND PROVISION

A. J. CULYER

University of York

I. INTRODUCTION

Whereas in many countries the ‘crisis’ in medical care has been seen in terms of ‘excess’ spending on health services, in the UK it has been seen (at least by most of those who manage and work in the NHS) as a crisis of ‘underfunding’. This has come about as the result of the government’s successful attempts to restrain the rate of growth of real spending on the NHS, which in turn reflects the government’s belief that its principal effective weapon against what it perceives to be inefficiency in the NHS is to challenge management (itself reformed and to some extent liberated) by systematic financial squeezes. The same concern has given rise to a host of proposals for reform of the NHS most of which involve a much greater role for private insurance, private finance of other kinds (such as out-of-pocket payments), and private provision of health care itself.

Few of these contributions have been informed by the work of health economists, particularly their normative work. There have been several reviews of the empirical literature (e.g. Culyer, Donaldson and Gerard, 1988; Culyer, Brazier and O’Donnell, 1988). This essay is a review of the main contributions of a more conceptual kind.

I have adopted the rather general term ‘normative’ in my title, rather than ‘welfare economics’ for a reason. This is that much of the modern systematic approach to policy questions in the health territory is based on an approach that transcends (and may

---

1 I am grateful for the helpful comments of Richard Arnould, Gwyn Bevan, Paul Fenn, Alistair McGuire, Gavin Mooney, John Posnett, and Alan Williams. The usual disclaimer applies.
come to supersede in part) the ‘welfarist’ tradition of normative economics.

A key concept underlying the current discussion is ‘efficiency’. This is, however, far from being an unambiguous notion and, because it is also so central a concept (one indeed that pervades the entire debate) it is the one that preoccupies me in this paper. The practical context in which the concept can illuminate is, evidently, scarcely less important. For the purposes of this paper, the practical context is taken to relate to broad policy questions about the ‘efficiency’ (somewhat defined) of ways of financing health care (e.g. through private, competitive insurance or via compulsory public insurance), and of the user-prices that it should carry (e.g. whether ‘free’, partially subsidized, not at all subsidized, differentially subsidized according to particular categories of user).

These are the principal ‘demand side’ contextual questions whose discussion forms the bulk of the paper. There is, of course, a host of related questions that arises in specific policy formulation and a whole raft of supply-side ‘efficiency’ questions. These may from time to time be alluded to here, or cross-references made to other papers in this volume. They are not, however, the main focus.

The conventional ways in which economists use the term ‘efficiency’ (which at this level of generality are illuminating and ideologically innocuous) are the following:

**technical efficiency**: obtaining when for a given output the amount of inputs used is minimized or (what is the same thing) when for a given combination of inputs, the output is maximized. This is a supply-side concept that is a necessary condition embodied in the subsequent concepts of efficiency. In economic jargon it means ‘being on an isoquant’. Since there is usually more (many more) than one way of producing an output that meets this condition, the implied balance of resource use even for a given target output is not unique.

**cost-effectiveness**: obtaining when for a given output the cost is minimized or (what is the same thing) when for a given cost the output is maximized. This is also a supply-side concept. It embodies technical efficiency (clearly if one were using more inputs than were technically necessary to produce an output, that cannot be cost-effective) and the cost-effectiveness version of efficiency is also embodied in the next concept of efficiency. Being cost-effective necessarily eliminates as ‘inefficient’ some technically efficient combinations of inputs, but it leaves unsettled the question of the efficient rate of production. In economic jargon, cost-effectiveness means ‘being where an isocost line is tangential to an isoquant’.

**ideal output**: obtaining when cost-effective outputs are produced at a rate that is ‘socially’ optimal and allocated to individual members of society in a ‘socially’ optimal fashion. This is ‘top-level’ efficiency and arises from combining supply-side and the demand-side considerations. In a society with limited resources it entails establishing rates and allocations of outputs that are such that no alternative rates of reallocations can be perceived that are ‘better’. In economic jargon it entails setting marginal rates of transformation on the production-side equal to marginal rates of substitution in consumption.

While this trio is an extremely useful sorting device in discussions of efficiency, the main focus in this paper is on the third, though (as will be seen) ambiguities about the concept of ‘output’ will force us also to look to some extent at the other two notions of efficiency as well. Of the various tricky things embodied in ‘ideal output’ as a concept of efficiency, one that has come to the fore in health economics in a distinctive way concerns the meaning of ‘socially’ and ‘better’ (or ‘optimal’).

(i) Welfarism and Extra Welfarism

One approach in health economics, which has become the traditional one in economics as a whole, is what Sen (1977) calls ‘welfarist’. This is very much in accord with liberal political opinion and asserts that social welfare (any increase in which is ‘better’ than none) is a function only of individual welfare (or utility) and judgements about the superiority of one state of the world (defined by reference to these utilities) over another are made irrespective of the non-utility aspects of each state. Moreover, the individual welfares (or utilities) are a function only of goods and services consumed.
The other approach, which might be termed ‘extra-welfarist’, relaxes what its adherents see as an undue information restriction in welfarism so that other aspects of each social state are also embodied in the judgement. Since, in the welfarist approach, the basis of social welfare—or dis-welfare—is only the utility got from goods and services (including labour services), an important class of ‘extra’ welfare sources is the non-goods characteristics of individuals (like whether they are happy, out of pain, free to choose, physically mobile, honest). Extra-welfarism thus transcends traditional welfare: it does not exclude individual welfare from the judgement about the social state, but it does supplement them with other aspects of individuals (including even the quality of the relationships between individuals, groups, and social classes).

This distinction is actually quite old in the history of (the so-called ‘new’) welfare economics. In Bergson’s (1938) classic theoretical article, for example, his social welfare function included unspecified terms that could be interpreted as extra-welfarist elements of the sort just described. But he dropped them in favour of an explicit ‘partial’ analysis after a page.

Another famous extra-welfarist strand in the wider literature is the notion of merit goods (Musgrave, 1959)—goods whose consumption is considered so meritorious (by government) that they are made available on terms that are more generous than in the market place. Despite (unsuccessful) attempts to bring merit goods into the welfarist scheme of things (e.g. Culter, 1971), it seems altogether more preferable to adjust the scheme of things so as to incorporate such considerations fully rather than leaving them as a kind of ad hoc ‘escape clause’ (Margolis, 1982) lying outside traditional theory; not fitting into it but necessary in order to prevent theoretical emasculation (inability to explain why some common phenomena are observed, inability to discuss in a consistent normative framework some matters that are of evident normative importance). This plainly involves the possibility of overruling individual judgements of value and raises the question not only of the weights to be attached to individual utilities in a social welfare function but of who should be assigning those weights. Should the values of some members of society count for more than those of others?

More recently an explicit departure from welfarism was advocated by Williams (1972) and discussed further by Sugden and Williams (1978) in the context of cost-benefit analysis. This is the ‘decision-making’ approach which is contrasted with the ‘Pareitian’ approach. The latter is one example of welfarist analysis in that social welfare is a function only of the utilities to individuals of goods and services. The former allows that governments (and other ‘decision makers’) may have other objectives than the making of (actual or potential) Pareto improvements, which may involve not only imposing their own values on the consumption of individuals (rather than those of the individuals in question) as with merit goods, and their own view of the appropriate intertemporal discount rate, but also takes into account some extra-welfarist elements of choices that pure welfarism excludes. In this context the answer to the question ‘who decides what entities with what weights go into the social welfare function?’ is ‘decision makers’.

The bulk of the health economics literature is decidedly welfarist in orientation—and Pareitian to boot. This is discussed in Section II. Section III addresses the more recent extra-welfarist tendencies of health economists, and also (since the approach lends itself so readily to this) looks at the ways in which extra-welfarism has infiltrated the supply-side notions of efficiency.

Although the discussion is of a normative literature, it should be pointed out that in neoclassical economics it is easy to slip between positive and normative (welfarist-style) according to whether one treats utilities as behaviour-generating indices or ascribes direct normative value to them. Thus, some of the literature on the NHS (e.g. Lindsay, 1969; Culter, 1971b) was ostensibly explanatory: the authors sought to provide empirically falsifiable accounts of some of the institutional features of the NHS by postulating the presence of particular arguments in individual utility functions. Under welfarism, it is quite easy to re-interpret these theories as claims for the Pareto-efficiency of the NHS. In the positive interpretation, the institutional features are phenomena shown (or so it is claimed) to be the predicted results of utility-maximizing individuals’ behaviour in particular environments; in the normative interpretation, the institutional features are seen as desirable attributes.
rather than merely explained phenomena: the theory justifies what is or implies what ought to be, prescribing rather than predicting.

As a final prefatory remark it is worth pointing out that health economics (in common with many other fields of application within economics) has by and large addressed only a part of the agenda of the ‘political economy’ of health care, to do with finance, public or private. There is very little on the normative implications of different forms of ownership or organisation on the provider side. Indeed, the aforementioned positive theories of the NHS (Lindsay, 1969; Culyer, 1971b) are typical in that they are concerned entirely with the terms of access to health care. Although it may have frequently been thought in the past that arguments for ‘free’ or subsidized or publicly financed health care were ipso facto arguments for publicly provided care, that is not so. Public finance can in principle be combined with public or private provision, just as privately financed consumption can be combined with public or private provision. That is why it is possible for people who support the general idea of free health care available at the time it is needed also to support private provision, or a public/private mix in provision, or competitive ‘internal markets’ within the public system. However, the focus here is on the demand side, and so no further development of these possibilities is attempted.

II. THE WELFARIST APPROACH TO HEALTH CARE FINANCE AND PROVISION

It was (and is) common for these issues to be addressed by considering a set of factors that make health care ‘different’ from other goods or services and which therefore may constitute a reason for wanting to allocate it differently from these ‘other’ goods or services (and to evaluate the efficiency of different allocations differently) (Culyer, 1971). Implicit in this approach is the presumption that goods and services are in principle best allocated by market mechanisms and that departures from that mechanism require special warrants. This seems to be the product of a particular bit of cultural conditioning to which most (Western) economists are prone but, since it is my purpose more to outline an approach than to make sociological comments on its culture-contingency, I shall say no more on that here. The following sections embrace the usual ‘list’ of factors that make health care ‘different’ and I shall indicate the kind of reasoning that relates to each.

(i) The Competence of the Consumer

The welfarist approach requires that choices made by or for consumers be rational in the particular sense of that term used by economists. Should choices not be founded on axioms that include, for example, an ability to compare alternatives and to rank them consistently (transitively, if weakly) in order of preference, the entire edifice tumbles because it no longer becomes possible to infer from actual behaviour that the choices made were (subject to resource constraints) those most preferred. Even if consumers were better informed than they typically are about the pros and cons of alternative actions (for example, choices between alternative strategies of personal medical treatment) there are occasions when even mentally healthy people prove incapable of choosing (especially between alternatives that involve horrid consequences). Mentally ill and mentally handicapped people may more frequently be found to be ‘irrational’ in this sense. There are also occasions, of course, in which they are incapable of making any choices at all (for example, when they are the traumatic and unconscious victims of accidents).

It is common for the consumer, who is almost invariably incompetent in some degree, to defer some of the judgements involved in clinical choices to professionals, especially medical doctors, who act as agents on his or her behalf: ideally choosing in the way the individual would, had he or she been possessed of the same informational advantages as the professional, in a system characterized by what Evans (1981) has termed ‘incomplete vertical integration’ between consumers and producers.

The particular technical skill possessed by the professional that is relevant here is a better knowledge of the effect that health care will have on health. If one supposes that it is health which generates utility for individuals, both directly via a sense of well-being and indirectly in the sources of welfare of which better health enables one to take advantage, then health care is itself only instrumental:
a means to an end. Even if the consumer is able rationally to value health, he or she will usually have much less information about the process of medicine and the risks that different processes imply. Contrary to the usual welfarist assumption, the buyer is not the best judge of his or her interests but must rely on the seller's advice. It is quite easy to imagine circumstances in which the selfish interest of the professional can conflict with the best interests of the patient. Health itself is not a traded commodity, yet that is what the rational consumer may be expected to seek. The traded commodities are information and health care, which the consumer cannot normally be expected to evaluate, and neither of which are valued for themselves.

The technical expert is thus required to be more than 'merely' a technical expert. In acting as agent, the expert as it were 'enters the skin' of the patient. This is what seems to endow a professional relationship with its most important—but elusive and delicate—characteristic (Trebilcock et al., 1979). Being delicate, it is vulnerable. It is vulnerable to cultural assumptions not shared by agent and client (the male-dominated specialty of obstetrics is often charged with sexist disregard of the interests of its wholly female clientele). It is vulnerable to misplaced technological zeal, in which the expert subjects the patient to painful diagnostic testing that yields little usable information, or makes presumptions about a patient's trade-offs between short-term risks and long-run benefits. It is vulnerable to class bias, in which the professional, as a member of a higher class than at least some patients, is incapable of 'entering their skins'. It is also vulnerable to financial distortions, whereby some systems of medical remuneration encourage the supply of services whose principal justification is the income they bring the expert rather than the benefit they bring the patient. This is sometimes referred to as supplier-induced demand (Evans, 1974, 1976). These 'vulnerabilities' are forms of market failure and are widely regarded as justifications for regulatory measures to protect both consumers and doctors; the former from quacks and the exploitation of a professional monopoly, the latter also from quacks (unfair competition) and unreasonable demands of dissatisfied (and possibly litigious) customers. They also provide the basis for universally observed systems by which medical education is determined and entry to the profession regulated, and for subsidized in-service training systems and other professional activities. The existence of such market failures means that there are inevitably nice balances to be struck here between, for example, the self-regulating monopoly awarded to doctors in medical practice and its possible abuse by inhibiting innovation (e.g. alleged but not actual quackery, inefficiency in doing whatever is done, and the earning of monopoly rents). While economics is helpful in identifying the risks, and in predicting the consequences of policies designed to mitigate them, its role is mainly qualitative: much judgement is required in making assessments of the pros and cons, and their respective strengths, of alternative possible arrangements.

(ii) Supplier-Induced Demand (SID)

Evans' idea that physicians have a target income and adjust workload (under a fee-for-service system of paying doctors) in response to changes in the environment, seems to have grown out of the empirical observation that regional utilization of health care is positively associated with the annual stock of doctors, holding price and other variables constant (Fuchs, 1978; Cromwell and Mitchell, 1986; Phelps, 1986). The thesis is that physicians will induce patients to use more services in order to maintain income. A positive association has also sometimes been found between physician stock and prices, though this result is even more disputed than the fundamental utilization effect (Sloan and Feldman, 1978; Auster and Oaxaca, 1981; Green, 1978). There are, of course, huge econometric and empirical problems in testing for SID: only one study for example (Pauly, 1980) controlled (approximately) for patient health status, but Rice's claim (Rice, 1983, 1987) that experimental rather than routine data strongly support an inverse relationship between reimbursement rates and use of services seems persuasive.

SID is an area in health economics where we suffer an embarass de richesses. The target-income hypothesis is one possible explanation. But there are others too that have never been rigorously compared and tested: increasing the numbers of doctors increases their availability (less distance to travel, less time to wait) and hence reduces the 'time
price' to patients; there may be 'excess\(^2\) demands at the existing stock and a non-market-clearing price, so that increasing utilization following increases in the stock is simply the meeting of previously unmet 'excess' demand. Doctors may be less vulnerable to the effects of consumer detection of SID in communities with a high doctor/population ratio since, given extensive consumer ignorance, new patients can always be found to replace any who leave because of the SID they detect—if so, cross-sectional analysis will produce the observed correlation between physician stock and utilization. It is unfortunate that this unsatisfactory state of the art persists—and is compounded by empirical uncertainty (at least in some minds) about the very existence of SID at all—given the huge potential threat that SID constitutes to the welfarist approach and especially to the pro-market camp within the welfarist school. If demand curves really do reflect the health-irrelevant preference of suppliers then, depending on the depth of this contamination, the use of willingness-to-pay as an indicator of consumer welfare is more or less illegitimate. Hence the considerable amount of passion found in the literature on this subject.

(iii) Where Does this Get Us?

The chief lesson for health care finance to be drawn from this catalogue is not that the 'market' is inherently more flawed than the 'state' in the way in which medical care is financed (or vice versa) but that there are no simple lessons to be drawn. Market systems and their (usual) fee-for-service methods of paying doctors are prone to violate the underpinnings of welfarism in one way. State systems and their (usual) salary or capitation systems are prone to violate the underpinnings in another. If fee-for-service may encourage an excess of interventionist zeal then salaries and capitation fees may do the opposite. Under either system it is common to see a lot of interregional variation (McPherson et al., 1981; Vayda et al., 1982).

The agency relationship must inherently be incomplete. The professional may know best about the instrumentality of health care and, if working in a system that encourages a concern for the whole patient (as is usually the case in general practitioner services in those medical care systems like the UK's that still have GPs), may also come to know a good deal of relevant information about the patient's values, financial circumstances, working life, and family context. In such circumstances the agency relationship is likely to be as perfect as it probably can be. But even in this situation, the agent too has his or her own values, financial interests, working relationships, family and social characteristics which need not be congruent to those of the patient (Evans, 1984). In both market and non-market systems of health care finance and provision, the medical monopoly is strong which seems to produce higher monopoly rents in the more market-orientated systems (but see Lindsay, 1973).

As regards the terms of access for consumers, much of this discussion casts doubt on the usefulness of marginal willingness to pay as an adequate representation of the marginal benefit to consumers, but it is not clear what the notion of welfarism requires to be put in its place, nor is it clear whether 'free' care under the NHS or under first-pound zerodeductible health insurance encourages inefficient rates of use relative to the 'true' optimum. On this issue, as will be seen, the extra-welfarists are able to be more forthright.

(iv) Caring and Sharing

In the early days of the welfarist literature, the idea that health care was 'different' because the medical care consumption of others, or the health of others, might have a direct influence on the welfare of oneself, was given short shrift. There was early recognition (e.g. Weisbrod, 1961) that a direct physical externality might exist in cases of communicable disease (via infection or contagion): an individual in choosing or rejecting vaccination may fail to take account of the benefits accruing externally (viz. to others) in the form of a reduced probability of the others contracting a disease. This was concedé to provide a case (abstracting from the problems arising from consumer incompetence) for subsidized prices for such services and, in some cases, for making them compulsory (e.g. for immigrants). But the scope

---

\(^2\) I put the term in quotation marks as I do not wish to imply that the demands are necessarily in excess of a Pareto optimal rate of use, merely that they are excess in the (positive economics) sense of demand exceeding price at the going full user-price.
for interference with the market on this ground alone was evidently highly limited and the marginal externality probably falls quite fast as the proportion of immunized individuals rises.

The subtler kind of externality, direct interdependence of the form that makes one person's welfare directly dependent on the consumption of others, was not taken very seriously (e.g. Lees, 1960, 1962; Jewkes et al., 1963; Klarman, 1963, 1965; Buchanan, 1965). The view is well-encapsulated by Lees (1967):

It is argued that there is another relevant externality, namely, the disutility felt by an individual at the thought that others are not getting adequate medical care. This is no doubt so and is the basis of philanthropic support for health services. But the important point here is that there is nothing special about health services in this regard. Similar disutility is felt at the thought of others not getting adequate food, clothing, housing, and other goods commonly regarded as necessities. Apart from philanthropy, the community approach to this problem has typically been public subsidies to those in need. Externalities of this kind do not establish even a prime facie [sic] case for the abolition of markets and the substitution of collective arrangements.

Even where externalities exist, recent studies have shown that the scope for government spending to correct market imperfections and raise the general level of economic efficiency is less than conventional analysis had led us to believe. The implications of these studies may be summarized thus:

(a) many externalities are irrelevant to human action and the achievement of optimal solutions;

(b) many relevant externalities can be, and are, dealt with voluntarily;

(c) the costs of governmental intervention, even when it is 'perfect', may outweigh the benefits;

(d) the imperfections of government as a decision-making and choice-making process may make an 'imperfect' market situation more imperfect.

Such was the dominant idea at that time: the 'nothing special' about health argument; the distinction between philanthropy approved at one level of collectivity (private charity) but not at a higher (governmental) level; the reference to studies showing the scope for corrective government action to be limited; the mention of things which 'may be' (but which are not) measured to find out what actually is.

One of the ironies of the period was that these remarks and others in the same genre were being made in societies that had substantially altered the terms of financial access to health care, possibly for reasons (in part) of externality. At least, such a view would imply the possibility that the subsidy arrangements were a rational response to a real situation rather than the irrational constructions of ideologues versed in the eternal truths of neoclassical economics. It might be thought that the reasons for this intervention were like those for similar interventions in education and housing. Since even crude policies might have gone sufficiently far to eliminate most Pareto-relevant marginal externalities, casual observation could not have revealed massive failures to internalize at efficient rates of utilization: one needed both observation of what was and what otherwise would probably have been to make an assessment of the significance of externalities; at the very least one needed the imagination to envisage what a counterfactual 'uninternalized' world might look like.

What sort of evidence might settle this issue of the existence and size of what some have called 'caring' externalities (e.g. Culyer, 1980)? One is introspective. Are sufficient numbers of introspectors prepared to sacrifice some of their own consumption so that others may have more? If so, they care in the externality sense. In particular, invite introspectors to try to imagine a society in which health care was supplied entirely by commercial organizations without public subsidy or by private charities. Ask each, given that they had imagined themselves into such a society, whether they would be willing to sacrifice some of what each has so that others would receive more. If each would (and they may want to add the side condition that each would sacrifice only if others similarly placed also made sacrifices) then they would be providing evidence for the existence of caring externalities of this sort.

One of the issues that arises in this context is the question of whether subsidies should be specific (e.g. in-kind or directly tied to purchases of specific goods and services as with voucher schemes) or general (essentially income support). The basic line of argument has tended to be that it seems that
generalized support may be appropriate for goods and services that are highly income elastic at low levels of household income, where there will be a strong presumption that income supplements will be spent by most families on basic necessities like food (so there is no need for a ‘National Food Service’), but that where this is not the case, or where people are held to be poor judges of their own welfare, specific subsidies are a more cost-effective means of promoting consumption of particular goods since they utilize substitution effects as well as income effects. It should be clear that the potential superiority of specific subsidies arises from the welfare gain to the subsidy-provider: in any comparison of alternatives, provided that the subsidy-receiver is not worse off under one than the other, that yielding the largest benefit to the subsidy-provider is Pareto-preferred.

There has been some speculation about the way in which inequality in health care consumption enters welfarist utility functions. This seems mostly to have arisen in the context of positive rather than normative economics: to provide explanations of the patterns of subsidy actually observed in various societies. From the late 1960s into the early 1970s, focus continued to be upon the consumption of health services as the principal source of external ‘concern’ (rather than, say, the effect of such consumption on health), though for some the focus was on absolute rate of consumption while for others it was relative.

Pauly (1971) argued that the (negatively sloped) external demand (marginal valuation) for a person’s care was invariable with respect to the identity of the person of concern and that the demand for health care was, in general, income elastic. Detailed microempirical studies (all North American) on this latter question yield little unambiguous evidence about direct income elasticities (though they seem to be positive) because of income-related upper limits on out-of-pocket expenses under health care insurance (see, e.g., Manning et al., 1987), though aggregate studies show income elasticities of between 1.18 and 1.36, so that a 10 per cent increase in (aggregate) income can be expected to lead on average to a roughly 12.5 per cent increase in health care spending (for a review see Culyer et al., 1988 and, for a critique, Parkin et al., 1987).

There seems little doubt about the overall negative slope of demand curves. The econometric evidence is not easily interpreted, partly because in those systems which have been most studied—North American—fee-for-service and the possibility of supplier-induced demand can cause supply-side contamination of the ‘pure’ substitution effects of changing user-prices. Beck (1974) estimated that the reduction in use following the introduction of charges of $1.50 per surgery visit and $2.00 per house call in Saskatchewan in 1968 was about 7 per cent for the whole population but 18 per cent for the poor. Early studies of price-elasticity found values in the range -0.4 to -1.0. Subsequent work by Phelps and Newhouse (1974), Newhouse and Phelps (1976), and, most recently and most authoritatively, the Rand study reported in Manning et al. (1987), found price-elasticities for all health care in the -0.2 to -0.1 range (i.e. a 10 per cent increase in user price causes demand to fall by one to two per cent). The effects in the Rand study were, however, much stronger for the poor, for children’s demand, and especially for the children of the poor (see Lohr et al., 1986). These results have tremendous significance: the handicapping effects of, say, untreated otitis media in children far outweigh the short-term functional impairment, as further socializing and educational handicaps will almost inevitably become added for children already disadvantaged from birth.

The implications of the Pauly model are quite explicit: given a similar cost of care per person in a particular diagnostic group, efficient internalizing of the external effect requires a variable subsidy: varying from 100 per cent in the case of the lowest income group to 0 per cent at a sufficiently high level. This is archetypical ‘selectivity’ in social policy.

The argument can be seen quite simply in terms of Figure 1. In this figure the three marginal willingness-to-pay curves (marginal valuation curves or real income constant demand curves) of individuals A, B, and C are shown. Each is supposed to have a common ‘taste’ for health care and the difference between is postulated to arise from income differences (ability-to-pay). The curve EMV is the external marginal valuation curve reflecting the value that members of society other than these three place upon their consumption. It is assumed to be the same for all (anonymous) individuals. At a market price $\text{P}_\text{M}$, A consumes $x_A$, B consumes $x_B$, and C consumes no $X$ at all (when a particular event strikes, such as a specific illness). The socially
optimal amounts (assuming price represents marginal social cost accurately) are, however, \( x_a^* \) and \( x_b^* \), determined by the condition \( MSC = MV_a + EMV \) for all \( i \). This implies a variable subsidy such that \( C \) faces a marginal price of \( P_C \), \( B \) faces \( P_B \), and \( A \) pays the full price: the subsidy varies inversely with ability-to-pay. In practice, of course, fine distinctions between individuals will not be possible and general approximations will have to be made—but on this rationale they should clearly have the selective character shown.

Culyer (1971; for a simple version see Cullis and West, 1979) attempted a model (again assuming conventional signs on the income and price elasticities) that implied zero prices, as in the NHS, and showed, using a triangular Edgeworth Box, that in a two-person world free care will be preferred by both the subsidized and the subsidizing parties over cash transfers, provided that the preference functions of the two individuals produced an appropriate configuration of offer curves.

In contrast to these ‘absolutist’ externality models of ‘caring’, Lindsay (1969; see also Cullis and West, 1979) offered a ‘relative’ model of ‘sharing’ in which equal treatment for equal (medical) need was implied. He showed that the most cost-effective method of achieving a given level of equality is by a combination of standard subsidy (but not in general to reduce price to zero) and enforced denial of access via nonprice rationing (‘abstention’ in his terminology) to the better off.

One set of problems arising from this set of externality models is that, while each provides an account of some of the features of health care subsidy (e.g. some selectivity, some ‘free’ care, some degree of reduced inequality in access), their practical implementation (assuming that any one of them underlies the actual practice of democratic societies seeking to implement institutional arrangements that promote Pareto optimality more closely than would the market unaided) means that the apparently sharp distinctions between their implications at the theoretical level become blurred in practice. It is consequently hard to discriminate between them on empirical grounds.

There are also more deep-seated difficulties. One arises from the characteristic that the externality is in each case public so that its internalization is of benefit to all whether or not one contributes. It is in everyone’s interest to ‘free ride’. While, on the one hand, the ‘club’ view of government as a mechanism by which individuals voluntarily agree to be coerced

---

**Figure 1**

**Consumption Externalities and Selective Subsidies**

[Diagram showing consumption externalities and selective subsidies with labels for Price and marginal value, \( P_a \), \( P_b \), \( P_c \), \( MV_a \), \( MV_b \), \( MV_c \), and EMV, with points \( x_c \), \( x_a^* \), \( x_b^* \), and \( x_A = x_a^* \).]
to provide more optimal levels of public goods can be seen as a collective solution to the free-rider problem, it is quite clear that there are several types of activity (most notably private charitable giving) that are not collectivized and that are fully exposed to the problem of free-riding. Strictly, the free-rider problem implies that no-one (save one whose marginal value of the external benefit most exceeds its marginal cost) has an incentive to contribute. Consequently, charities will have either no supporters or only one. Since this is plainly not so, there seems to be something wrong with the theory. And if each is therefore a wrong theory, none can be used to justify the kinds of public subsidy often seen in health care. The free-rider theorem is an example of ‘over-kill’: its implications are too strong. The problem is further compounded by the well-known puzzle that a democratic collectivist solution requires individuals to vote. Yet the same welfarist free-riding argument predicts that no one will ever vote.

One solution to this sort of problem has been offered by Collard (1978) among others, which appeals to the rationality rather than the public spiritedness of individuals. This is the Kantian principle: ‘if the interest of the action can without self-contradiction be universalized it is morally possible’ (Kant, 1930). A non-Kantian altruist will consider only his own (usually) negligible contribution to the financing of public good and free ride. If all are non-Kantian, all free-ride and the classic prisoners’ dilemma result ensues. The Kantian altruist, per contra, reckons with the behaviour of others: if all do as he does and he free rides he knows (rationally) that the worst outcome will result, so he (and all other Kantians) behave morally and pay up. While it is in many ways attractive to suppose that rules of morality actually may affect behaviour, it is plain that this approach is, in the welfarist context, ad hoc. It is, in fact, extra-welfarist.

Another, rather different, attempt to find a solution is to posit that individuals attach utility not only (possibly not at all) to changing the consumption of goods and services by others, but to the act of contributing itself: utility is derived from what one gives away rather than that what is used for. This would get one quite neatly out of the free-rider problem, for the source of utility is no longer public: if B gives to the charity rather than A, A derives no utility. However, this seems to lead nowhere if one wishes to build a welfarist model of collective health care subsidies, for it is the very publicness of the external effect that generates the public subsidy argument. Perhaps a mixture of what Margolis (1982) calls ‘goods altruism’ and ‘participation altruism’ might be developed that helped to account for the simultaneous existence of collective and private acts of altruism.

Another problem has been pinpointed by Sugden (1980, 1982) again attacking the roots of the theory of welfarist altruism. Again reverting to the private charity case, suppose you have decided to donate £100 to a charity. Under welfarism, you have selected this as your ‘best’ choice given your income, what you know about the charity, etc. You are on the point of writing out your cheque when you discover that your neighbour has just posted a cheque for £100 to the same charity. Since the charity is now as well off as you thought it would be with your own contribution, but you can now be £100 better off, your welfare is undoubtedly higher if you no longer write the cheque. In fact, since you are a bit richer than before, and charitable giving has a positive income elasticity, you may want to revise your initial view about how much to give slightly upwards, but even if you gave £10 instead of nothing, 10 per cent is a very large proportion of extra income (£100 in this case) for anyone to give away (see Collard, 1978 for a survey).

Put it even more dramatically. Suppose the charity’s annual income is £10,000 and you are contributing £100 each year out of personal income of £1,000 (you are unusually generous). Now suppose, entirely by coincidence, that your income falls by £1 at the same time as everyone else’s contributions rise from £9,900 to £9,901. With your preferences and the terms of trade constant, your initially preferred combination (£10,000 for the charity and £900 for yourself) is still available and will therefore still be preferred so you give £9: a fall in an altruist’s income that is exactly matched by an increase in everyone else’s contributions means that an altruist, as welfarists see him or her, will choose to reduce his or her contribution by the full amount by which personal income has fallen!
Once again, a theory that has such counter-intuitive (and counter-factual) empirical implications has to be regarded as an unusually weak foundation upon which to build normative propositions. Yet this is the same theory that underpins all these externality arguments for public subsidy (of one form or another, depending on which of the rival externality theories one is using).

Sugden suggests that it is preferable to invoke another (extra-welfarist) theory (specifically, of 'duty'). Unfortunately, he does not develop this in the context of individual charitable behaviour, let alone use it to mount any arguments (welfarist or other) for the welfare state or 'free' or subsidized health care, etc. It is worth noting, however, that it was precisely in the language of duty (as well as altruism and reciprocity) that Titmuss couched his defence of the NHS (Titmuss, 1970). This may not be the welfarist way—but it may be (pace all the economists who attacked Titmuss) the right way! Thus we reach a rather destructive end of a chain of welfarist attempts to build rationales based on altruism or philanthropy.

One interesting possibility that may yet rescue welfarism has been developed by Margolis (1982). He postulates that individuals have a split preference system in which one set of preferences relates to group-interests and another relates to selfish ones. This develops hints dropped by, for example, Harsanyi (1955), Pattanaik (1968), Meade (1973), and Rawls (1972), in which the ancient distinction of Plato between man as citizen and man as individual is developed in various ways, but each having the characteristic that one (higher) set of 'preferences', or morals, constrains or interacts with another (lower) set.

Let, then, a representative individual derive utility from s, expenditures on self, which are subject to the usual diminishing marginal valuation and from g, expenditure by the individual on the group (participation altruism). Since the individual's contribution is small relative to the group's as a whole, the marginal utility of g can be taken as constant. Thus, as g increases (and s correspondingly falls) the 'participation ratio' g/s rises, and the G'/S', the marginal rate of substitution (marginal utility of g-spending to that of s-spending) falls. What determines the individual's preferred balance of g and s? Margolis suggests that the higher is g/s the greater the weight given to s and the higher the ratio G'/S' the greater the weight given to g. This he terms the Fair Shares principle (FS). If the weight given to g relative to s is W, then a stable equilibrium occurs at E in Figure 2, with the consequent equilibrium of g occurring at g*.

As a model of altruistic behaviour, this has several attractive welfarist features. Since it does not appeal to externality arguments, it is not flawed by free-rider problems, nor does it have any of the Sugden anomalies. It does not require that the individual be indifferent to the uses to which g is put (which are determined by the individual's preferences over the set of possibilities that exist in the usual utility-maximizing way once the size of g has been decided). It does not elevate g-preferences to a higher status than s-preferences (they are both preferences—one is not a set of 'moral rules') so it is amenable to the usual welfarist interpretations. The W function, while possible to derive from some form of 'super' welfare function, is not required so to be derived (and Margolis advances some attractive reasons for not deriving it in this way). It gives rise to the possibility of private charities (since gains from trade are implied by the g-preference function) and, via scale economies and the saving of transaction costs, to governmental social spending and 'coercive' taxation.
The implications of the model have not been developed specifically for the health care sector (for a beginning, see Mooney, 1986) but, given the weakness of other welfarist models, it may be that it holds out the most promise for a satisfactory such model (Margolis himself developed it in a positive context, a positive-normative flip is thus required as was noted before in connection with Lindsay, 1969 and Culyer, 1971b).

(v) ... and Where Does this Get Us?

Since the literature on externalities and 'caring' is largely a priori in nature, its value is primarily as an aid to introspection and judgement. One needs first to form a view about the existence of the general events that may fall into this class by imagining a state of the world devoid of attempts to optimize via subsidy and regulation, and to determine whether externalities are Pareto-relevant in total or at the margin (Buchanan and Stubblebine, 1962) or whether g-orientation is likely to be pervasive. Beyond this there are questions as to whether group concerns are better conceived as relating to absolute or relative levels of consumption and whether they are better seen as questions of preference (arguments of utility functions) or of morals (constraints on selfish welfare maximization). These rather useful sorting devices are further supplemented by the enrichment the literature affords of the concepts of 'altruism', 'selfishness', and 'caring'. As an aid to clarity of thought, the welfarist literature is valuable; it cannot be claimed, however, that it has yielded settled conclusions on those issues I have identified as requiring judgement. The view that the externality or group concern is derived not from health care consumption but from 'health' itself is not directly addressed in this literature but occurs in what I shall discuss later in the extra-welfarist section.

(vi) Uncertainty

The stochastic nature of disease and ill-health has been held to be another respect in which health care is 'different' from at least some other goods and services. The standard welfarist economics of health insurance (e.g. Arrow, 1963; Culyer, 1979; Evans, 1983, 1984) is based on the expected utility maximizing model of risk-aversion (for a survey see Schoemaker, 1982). The aspect of health-affecting events that is insured against is the financial cost only of medical care (loss of employment income is not considered here). In this analysis, the welfare gain to the insured risk-averse individual arises from the implication that the welfare loss from paying a certain premium is less than the welfare loss of the expected financial loss (having the same actuarial value as the premium) of the
uncertain prospect if the individual ‘self-insures’ by retaining net income in health and by paying medical expenses out-of-pocket if sickness strikes. This in turn arises because the marginal welfare loss from having to pay an uncertain but large sum is greater than that from having to pay a certain but smaller sum, due to the assumption of a diminishing marginal utility of income or wealth. It is usually assumed that the utility of income is independent of the stochastic states ‘healthy’ or ‘sick’. If it is not, it is not clear what the optimal level of insurance is (Shavell, 1978).

Thus Figure 3 shows an individual’s welfare as a function of income (diminishing marginal welfare). Let the individual’s income be $I_x$, yielding a welfare level $W(I_x)$ on the vertical axis, and consider a choice between insuring against an uncertain event (illness), which will require an expenditure of $H$, or not insuring (self-insurance). If the event occurs, expenditure will be $H$, income net of expenditure $I_x-H$ and the associated level of welfare is $W(I_x-H)$. Suppose the probability of the event occurring is known to the individual and insurance companies is $p$. With self-insurance, income net of expected health expenditure is $I_x-pH$. The expected welfare of this option is $pW(I_x-H) + (1-p)W(I_x)$: the sum of the expected welfare if sickness occurs plus the expected welfare if it does not. The point $c$ on the chord ab is located proportionately to $p$. If $p$ is, say, 0.4, then $bc/ac = 0.4$.

If an insurance firm offers insurance at an actuarially fair premium $pH$ (the cost of the event times its probability), the welfare of the individual paying the premium is $W(I_x-pH)$. The choice confronting the individual is thus to self-insure or to purchase insurance, with payoffs as in Figure 4.

It is clear that, with diminishing marginal welfare from income (risk-aversion), welfare is higher with insurance, at d, than without it, at c. The certain monetary loss $pH$ is preferable to the uncertain loss of $H$ with the same probability. Fair insurance thus increases welfare.

The important question for health care financing is, of course, whether a voluntary competitive health insurance market is better able to maximize these welfare gains than a compulsory system of social insurance or tax finance.

(vii) Loading

The analysis so far assumes that there is a competitive insurance market in which many insurers compete for trade (or potential entrance deters oligopoly premium-setting) and premiums are set at actuarially fair rates: the product of the probability and the estimated expense. In practice the genuine opportunity costs of insurance provision, together with any X-inefficiency and monopoly rent, will also be ‘loaded’ on to premiums which consequently rise above their actuarially fair level. In large group plans in the USA the loading is about 10 per cent (Pauly, 1986). This implies that the welfare gains for consumers of insurance will fall as premiums rise and some risk-averse individuals will no longer insure.

Let loading charges, $L$, be proportional to the premium $pH$. The premium now becomes $(1+L)pH$ and the insurance outcome (in this case) now has a lower welfare for the individual, $W[I_x-(1+L)pH]$, than self-insurance. Any loading large enough to bring the individual on to the section of the welfare function below $f$ on the horizontal through $c$ will reduce welfare relative to the self-insurance option. Any loading less than $c$ will not.

To the extent that the loading represents costs in excess of opportunity costs, this will thus cause deadweight welfare losses. To the extent that
opportunity costs themselves can be lowered by scale economies and other measures that might be taken to reduce costs, the competitive insurance system itself may thus cause unnecessary welfare losses.

Loading also leads to a preference for deductibles (Arrow, 1963). Suppose, as seems realistic, that the loading is not simply proportional to the size of the premium but is a relatively high proportion of small claims (say, checking for fraud, etc., is equally costly for probable and less probable events). Then, as the limit of acceptable (to the consumer) loading is less than \( fc \) in Figure 3, as \( b \) is approached and as the cost of the event falls (point \( a \) is closer to \( b \) on the utility function) the probability of the loading exceeding the welfare gain rises: very probable events (e.g. routine dental care) will be more self-insured via deductibles and very expensive events (e.g. cancer care treatment) will be more likely to be fully insured (no deductibles).

(viii) Adverse Selection

Another problem that arises has its roots in another kind of informational asymmetry—this time, however, the informational advantage lies with the consumer rather than the supplier of the service (financial in this case). Premiums are set according to the calculated risks for groups in the community. In the simplest case of ‘pure’ community rating, the premium is set according to the population-wide probability of consuming health care. The typical consumer will, however, usually know better his or her own probability. Those whose probability times expected expenses is greater than the premium will gain welfare in excess of that expected in the model (viz. \( > cd \) in Figure 3) and may gain even if they are not risk averse! Those whose true probability times expected expenses is less than the premium will tend (depending on their risk aversion) to self-insure. This phenomenon is known as adverse selection. It leads to a progressive upward pressure on premiums (as low potential users are driven out), with insurance cover increasingly restricted to the very worst risks. The upshot is a pool of uninsured people, many of whom are risk averse, in addition of course to those who are uninsurable by virtue of chronic disease or other disqualifying features, and a heavy financial burden on the sickest members of society who remain insured. In the United States, estimates vary of the numbers who are uninsured or inadequately insured, but seem to indicate a total of around 50 million (Farley, 1985; Mundinger, 1985). In any event, adverse selection, unless checked, is likely to cause massive externalities of the sort described earlier and become a major affront to most principles of equity.

Left to itself, adverse selection would destroy the market entirely. In practice this does not happen, partly because insurance plans are often based on employee groups in which a condition of employment may be that each joins the company plan; so opting out is not possible unless the gains from doing so would outweigh the losses from changing one’s employment. Partly it does not happen because the presence of adverse selection itself affords gains from trade between self-insurers and profit-seeking insurance agencies (and non-profit seeking too).

It pays insurers to offer policies with premiums that better approximate the true expected expenses of those driven out by adverse selection, even though the acquisition of this information is not costless. This erodes the principle of community-rating which becomes replaced by experience-rating and by packages of cover that fall short of fully comprehensive (e.g. ‘major medical’ only). In this fashion ‘good risks’ are creamed off. While there will remain some inefficiency in that some very low risk groups may still fail to find a suitable package at an actuarially appropriate premium, experience-rating is more efficient than community-rating. It differs also from community-rating in that, whereas under community-rating wealth is redistributed from low to high risk individuals and families in advance of any health care consumption, under experience-rating, there is redistribution \textit{ex post} from those (insured) who are well to those who are sick (as does all insurance).

Though more efficient from a welfarist perspective (Arrow, 1963), experience-rating is likely to violate the usual distributional criteria of equity: those with a history of sickness will face the highest premiums, have the less comprehensive cover, be most likely to pay deductibles (both of these latter options have the effect of reducing premiums as they reduce probable pay-outs by the insurer)—and, since ill-health and income are correlated, will on average be the poorer members of the community.
Many, moreover, can be expected to find their risk-aversion insufficient to warrant insurance of any kind.

(ix) Moral Hazard

Another difficulty with insurance systems is known as moral hazard. One form of this has already been met in the shape of supplier-induced demand, which is a kind of producer's moral hazard, whereby producers (on a fee for service system of reimbursement) have an incentive to adjust the client's demand in pursuit of their personal income objectives. This is easier done (or at least is less likely to offend the doctor-patient relationship) if the cost of supplier-induced demand can be passed on to the insurer.

More commonly discussed in the literature, however, is consumer's moral hazard: in which the fact of being insured encourages the individual to take less care in ensuring that the undesired state (illness) does not occur (ex ante moral hazard) and, when sickness occurs, encourages the consumer to maximize the consumption of services beyond the point at which marginal cost (assuming that services are priced at marginal cost) equals marginal value (ex post moral hazard). In health insurance, ex post moral hazard seems to be the principal problem on the consumer side. Two caveats are worth noting here. One is that consumer demand is, as we have already seen, mostly interpreted by an agent, so the distinction between consumer's and producer's moral hazard is at best fuzzy. The other is that if hospital and other services are priced at above marginal cost, second-best considerations would dictate that the usual welfarist efficiency condition of \( P = MV \) may no longer apply. Consumers may be receiving false signals about marginal cost from market prices and should probably be encouraged to consume beyond the rate at which \( P = MV \), quite apart from any externality considerations, even if this implies even bigger rents for monopoly suppliers of health care, as income is redistributed from the insured to the suppliers via the insurers.

In any event, the effect on the insurance market is predictable: with moral hazard, consumption exceeds the rate at which premiums have been (historically) set, the (retrospective) reimbursement of suppliers rises to levels higher than predicted, and premiums rise. Higher premiums will drive out some risk-averse individuals from the insurance market (dead-weight losses) and may (but only 'may') cause an insufficiently large rate of consumption.

To retain their market, insurers offer packages that do not include 'first pound coverage' and the consumer has to pay deductibles (e.g. the first £x of any expense) or coinsurance (e.g. x per cent of the total bill). These out-of-pocket payments in the US currently amount to about 30 per cent of non-hospital expenses and 10 per cent of hospital expenses.

Figure 5

The Welfare Loss of 'Excess' Health Insurance

![Diagram showing the welfare loss of 'Excess' Health Insurance](#)
While the likely offence to equity principles is plain, the efficiency implications are less obvious in a welfarist context. With full ‘first-pound’ cover, and zero marginal (money) user price, the sick individual may be expected to consume QQ units of health care in Figure 5, given that he or she is sick, at which point MV = 0. This generates a welfare loss of abQ: the amount by which the cost of providing QQ units exceeds their value to the individual (qqabQ − q*aQ). This analysis assumes that consumer marginal willingness to pay is adequately reflected in the demand curve and that OP is the marginal social opportunity cost.

The excess ‘burden’ can be reduced by co-insurance. Thus, if the consumer pays a proportion, p, of the daily cost P, and that is 50 per cent, user price becomes pP, consumption falls from Q to QQ (closer to the ‘optimum’ QQ), the excess burden falls by more than half to acd, and total expenditure falls from OPbQ to OPcQQ implying a lower (future) premium.

In one of the few thorough empirical attempts to measure welfare effects in the literature, Feldstein (1973) estimated that the maximum reduction in the excess burden of health insurance in the US, by raising co-insurance from an average of 0.33 to 0.5, would have been $10 billion in 1969 prices after allowing for the loss of welfare to those who would no longer choose insurance. However, the implications of this analysis hinge crucially on the adequacy of prices as measures of marginal cost and on the absence of externalities. They also depend on the welfarist assumption that MV adequately represents the consumer’s true estimate of the worth of medical care. (The actual calculation was also based on elasticity estimates that were almost certainly too high by a large factor.)

(x) . . . and Where Does this Get Us?

Insurance is shown plainly to have welfare-increasing properties though it seems clear also that optimal insurance may be less than complete insurance: some events should be uninsured and for others less than 100 per cent of the risk will be optimal. This analysis does not, however, take account of any contrary indications suggested by externality or equity factors and is entirely dependent on the strong assumption that consumer choices adequately reveal their preferences (viz. that the agency relationship is perfect).

Provided that it is not accompanied by compensating X-inefficiencies, a system of compulsory universal public insurance operated through the general tax system may have substantial cost advantages over competitive insurance: compulsion avoids adverse selection and enables scale economies to be gained (if they exist), reducing loading and increasing the welfare gains from comprehensive insurance cover. Universality and tax finance avoid the necessity of risk assessment and premium setting, billing, reimbursement, checking for fraud and so on. These costs appear to vary substantially across different systems, being less than 3 per cent of total expenditure in countries with a tax-based social insurance system, like the UK and Canada, and in excess of 10 per cent in private insurance systems, or in systems with public finance but with complex systems of billing and reimbursement (e.g., USA and France) (OECD, 1977). It may be technically possible for a large country (like the USA) to operate a competitive system that would avoid monopoly exploitation and that would also realize scale economies (though this does not seem to have happened) but may be implausible in smaller countries like the UK. Moral hazard may be controlled in public or private insurance systems by the adoption of (private or public) regulatory schemes defining the appropriateness of packages of care, and by prospective reimbursement according to an agreed charge per case-type rather than retrospectively according to whatever the suppliers happen to have charged. Given a non-competitive health care industry (whether public or private), such regulation is commonly observed, and it is a question of judgement whether the regulation is better done by a publicly accountable agency or by the (insurance) industry itself.

(xi) The Welfarist Approach: an Overview

The welfarist techniques of analysis have been the traditional way in which issues of health service finance have been addressed in the economics literature. Granted the acceptability of welfarism’s value assumptions, its implications for policy hinge on judgements about the empirical significance of consumer rationality, the ‘purity’ of the agency relationship, the nature of any externalities (physical or utility interdependence or group-concern),
the extent of adverse selection, moral hazard, supplier induced demand, unnecessary premium loading under insurance and the empirical validity of the neoclassical behavioural model that, in its normative version, is welfarism’s centrepiece. Each will draw his or her own conclusions based on the (patchy) evidence and more casual experience.

III. THE EXTRA-WELFARIST APPROACH

Whereas welfarism holds that standards of living, the efficiency of social arrangements and the justice of distributions and redistributions are all to be evaluated in terms of individuals’ utilities (or welfares), an extra-welfarist approach (Sen, 1979) admits non-utility information about individuals into the process of comparing social states. Using an illustration from Sen, consider three social states x, y, and z with the following (cardinal and even interpersonally comparable) utility numbers for persons 1 and 2.

<table>
<thead>
<tr>
<th></th>
<th>x</th>
<th>y</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>person 1’s utility</td>
<td>4</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>person 2’s utility</td>
<td>10</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>

In x, person 1 is hungry while 2 eats amply. In y, 2 has been forced to surrender some food to 1 and 2’s utility loss is less than 1’s utility gain. Under welfarism, the fact of coercion is an irrelevance: the sum total of utility is higher in y than x so y is socially preferred. This is one example of the exclusion of non-utility information in the making of social comparisons. In z, person 1 is as hungry as in x and 2 is as amply fed but 1, who happens to be a sadist, is allowed to torture 2 (who is no masochist). It so happens that 1’s utility gain and 2’s utility loss are the same as under the food transfer programme. Under welfarism y is socially preferred to x but z is the same, utility-wise, as y, therefore z too is socially preferred to x. Again, the relevance of non-utility information (in this case the fact of torture and my disapproval of it) is denied.

Sen (1980) argues that a particularly important class of non-utility information about individuals is ‘basic capabilities’: a person being able to do certain things. It is, he suggests, because a cripple is unable to perform particular activities that he or she is seen as having special ‘needs’ that are independent of his or her total or marginal utility. Culyer (1989) advocates the more general notion of ‘characteristics of people’—for example, their genetic endowment of health, their relative deprivation independently of the absolute consumption of commodities or the characteristics of commodities, their moral ‘worth’ and ‘deservingness’, whether or not they are in pain, or stigmatized by society. Characteristics may also relate to the character of relationships between people such as the quality of friendships, community support for the individual when in need, social isolation, or changes in them, such as becoming (as distinct from being) crippled.

Only some of the characteristics of people (which will include some of their capabilities) will be deemed relevant and the list of such relevant characteristics is likely to vary between cultures, climates, historical periods, and so on. It is, in short, contingent. It is related to a concept of need. If the characteristics of people are a way of describing deprivation, desired states, or significant changes in people’s characteristics, then commodities and characteristics of commodities are what is often needed to remove the deprivation or to move towards the desired state, or to help people cope with change. They are the necessary means to a desired end. To compare the ill-health of different individuals or groups is not the same as to compare the health care they have received (they could receive the same amounts and still be unhealthy, or different amounts and be equally healthy).

(i) Need and Extra-Welfarism

Whereas the notion of ‘need’ has received a bad press from many welfarist economists, extra-welfarists have been able to use the term with some precision and confidence. In its most rudimentary form ‘need’ for health care would seem to imply that someone is better off with the ‘needed’ treatment than without it (Williams, 1974) and that the ‘better offness’ has to do with persons’ health (Culyer, Lavers, Williams, 1971). Thus health services are needed (viz. are a necessary condition for achieving a particular outcome) only if the outcome is desired and there is no alternative (or
more cost-effective) way of realizing it. Since inputs are nearly always substitutable, it will not normally make sense to say that a specific resource in a specific quantity is needed. Since there is no effective treatment for some conditions, it is nonsensical to say that persons suffering from such conditions need health services (they may need the fruits of research and they may need love and comfort, but they cannot need ineffective care, even though they may demand it, and such care may also damage health). Since health services are needed only for what they enable to be accomplished, in a world of scarcity judgements must be made about the value of what might be accomplished. Some services are bound not to be supplied in the light of such judgements and so it does not make much sense to require that all needs should be met. One of the first lessons of ‘needology’ is the lesson of the ethical acceptability of unmet need (Wiggins et al., 1987).

Since health services can be needed only if their outcomes are desired, the important question arises as to whose judgements ought to be decisive in assessing desirability of outcome. While the judgement of technicians (such as doctors) may be appropriate in evaluations of effectiveness, technical experts have no particular authority for making value judgements so it does not make much ethical sense to pretend that these essentially political decisions can be desanitized by being left to the experts. At the centre of this problem lies the issue of how one person’s needs (viz. the potential improvement in his or her health attributable to health care) are to be weighed against another’s, an issue taken up below.

In health economics, the extra-welfarist approach has taken ‘health’ as the proximate maximand. This does not imply the complete ousting of ‘welfare’ with its usual normative connotations, but the use of both sets of ‘data’ to evaluate alternatives. ‘Health’ is itself a descriptive characteristic of people, but in practice has been interpreted in the literature as a composite ‘bundle’ of other characteristics such as pain and restriction of activity (Culyer, Lavers and Williams, 1971).

The implications of the extra-welfarist approach for the finance of health care seem partly to take negative forms: insurance arrangements, user prices (money, time, etc.) should not act so as to discour-

age use of care that contributes to the objective of maximizing health. There are also implications for rationing care (equalizing marginal products in terms of health per unit of resource), selecting patients from waiting lists, conducting cost-benefit analyses in the health service.

At the core of the extra-welfarist approach is, of course, the issue of how the maximand is to be measured. This has proved to be an issue involving much cross-disciplinary collaboration between (for example) economists, physicians, psychologists, and political scientists which has exposed the complexities involved in defining ‘health’ and measuring it, and which has also developed a battery of experimental techniques designed to test theories and quantify the hitherto unquantified.

(ii) Measuring Health

Extra-welfarists identify ‘health’ as the principal output of health services and its efficient production (technical efficiency, cost-effectiveness) as an issue upon which its insights can be particularly valuable, in contrast to welfarism, under which it is natural to take goods and services as the natural units of output. Much of the cost-effectiveness literature in health economics is implicitly extra-welfarist in seeking to identify the least-cost method of delivering a given health improvement (or prevention of deterioration) for a given patient group (or across patient groups). In many cases, extra-welfarist economists have joined hands with non-economists in the search for better measures of health and, of necessity, in identifying the production functions that underlie cost functions and cost-benefit relationships.

One approach to health measurement has been to use ad hoc numerical scales to quantify the bundle of characteristics (usually in the context of planning exercises or studies of the effectiveness of medical procedures). This approach involves the individual being assessed in several dimensions, with numbers being associated with each assessment and the resultant scores (sometimes weighted) being added up (e.g. Harris et al., 1971; Grogono and Woodgate, 1971). The arbitrariness of such procedures has been spelled out several times (e.g. by Culyer, 1978a and b; also see Drummond, this issue):
(1) criteria for selecting characteristics were usually unspecified

(2) the scaling systems often implied only order but were subsequently used to construct a cardinal index

(3) the possibility that combinations of characteristics may have higher or lower numbers than the sum of the separate scores was often excluded

(4) increasing marginal severity was rarely allowed

(5) criteria for selecting those making these (value) judgements were usually unspecified.

However, such measures present a potential problem. The literature frequently characterizes measures of health as utility measures (e.g. Torrance, 1986). Here it is important to distinguish between the welfarist notion of utility as welfare and the extra-welfarist notion. Under Paretianism, for example, the notion of welfare relates to goods and services and is the utility of the individual affected by their consumption. Under extra-welfarism, while this notion of utility may still apply, there is the further idea that uses utility theory in order to derive measures of characteristics of individuals that are not goods, not services, nor necessarily having a value content that corresponds to the Paretian notion that ‘the individual is the best judge of his/her own welfare’. Confusingly, however, these too are called ‘utility’ measures. The use of the standard gamble based on Neumann–Morgenstern assumptions described above is a good example. The possibility that one can have utility measures that are not welfarist is thus important to appreciate.

(iii) “Utility” is not ‘Utility’ is not Utility ...

A related puzzle that pervades some welfarist theory is the meaning to be attached to the word ‘utility’. While it is commonplace to distinguish between utility as ‘welfare’ and utility as ‘an index of choice’ in normative and positive analyses respectively, some have difficulty in distinguishing (in a welfarist context) between utility-maximizing behaviour that is altruistic and that which is selfish: is not one who maximizes utility by giving, in some sense, selfish? In positive economics this poses little difficulty: altruism means simply that a person’s allocation of wealth is influenced by the effects that the allocation has on others and that the person in question has benign intentions. Since both behaviour and motive are different from that of one for whom neither of these things is true, this is a distinction that needs to be made. It fails, however, to capture another distinction: that between fulfilling one’s selfish preferences and acting contrary (at least in part) to those preferences by adopting a set of moral rules that may constrain. Such altruism, that is contrary to preference, is of a different kind and, perhaps, requires an appropriate term that differentiates it from purely ‘welfarist altruism’: it is, after all, extra-welfarist.

The ways in which the quality weights have been developed and interpersonal comparisons made are of particular analytical interest. A historical review of the health index literature (of which these quality weights form a part) is Rosser (1983) and a review of the techniques for measuring health indices is Torrance (1982).

There are three commonly-used methods of measuring scales of health (on this and other important distinctions see Culyer, 1986): the rating scale, the standard gamble and the time trade-off (see Drummond et al., 1987). These are defined in detail in Drummond (this issue).
As a result, and given the difficulties in defining the relationship between health measures and utility measures discussed above, one of the normative issues that the extra-welfarist approach identifies (but does not resolve) concerns who shall decide the weights to be applied to different health states and to the components of health states. Who shall decide the categories of functioning, etc. to be considered? Who shall decide who shall decide?

The answers may well depend upon the nature of the problem under consideration: politicians, civil servants, managers, representatives of the public, persons at risk of particular disease, patients, doctors, nurses . . . all may have some claims by virtue of identity, skill, or position of trust. It does not, however, follow that those judged best able to exercise a judgement about, say, the effectiveness of a medical procedure are necessarily those best qualified to exercise a judgement about how pain and disability are to be traded off.

(iv) Extra-Welfarism and Distributional Weights

The non-welfarist literature clearly sees information of the cost-per-QALY sort as a means of guiding resource allocation decisions in the NHS, whether on equity or efficiency grounds, and the literature is quite explicit in its departure from consumer-based willingness-to-pay as the basis for benefit valuation (Williams, 1988).

The question that naturally arises next, of course, relates to whose willingness-to-pay is to be substituted for the consumer's. The most recent work has addressed this issue in the following way: ' . . . is a particular improvement in health to be regarded as of equal value no matter who gets it; and, if not, what precisely is its relative value in accruing to one kind of person as opposed to another?' (Williams, 1988). The current work has begun by seeking to find out what views are actually held by surveyed individuals on the matter. It would seem from the early results that there is a consensus that particular phases of the life-cycle are regarded as times when health is of greatest value. 'Two stood out of the ten phases used in Williams' survey: 'as infants' and 'when bringing up children'. The information available to date is only qualitative and a next phase of the work will address the issue of how much good health is worth at points in the life cycle relative to others. It is also planned to unscramble the life-cycle phases into elements that are age-, role-, and sex-related.

If ethical authority is to be accorded these (or similar) results, a departure is implied from the distributive value-judgement normally (if only provisionally) embodied in measures of health such as QALYs: that a unit of 'health' is of equal value no matter who gets it. It thus seems possible that distributive judgements will be able to be built into outcome measures and, via cost-effectiveness analyses, into efficiency analysis (e.g. Culyer, 1988b). Indeed, if all of the features of distributional equity that are of (legitimate) concern are built into outcome data in this way, then a full integration of equity and efficiency will have been achieved in health policy; given routine information about population characteristics (disease incidence, etc.), medical technology (the possibilities for changing health states for the better), and cost information, it will be possible to make informed routine judgements about resource allocations to providers of health care.

(v) Objections to QALYs

There is, however, quite a long way before that dream becomes a reality—and some have seen it more as a nightmare possibility. One type of objection is represented by Smith (1987) who argues that the use of a quantitative algorithm obscures a process by which essentially arbitrary assessments of the values of people's lives are being made. To this Williams' (1987) retort seems compelling: that far from obscuring the need for value judgement the procedure highlights the value-judgemental elements and offers techniques by which they could be made more explicitly; that far from imposing essentially arbitrary values, the process of quantification is, by virtue of its explicitness, open to criticism and change at every stage. Indeed, it is hard not to see Smith's objections as being much more aptly directed at present modes of making decisions (e.g. about health service cuts across the board).

Rather deeper is an objection of Broome (1985) that the measures ignore 'population': the outcome of health services may include not only the extra health for people who may also be having extra time, but also more people if an additional outcome
includes children being born who would not otherwise have existed. In the welfarist tradition this poses a problem because although it might be possible to find out how extra QALYs are valued, it is not possible to find out how the unborn value being born. But this also poses a problem for the non-welfarists in so far as they too wish to ascribe a value to unborn lives—as well as the lives of the children of the unborn, and so on ad infinitum. Unfortunately, we have no basis at present for valuing population changes, therefore (says Broome) we have no basis for valuing life or QALYs.

Earlier Broome (1978) argued that the only appropriate value for a 'statistical' life was infinity on the grounds that, eventually, statistical probabilities of death (or of opportunities for life extensions not taken advantage of) translate into deaths of actual individuals who might reasonably be expected to exercise a veto. This clearly poses something of a challenge to Paretian welfarists though less of one, of course, to non-welfarists. A welfarist may find something of a defence in the reflection that he might himself agree to an option offering some benefit but with a very small prospect of its entailing his own death, so why should not a society of like-minded folk feel similarly? An extra-welfarist might take the view that she would be guided by the majority view on the value of (or differential values of) life.

It is worth, perhaps, reminding everyone (welfarist and extra-welfarist alike) that there is nothing to be gained in the context of resource allocation decision-making from taking an ontological view of QALYs, or life, or lives. One is not concerned with the inherent cherishable worth of people but rather with the value of resources that we might spend in order to gain better health or prevent (or postpone) death or change the prospects of either for the better. If we spend £2,000 per person to protect them from the consequences of some risk that is fatal for say one in five hundred, it is merely arithmetic that we shall spend, on average, £1 million per life saved: only in this sense is a life 'worth' £1 million.

Loomes (1988) mounts a far more powerful assault on both welfarist and non-welfarist traditions by attacking the usual behavioural axioms that are shared in both. He focuses on the systematic differences in measure that have been observed as between rating scales, standard gambles, and time trade-offs (e.g. Torrance, 1976; Bombardier et al., 1982). His analysis hinges on a distinction between the utility gained as the result of one's own choice and 'choiceless' utility: viz. the utility 'experienced' as the consequence of a happening generated in any other way. The importance of the distinction lies in the fact that only in the former is there the possibility that, in an uncertain world, you may come to regret or rejoice over a decision you have made. The rating method does not involve choice; it requires the subject only to locate a state on a utility scale. The standard gamble, by contrast, does involve choice and so does the time trade-off method.

'Regret theory' is recommended as the prima facie better foundation for future work in the QALY territory. It seems clear that research should be expanded to incorporate regret theory into health status and QALY measurement experiments in order to compare results systematically with the other techniques. The potential from exploiting other substitutes for expected utility theory (such as prospect theory) remains to be explored.

(vi) Where does Extra-Welfarism Get Us?

As far as the demand side is concerned, extra-welfarism in health economics may be seen to take 'health' output as the maximand. The emphasis is not in principle exclusive, as extra-welfarism is not exclusive, and it seems unlikely that any extra-welfarist would assign zero weights to such factors as consumer choice, privacy, speed of service, hospital hotel-services, and other factors that may be not only remotely causally linked to health.

Extra-welfarism thus immediately implies another notion of efficiency, in which explicit (but not welfarist) value judgements are incorporated into the maximand which is a cardinal 'utility' index of health (the welfarist tradition, the Margolis model, and the insurance literature are also cardinal in the sense that 'utility' as used there is measurable up to a linear transformation). This index is extremely useful in supply-side efficiency studies. It is worth noting that it is not uniquely applicable under socialized systems of care and may, for example, be used in market systems where insurance companies
seek to control producer moral hazard by reimbursing providers only for procedures that are demonstrably relatively cost-effective in restoring health and in any clinical research with a similar objective. In this territory, extra-welfarists have been more active empirically than welfarists typically have been.

As a matter of necessity, the literature has focused on the difficult issues of measuring health itself (or changes in it) in order to improve the ability of the system to produce health cost-effectively. It is quite possible for these efforts also to serve the cause of welfarism: after all, better information on outcomes can enhance the physician’s ability to act as a good agent for the patient. But it is quite clear that other causes are served as well and that the implications of extra-welfarist health maximization are not the same as welfarist analysis.

Extra-welfarism resembles the welfarist externality arguments in implying free or subsidized terms of access to health care. Under welfarism, however, the reason for the subsidy lies in the optimal internalization of externalities. In extra-welfarism, the reason lies in more ‘engineering’ sorts of concern: optimal resource use is determined by equality of marginal health output per unit of resource in various activities and across various client groups. Willingness-to-pay is an irrelevance and may be directly counter to this objective if willingness-to-pay is positively associated with ability to pay, but ability to pay is inversely associated with ‘potential for health improvement’. The extra-welfarist approach therefore attaches great importance to the identification of potential for benefit. Indeed, it is possible to see some traditional policy arguments cast in equity terms as possibly extra-welfarist health-maximization arguments. Assume, for example, that a (satisfactorily measured) QALY is judged to be equally valuable socially to whomsoever it may accrue. Now allow that the sickest in society are by and large those for whom the marginal product of health care in terms of QALYs is highest, that these are also the poorest, and that when (ceteris paribus) health service per capita rises, the marginal product in terms of health falls. It evidently follows that efforts to equalise the geographical distribution of resources, to channel more of them to the sick and more of them to the poor, might be seen not as distributional policies to be justified by equity arguments but efficient policies justified by health maximization.

There is a danger in extra-welfarism of becoming too fixed on the ‘bottom line’. The great advantage the approach can claim in issues like outcome measurement is rather like the claim made earlier on behalf of welfarism: it provides a conceptual framework for handling extremely complex issues in a systematic fashion and that exposes each aspect of an argument clearly. It is less important what the cost-per-QALY is, than that individuals with responsibility for resource allocation in health care have a means of working through the issues so that they can come to their own informed view about the pros and cons of different resource allocations. It is in this sense that it is important to emphasize that the method is intended (in an archetypical ‘decision-makers’ approach) as an aid to, rather than a substitute for, thought.

IV. CONCLUSIONS

It will be clear from the foregoing that there is a paradigm clash in the normative economics of health, though it would be wrong to overdraw the differences. The extra-welfarist approach is, after all, inclusive of welfarism. There can be no question that the extra-welfarist approach is more tolerant of what may be seen as ‘ paternalism’ and can be readily enlisted on behalf of the sorts of access terms and distributional issues that have lain at the heart of the ideology of the British NHS—and increasingly of policy towards it. Extra-welfarism is also providing a theoretical basis upon which usable output measures can be derived.

Ultimately, however, neither approach can yield final answers. They provide frameworks whose usefulness remains a matter for judgement. An earlier judgement by me upon the field, ‘The heady atmosphere of grand designs has to be replaced by the mundane, but ultimately more fruitful, ground of systematically applied economics . . . In this scheme of things the role of welfare economics is to provide an appropriate theoretical base in which to build empirical studies and not to prejudice the facts’ (Culyer, 1971) needs amendment in only one respect: add ‘or extra-welfare’ after ‘welfare’.