UNIVERSITY
OF YORK

CENTRE FOR HEALTH ECONOMICS
HEALTH ECONOMICS CONSORTIUM

Health, health expenditures and equity

by A. J. Culyer

DISCUSSION PAPER 83

Health, health expenditures, and equity

by

A J Culyer

April 1991

The Author

A J Culyer is Professor of Economics in the Department of Economics and Related Studies, University of York and Medis Institut fuer Medizinishe Informatik und Systemsforschung, Gesellschaft fuer Strahlen- und Umweltforschung, Muenchen).

Acknowledgements

The author gratefully acknowledges the detailed comments of Andreas Mielck, Owen O'Donnell, Adam Wagstaff and Alan Williams on drafts of this paper, and for the comments also of other participants at the Bellagio conference. The usual disclaimer applies, of course.

Further Copies

Further copies of this document are available (£4.00 to cover costs of publications, postage and packing) from:

The Publications Secretary
Centre for Health Economics
University of York
Heslington
YORK YO1 5DD

Please make cheques payable to the University of York. Details of other Discussion Papers can be obtained from the same address, or telephone York (0904) 433648 or 433646.

The Centre for Health Economics is a Designated Research Centre of the Economic and Social Research Council and the Department of Health.

ABSTRACT

This paper offers some reasons why it may be interesting to examine the distribution of health, health care, and payments for health care. The reason for a legitimate concern for these distributions is mainly because they relate to a more fundamental concern with the distribution of health itself. It is argued that an equitable distribution must take account of the whole distribution and that it is insufficient to discuss equity in terms of minimum standards. A concept of need based on capacity to benefit from health care is clarified and applied. Horizontal equity is considered in terms of three alternative principles, whose consistency with one another and with efficiency are examined using a diagrammatic technique that enables the simultaneous consideration of various equity principles and efficiency. It is shown that equality of treatment in the senses of outcome or input are mutually consistent under particular conditions, and also consistent with efficiency.

Why health?

What motivates studies dealing with the interpersonal (or inter family, interhousehold, interclass) distribution of income, wealth, health and so on? They seem usually to have been undertaken so as to enable judgments to be made (by either their authors or their readers) of the "fairness", "justice", or "equity" of such distributions with respect to some benchmark distribution (which need not, of course, be an equal distribution). It seems reasonable to ask two fundamental questions to do with their motivation. The first is: "what is the ethical theory that underlies the study?". This issue is examined elsewhere in the Bellagio conference papers by Alan Williams and I shall not examine it further here other than to note that, since there is more than one "ideology" serving as justification (for example, utilitarian, desert-based, Rawlsian) it would be in general helpful (though no doubt rather complicated) if studies were to be so conducted as to enable equity judgments to be made from a variety of viewpoints or ideologies (rather than that only of the authors - if they have one). When one is confident of there being a consensus amongst the targetted readership, this desideratum is less compelling, of course. The second is: "why study the distribution of whatever entity it is that has been studied?". In the case of the present volume, the entities are two - health care expenditures on the one hand, and health care payments on the other. In this section I shall explore some reasons why it may be interesting to examine the distribution of health care expenditures. That of payments will be discussed subsequently.

An initial answer may be because the entity in question is itself of direct ethical interest. This view seems implicit in some of the early economic literature on externalities in health care (for example, Culyer

1971, Pauly 1971) in which it was postulated that the health care consumption of one may be a source of utility to another; that there were, so to speak, two demands for health care: that of the individual in question (for preventive care when well and for cure, prevention of deterioration, or reduction in the speed of deterioration of health when sick) and that of "the rest of society" who could be conceived of as "caring" in some sense, being "sympathetic" to, or "solid" with, the individual (or group of individuals of a particuar type, such as those falling within a particular social class). This gave rise to a variety of recommendations for subsidy of access (for example, by subsidised insurance premiums) or of consumption (for example, by lower, or zero, user-charges). These had the apparent character of equity justifications. However, their justification was actually other than that. It was, in fact, an efficiency justification, to do with the quasi-utilitarian maximisation of a (usually Paretian) social welfare function in which interpersonal distributional considerations were actually explicitly ruled out.

This approach is not strictly to do with any distributional concern of the usual sort (other than such concerns which may - or may not - arise from the preferences of individuals). It is also vulnerable to challenge on the grounds of "what's special about health care?" or "What makes health care different from, say, bicycles?" The presumption is, of course, that there is no general concern with the equitable distribution of bicycles, nor is there any obvious reason to suppose that the consumption of bicycles by one produces any direct effect on the utility of others. The vulnerabilty arises because the natural way to respond to the challenge just posed is to answer in terms of some more ultimate entity, in whose distribution one is ethically interested, and which there are grounds for believing is either affected by

the distribution of health care (or bicycles) or is correlated with the distribution of health care (or bicycles), so the latter become useful indicators or tracer elements of the more ultimate entity that is of substantive interest.

In our case, a natural contender for this more ultimate entity is "health". It is widely believed that there are some entities, of which health is but one,[1] whose distribution is regarded as more important than the distribution of other entities - which may also be personal characteristics (like attractive hair) or consumption of goods (like bicycles). Such arguments have been put by economists of rather different political ideologies (for example, Buchanan 1968 and Tobin 1970). One reason for such beliefs may be to do with the important role these entities have in enabling people to fulfil their potential as persons: if it is possible for such entities not to be distributed in such a fashion as accords with perceived (or received) notions of what is equitable, then it is natural - and indeed necessary - to enquire about their actual distribution[2].

There may be other justifications for a focus on health care than this essentially instrumental view. One such, for example, may be a concern for the distribution of overall consumption, in which the distribution of health care (or, indeed, even bicycles) would be a part of some greater whole, which may itself be instrumental towards some more ultimate end. A utilitarian, for example, might take such a view: the more ultimate end being, of course, a concern with the distribution of utilities. I conjecture, however, that the most commonly held reason for having a concern for the distribution of health care is of the former kind - its instrumentality as an agent for the improvement of health (or the minimisation of ill-health). Indeed, other

elements in the common language in which public concern for health care is couched lend support to this conjecture, one such being - as will be discussed later - use of the word "need").

If this conjecture is correct, it has an important immediate implication: if the real (if implicit) distributional concern is with health then, since health care is only one of the instrumental variables which affect it, one is logically driven to examine the distribution of those of these other variables that are deemed most significant (such as housing, nutrition, public sanitation, social services, family support, health education, ...) and for the self-same reasons as one is examining the distribution of health care, whose marginal impact may often be less than that of some of the others (see, for example, McKeown 1976 for a historical review of the relative contribution of medicine to the reductions in mortality in Britain from infectious diseases). From this perspective, then, the current European Commission inequality studies presented at Bellagio, while amply justified in terms of the question with which this section began, can only be partial. They may provide important information for those with the distributional concerns described but they do not provide all the information required.

Why payments?

An ethical interest in the distribution of payments for health care may be motivated by (at least) four equity concerns.

The first is immediately implied by having an ethical concern (whether substantive or instrumental) about the distribution of health care: if the mechanisms of payment may possibly be such as to generate an inequitable

distribution of health care, then one had better examine them to see whether there seems good reason to suppose they actually do so. Since it is presumably the structural features of payment mechanisms which create this cause-and-effect link (such as eligibility under particular schemes of insurance, premium rates, out-of-pocket payments and deductibles), the focus ought properly to be upon such important details that may affect the demand for care. However, it may also be the case that a more aggregated approach which lumped together the various contributions for relevant social groups could yield useful insights. This is particularly likely to be the case when the possibly inequitable effect of the mechanism(s) in question were itself associated with, say, employment or income status. For example, if health care insurance were available only for the employed and their families, then employment status could be used as a tracer indicator of inequity in the mechanism (inequitable by virtue of its causing inequity in the substantive variable, health, or the proximate variable, health care consumption). Employment is, of course, usually associated with a higher income.

A second reason for concern with the distribution of payments arises because disposable income after tax and health care payments may increase inequity in the distribution of other entities having similar "ultimate" characteristics to health or, more specifically, in the distribution of consumption of goods that enhance these characteristics (for example, housing, education). Since these goods generally have the feature that brands them as "normal" goods in economists' terminology, having a positive income-elasticity of demand (so that their consumption rises as income rises), any reduction in disposable income will reduce these characteristics and any increase in the inequity of the distribution of disposable income will increase the other inequities: a regressive payments structure, for

example, will worsen the distribution of these other characteristics.

A third reason for concern with the distribution of payments may arise from a concern to assess the overall effect of a sector, such as the health care sector, on the general distribution of income in cash and kind in an economy. If such is the case, an analysis of the "net" contribution of one sector (and indeed of each) will be of interest, in order that a view may be formed about those sectors that contribute most to inequity in an overall sense and so that a view may also be formed about which inequities it might be most important to remove. This approach requires that inequities of some kinds may be "traded off" against others; they are not lexical or absolute ethical requirements.

A fourth reason for concern with the distribution of payments is of a similarly global type. Even if one has empirically examined and evaluated the equity of the distribution of ultimate characteristics or the utilisation of resources that affect them, there may remain a residual equitable concern with the distribution of disposable income (net of taxes and all payments for ultimate goods and services). A commonly adduced reason for such a concern is because disposable income and economic or political power are thought to be positively associated and because one has a concern with the distribution of power, which may be inequitable even though the distribution of - other - ultimate goods is equitable. One may, alternatively, be concerned with the extent to which the activity of the state redistributes real income between units such as individuals or households (is it on balance regressive or progressive?) in which case a focus on payments as well as (the market value of) health care as influenced directly or indirectly by the activity of the state (but not otherwise) is appropriate. Or one may simply have an

old-fashioned utilitarian interest in the distribution of net income in order to identify ways of shifting it in favour of those groups with high marginal utilities. As is well-known, such an interest need by no means also be an egalitarian one.

Minimum standards?

If this discussion of the ultimate entities underlying distributional concerns is (at least partly) correct, the question arises as to whether one is concerned with the whole range of the distribution or only part of it. If, for example, "protection from the elements" is an ultimate entity, and is met (at least partly) by being housed, then it seems unlikely that the concern will extend beyond that range of adequacy within which there is some likelihood that someone will not be protected. What the upper limit of this range will be is doubtless contentious, requiring substantive content to be put into the notion of "protection". However that issue is resolved, it seems unlikely that it would require the inclusion of the entire range of quality and quantity of housing which income units may enjoy.

Is there such a limit in health care? There are plausible reasons for supposing not. It may be plausible to suppose that relatively minor and relatively infrequent impairments of health would be of small distributive concern. It may also be plausible to suppose that the distribution of health care which is ineffective in its impact on health is also of small distributive concern (this is in principle an important implication of the difference between taking health care or health as the distributions of substantive concern). It is much less plausible to suppose, however, that the distribution of care that has an effective impact on life expectation,

disability, pain or other distress, is of small concern. It follows that the accessibilty and use of such care are also of distributive concern and that one should get as much of health care as one "needs" rather than merely some minimum amount. The quotation marks here alert us to the presence of a dangerous - and much abused - word to which we shall have to return shortly (when the seemingly strong implication just noted will be qualified).

What is health care?

The argument so far suggests that not all health care ought to be counted in studies of distributions when the ultimate concern relates to the health of individuals and groups (for example, they should exclude any identifiable ineffective care, of which epidemiologists tell us there is a good deal). Much of the expenditure identified in the accounts as health care expenditure is not, however, even in principle addressed to the improvement of people's health. Capital spending on the lavish atriums found in some hospitals is an example. Recurrent expenditure on the hotel services of hospitals is another. If one is not concerned substantively with the distribution of expenditure on hotels then, unless there is something special about being sick in a hotel called a hospital, which makes its hotel services of distributional consequence (here perhaps a minimum standard is called for), such expenditures ought to be excluded. No doubt this is a formidable empirical task, but it is nonetheless an appropriate one if the ethical justification for taking an empirical interest in the distribution of health care expenditures is indeed (as I have conjectured) provided and motivated by an interest in the distribution of health.

Types of equity: horizontal and vertical

The distinction between horizontal and vertical equity is as old as Aristotle's Nicomachean Ethics. Horizontal equity requires the equal treatment of equals; vertical equity requires the unequal treatment of unequals (in proportion, according to Aristotle, to their inequality). One needs, of course, to identify the relevant respects in which individuals or groups are unequal, and the meaning to be attached to treatment.

There are several ways in which the issue of respects can be addressed. A broad distinction can be made between those aspects of persons that relate to health and those that relate to wealth. Candidates here considered in relation to health are:

(a) the initial or presenting state of health
(b) the need for health care
(c) the final health state: the state of health after receiving health care.

These are not always carefully distinguished and the relationships between them are taken up in some detail later in the paper. The relation that each has to horizontal or vertical equity can be shown by the following assertions (or ethical principles):

H1 Persons having the same presenting state of health ought to be treated equally

V1 Persons having a worse presenting state of health ought to be treated relatively favourably (a weaker requirement than Aristotle's

proportionality rule)

H2 Persons having the same need for health ought to be treated equally

V2 Persons having a greater need for health ought to be treated relatively favourably

H3 Persons having the same expected final health state ought to be treated equally

V3 Persons having a worse expected final health state ought to be treated relatively favourably.

But what does equal or unequal treatment mean? It might be taken to mean "the same, or more/less, value of health care resources", but this seems pretty arbitrary since it takes no account of the effect that the consumption of health care (or, come to that, of any other health-affecting resources) may have on health. Nor does it take account of the fact that some highly effective health care procedures may be cheap (small expenditure per case) and other less effective procedures may be dear (large expenditure per case). The use of expenditures as the relevant measure of "treatment" will be discussed further below. At this stage it is more convenient (as well as being consistent with my own prejudices about the appropriate meaning of "treatment") to reject this "input" focus in favour of an "outcome" focus, which seems prima facie more attuned with the idea that equity in health care is at root concerned with equity in health. I shall therefore, for the moment, interpret "treatment" in terms of the effect that that health care has on health. Thus we have, rather more explicitly:

H1' Persons having the same presenting state of health ought to be treated so that each receives the same increment of health

or

V2' Persons having a greater need for health care ought to receive greater increases in health etc.

The implications of these interpretations of horizontal and vertical equity for efficiency in resource allocation, and the consistency between the various concepts of horizontal equity, will be discussed in some detail below. A fuller discussion of vertical equity and the conflicts that may arise between the various versions of it and the various versions of horizontal equity, and the difficulties that arise when persons are equal in some relevant respects but unequal in other relevant repects, are taken up elsewhere (Culyer 1990a).

As far as wealth is concerned, or its corresponding flow concept, permanent income (Friedman 1957), there are well-known practical difficulties of measurement (see, for example, Simons 1938). A more pressing issue, however, relates to the extent to which one is concerned with equity in the payments for health care rather than with equity in the distribution of permanent income and the way the tax/benefit system of the fisc "treats" those having equal or unequal permanent incomes. The distinction is important if, in the context of equity in finance, the relevant "respects" in which equity or inequity are to be judged are defined in terms of health. For example:

Persons having the same initial health state ought to pay the same

or

Persons having a greater increase in health ought to pay more.

It seems unlikely, however, that many would be inclined to describe the relevant respects in which to judge equity in finance in such a fashion. A notable and important exception to this would be the so-called "benefit principle" which would urge the equity of payment being proportionate to benefit and which leads rather directly to the view that an efficient allocation (in which those who value health care most get the most) is also an equitable distribution. I shall set this aside here on the grounds that readers of a volume such as this probably need no further convincing that there is something inherently inequitable about an allocation of health care that depends solely on willingness to pay, if only because willingness to pay and ability to pay are positively correlated, and ability to pay and health are also positively correlated.

Propositions more likely to command assent would be (for example):

Persons having the same initial permanent income ought to pay the same regardless of their health

and

Persons having a greater initial permanent income ought to pay more

regardless of their health.

These propositions sever, of course, any link between receipt of care, or health, and health care payments. Further, since the payments notionally made under social security and/or general taxation financing of health care are impossible to apportion in any sensible way between health care itself and the other purposes for which the revenues are used (is the marginal tax dollar a contribution to health care, or unemployment insurance, or national defence, or ...?) and prorating them according to the shares of each expenditure programme in total expenditure is as arbitrary as distributing them in any other way, one is really dealing here with the general equity of the tax/premium system. It makes sense to to look at the "contribution" to overall income equity made by the health care sector only when there is an earmarked tax/premium and/or specific out- of-pocket payments and, moreover, a balanced health care budget.

The rest of the paper will focus on equity in health and health care rather than in payments, not least because equity in health and health care has received much less attention (the case here is not itself an equity case but based on my conjectures about the marginal payoff to analysis here rather than there!).

What is need?

The second interpretation of horizontal and vertical equity in health (or health care) related to "need". Despite its frequent use in an ill-defined way - often barely cloaking special pleading (see Culyer et al. 1971) - the term seems irremovable from public, political and philosophical

discussion and, consequently, on the "if you can't beat 'em, join 'em" principle (but only on my terms!), it becomes necessary - stick though it may in the gullets of many economists so to do - to provide the word with suitable content. For consistency with the "output" orientation described above, I shall initially draw on the tradition at York which defines need as "ability to benefit" (Culyer 1976, 1978, Williams 1974, 1978). More health (somehow defined) is taken as a general ethical desideratum in much the same way, though less comprehensively, as more "welfare" is a commonly accepted ethical criterion in normative economics, and need is a kind of social (as distinct from private) demand for health, from which there may be a derived demand (or need) for health care. The question of who the arbiters of the social "demand" ought to be, and the extent to which consumer values are embodied in social judgments, is a major political question to which I, as an economist, cannot provide an authoritative answer. The matter is plainly an ideological one, to be determined by suitable political decisions, which may vary according to the tier of decision making one is considering. For example, the answer at the tier of decision making to do with determining the general resource constraints under which micro decisions at the hospital or clinical level are to be made may differ from the answers at these more micro levels. Distinctions between different individuals' or groups' needs thus become couched in terms of their ability to benefit from the consumption of health care, with benefit being measured in terms of expected health outcomes relative to what otherwise would have been the case. (The temptation to see benefit as a before-and-after comparison should be avoided).

Need, in the outcome approach, is both a supply and a demand notion. It is a supply notion in that it embodies the productivity of health care

resources (their potential for improving health) and a demand notion in that health (as will be seen) embodies value judgments about what it is that characterises "health", whose ultimate source, most will probably agree, ought whenever feasible to be actual or prospective patients, and about the relative value to be attached to a "unit" of the health of different individuals or groups, whose source must be some over arching social value judgment. It is, in fact, analogous at the level of individual need to the microeconomic concept of a general equilibrium demand curve, in which the demand function is derived from an indifference map subject to the constraint, not of the individual's income, but of the economy's production possibilities curve, and with social decision makers' preferences replacing consumers'.

Later, the input approach to need will be considered, as this is plainly relevant in the context of the present volume, where the empirical work has been largely conceived (so far as I can tell) in this way.

What is health?

For the sake of convenience in what is to follow, I shall take health to be measured (on a ratio scale) by some appropriate empirical measure such as the Quality Adjusted Life Year (QALY) (Torrance 1986, Williams 1985) or the Healthy Years Equivalent (HYE) (Mehrez and Gafni 1989) without taking sides as to which approach, or which experimental method for deriving the empirical measure, is most appropriate (and skating over the difficulty that a common characteristic of these measures is that they use interval rather than ratio scales). It is sufficient for present purposes to suppose that there exists some acceptable measure, embodying acceptable and consistent

value assumptions, which can be used to give substantive content to the word "health". (For further discussion of both "health" and "need" see Culyer 1990b).

The notion of "health" clearly involves value judgments (see Culyer 1978) relating, amongst other things, to the relevant characteristics in terms of which health is to be reckoned (such as ability to perform activities of daily living, freedom from pain) and the tradeoffs between these characteristics. Judgments about these elements have to be made at a variety of different tiers of decision making (for example, the amount of public expenditure to be devoted to health care, size and departmental composition of a hospital, clinical decisions about individual patients), each of which embodies an "agency relationship" between the decision maker or supplier of service and the ultimate beneficiary (the "patient")[3]. It is certainly not the case that the idea of health-as-QALYs necessarily involves the wholesale imposition of external values on individuals, or "paternalism" (though those who wish to impose, reject consumers' values, or be paternalist may certainly do all these three things via QALYs!). However, the necessity for establishing an interpersonal tradeoff between the health of different individuals or groups is inescapable, whether one uses willingnesses to pay or some other basis for weighting individual benefits. This can be done only by making (preferably explicit) distributional value judgments[4].

Health may be seen as both a stock and a flow: the stock being the sum of expected QALYs for an individual or group (I shall henceforth use QALYs as a suitable shorthand for "health") and the flow being any change in the stock brought about by "depreciation" of, "investment" in, or external "shocks" on, the stock (the language of capital and investment seems

compelling).

Equity and equality

With the foregoing as a preparation, it is now possible to compare the ideas of horizontal and vertical equity in health and to relate them in a context that enables them to be considered at the same time as efficiency in health production. Before doing so, however, it is necessary to be explicit about how interpersonal comparisons of health are to be made. I shall assert, as an egalitarian principle which I term "QALY egalitarianism", that a QALY is of equal social value to whomsoever it accrues. QALY egalitarianism is asserted in order to test its consistency with other equity objectives rather than because it is particularly compelling on ethical grounds (it is easy to imagine individuals whose QALYs ought possibly to command a relatively high weight, such as mothers with several dependent children; and it is not particularly compelling to treat as equally socially valuable 10 QALYs received by one person or one QALY received by each of 10 different, but equally meritorious, individuals). Nonetheless, QALY egalitarianism is the value judgment with which we shall work and, as will be seen, the method of analysis enables alternative assumptions to be inserted at will.

The three kinds of equality or inequality of treatment to be considered are:

Horizontal:

H1 Equal treatment of those with equal initial health

H2 Equal treatment for equal need

H3 Equal treatment of those with with equal expected final health

Vertical:

V1 More favourable treatment of those with worse initial health

V2 More favourable treatment of those with greater need

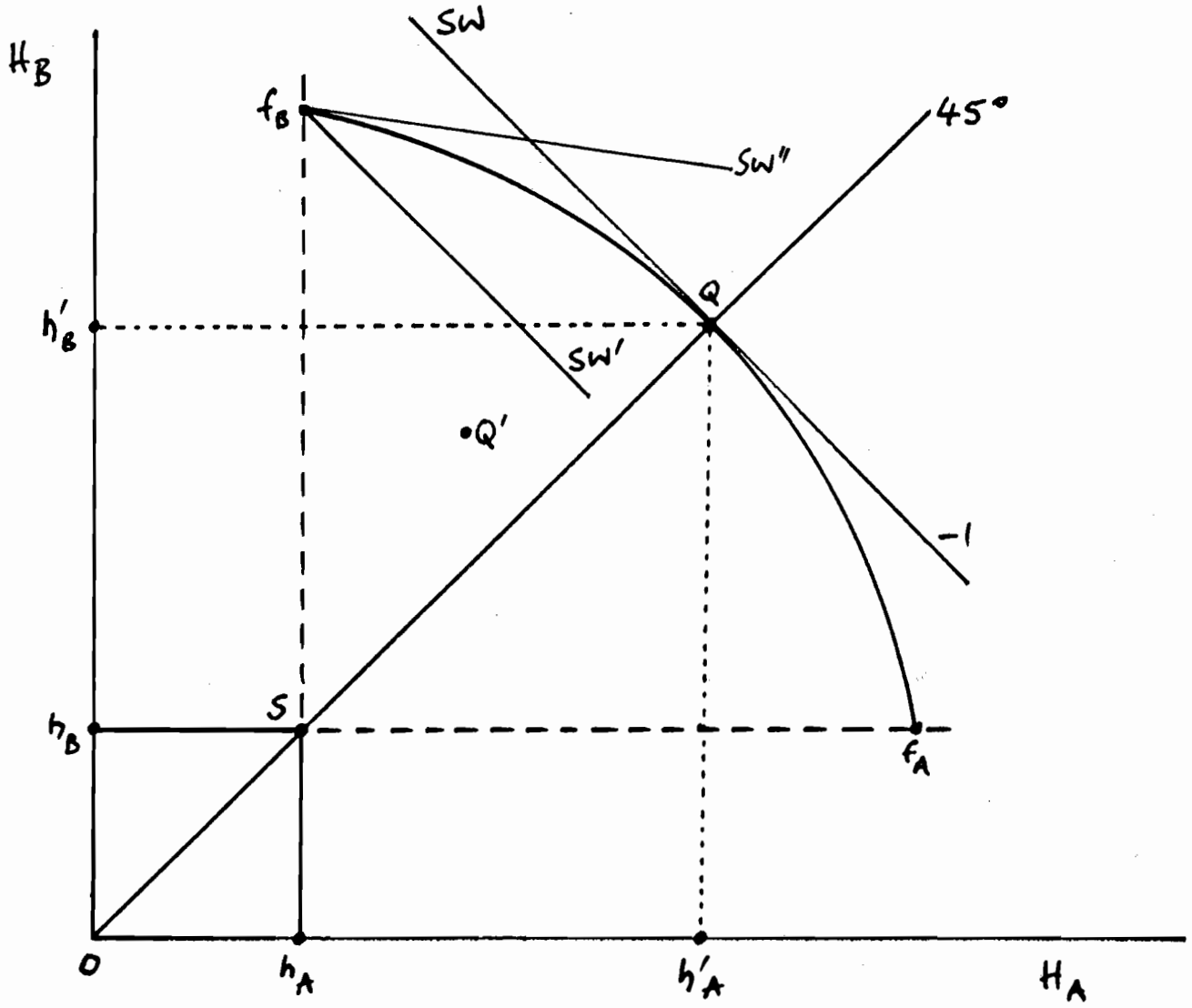
V3 More favourable treatment of those with a worse expected final health

It will be noted that vertical equity principles adopted are rather weaker than Aristotle's (who recommended proportionality). The detailed consideration of these principles is, moreover, a matter for another paper, and here we shall concentrate on horizontal equity.

Equal treatment as equal additional health

Figure 1 shows a special situation in which, in a world of only two individuals of equal age[5], there is equal initial health, equal need, and an equal expected final health outcome. The figure is based on one developed in Wagstaff 1990. The axes of the figure measure the stock of prospective health, H, in QALYs for each individual A and B. The 45° line is a locus of equality of health for A and B. Point S represents the initial stock of health of the two individuals ($h_B=h_A$). The dashed lines within the H-space represent additional H. Northerly movements from S represent increasing H for B, and easterly movements represent increasing H for A. Northeasterly

FIGURE 1



movements represent increases in the stock of H for both individuals. The convex (from above) locus f_Bf_A represents the technically possible increases in H that are possible, given an overall amount of resources (inputs) available to this two-person economy. These increases are to be imagined as computed in a fashion developed by Williams (eg Williams 1981). Its slope implies that increasing H for one individual can be obtained only by increasing sacrifices of additional H for the other. This may arise either under diminishing returns in the health care production functions for A and B, or under constant returns in each "technology" but with different input intensities in each. f_Bf_A is the "health frontier". It corresponds to, and can be derived from, the contract curve in an Edgeworth production box which is the locus of tangency points for isoQALY contours in the production functions for A's and B's health (see below). Given resource constraints, the maximum amount of additional H for B that is possible is Sf_B and the maximum amount for A is Sf_A. I take these limits as defining the "abilities to benefit" of A and B. Any productively efficient combination must lie on f_Bf_A. Any point on the frontier represents a cost-effective distribution of H between A and B in the sense that additional H for one can be procured only at the (opportunity) cost of less H for the other. All points below f_Bf_A indicate productive inefficiency and all points above it are unattainable. The f_Bf_A frontier is symmetrical about the 45° line, indicating that the capacity of each to benefit from health care is the same in QALYs (their needs are the same). The SW line with slope -1 is a social welfare contour embodying QALY egalitarianism and indicating that increments of H to either A or B are equally valued socially, independently of the amount of H that each has initially and that each receives additionally: any movement along a given SW contour represents a socially equal value of QALYs, though of course the distribution between A and B varies (if H were deemed to have a

diminishing marginal social value as each individual had more, the SW contours would take on a shape that would be concave from above).

The figure captures the elements of horizontal equity previously discussed. Equal initial health is indicated by point S ($Oh_B = Oh_A$). If health care is efficient in its impact on H, the horizontal equity requirement H1 of equal treatment for equal initial health (in terms of outcome) requires equal receipt of additional health. The only point in the figure satisfying both this equity criterion and productive efficiency is point Q, at which A receives h_Ah_A' equal to B's receipt of h_Bh_B'.

Equal expected final health is also shown by point Q, since it lies on the 45° line through the origin and is on the health frontier. Again, therefore, equal treatment in the form of equal additional health outcomes requires location at Q.

The <u>needs</u> of A and B are represented in Figure 1 by Sf_A and Sf_B, corresponding to their respective abilities to benefit. These abilities depend crucially on the productivity of health care (the shape of the frontier). In Figure 1, A and B have equal needs ($Sf_B = Sf_A$) and each ought therefore to receive equal amounts of additional QALYs according to the horizontal equity principle of equal treatment for equal need (output approach). This requires locating on the 45° line through S and, if productive efficiency is also to be realised, locating also on the health frontier. The point of intersection of the 45° line through S and the frontier is Q, indicating that this point is again that which satisfies the second horizontal equity requirement (H2).

Efficiency in meeting needs is interpreted here as selecting those needs that are to be met, or determining what may be termed "entitlements to health", by prioritizing the more "urgent" and so distributing health between A and B that, at the margin, the cost of A's and B's additional health is equal to the social value attached to the health of each. The relative social value (in QALYs) is given by QALY egalitarianism, expressed in the Figure by the social welfare contours having constant slopes of -1. Full allocative efficiency requires attainment of the highest SW contour, which occurs at Q, where the frontier is tangential to an SW contour. Efficient meeting of needs is thus consistent with the existence of some unmet need (cf. Wiggins and Dirmen 1987): the optimal unmet need is n_Bf_B and n_Af_A. The selection of any other point on f_Bf_A involves a lower social value of met need than that attainable at Q. For example, the selection of point f_B, at which B receives all the additional health, is efficient in cost-effective terms (A can have more only if B has less: it is still a point on the frontier) but it is not fully efficient. At f_B the social welfare contour SW' (also with slope -1) lies below the contour at Q and so does not maximise the social value of additional QALYs. This allocation of additional health between A and B would maximise the social value of additional health only if QALY egalitarianism were sufficiently relaxed, specifically requiring B's QALYs to receive a sufficiently higher social weight than A's to produce a tangency (or corner solution) on f_Bf_A with the much flatter contour SW".

Figure 1 represents, or course, a specially constructed case in which there is no conflict between any of the various horizontal equity requirements, nor with any of them and efficiency. Details of the conflicts that do emerge when individuals are equal in some, but not all, respects, are explored elsehwere (Culyer 1990a).

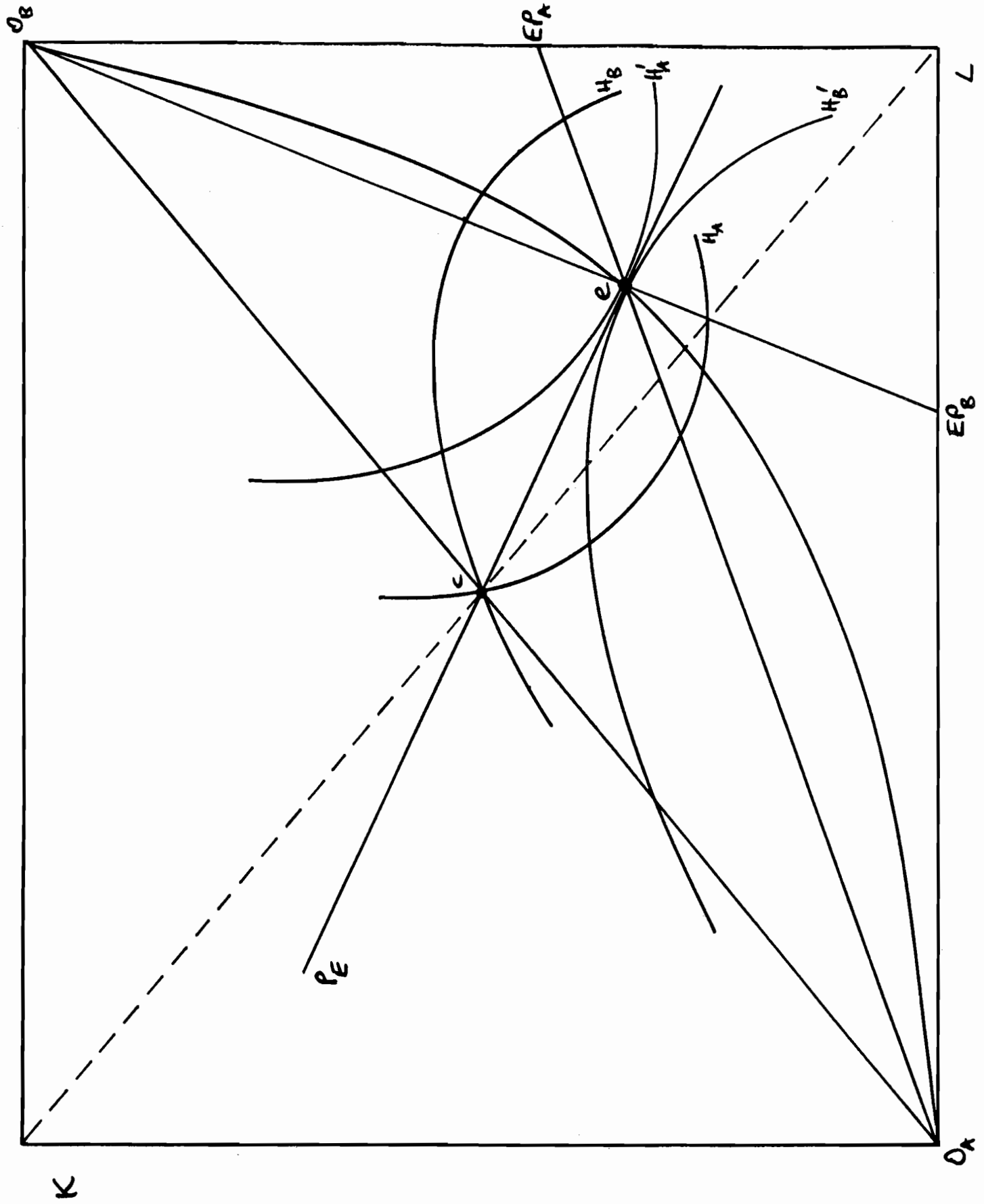
Equal treatment as equal expenditure

The alternative way of considering "treatment" is to regard it in input rather than output terms. The use of expenditures on health care is common in the literature (eg the pioneering work of Le Grand 1978) and is the approach selected by the author in the present volume. Under this approach, the horizontal eguity principles become "equal receipt of health care resources in value terms for equal initial health, need, or final health" and the question arises as to whether this introduces conflicts between equity and efficiency and whether the use of this approach implies different equitable distributions from those implied under the "output" approach.

This version of the problem can be explicated with the help of Figure 2. This is an Edgeworth Box in which the axes measure the available resources (two inputs, the services of capital and labour). The two production technologies for A's and B's health utilising these two inputs are represented by the two expansion paths $O_A A$ and $O_B B$, which trace out the the optimal input combinations at the equilibrium input price ration P_E as each activity expands, assuming constant returns in each but differing factor intensities (A's health-improving technology is assumed to be relatively labour-intensive). Output is measured in QALYs and each isoQALY contour is analogous to an isoquant. The contract curve $O_A O_B$ is the usual locus of tangencies of the isoQALY curves. It is from this contract curve that the health frontier in Figure 1 is derived, and productive efficiency requires the economy to be located on this locus.

Equal treatment in the sense of health care expenditures requires each technology (viz. that for A's health and that for B's) to share a common

FIGURE 2



isocost line. Such an isocost line must pass through the centre of the Box, c, since the cost of A's and B'a health care must be equal when they use identical inputs and the the input prices are the same for each technology. The relevant isocost line must also, for efficiency, be tangential to two isoQALY curves on the contract curve. One may conceive of some Walrasian process in which the relevant input prices are established to ensure optimality at the efficiency point e. All points along ce therefore indicate equal expenditure on A and B. It is immediately clear that equal expenditure on inputs is consistent with productive efficiency, so long as input prices can adjust appropriately, and that equal expenditure is thus consistent with being on the health frontier in Figure 1.[6]

There is in general, however, no reason to expect that this equal allocation of expenditure between the two individuals corresponds to an equal share in additional health. The two isoQALY curves through c (H_A and H_B) must represent equal additional QALYs for each on the assumption of constant returns and given equal needs. In Figure 2, equal needs (capacities to benefit) imply that B's isoQALY curve at O_A and A's at O_B have the same value. Since c represents half of the input flow at the origins, it must also represent half the output of QALYs, given constant returns. However, this point cannot be an equilibrium if factor intensities differ. The output of QALYs is higher for both individuals at e than at c but is even higher for B than A ($H_A' < H_B'$). In general, additional health will not be the same at e, even though expenditure is the same (they are on the same isocost line ce).

There is thus an inconsistency between the two versions of "equal treatment", which makes the distinction between them of substantive importance, and requires an explicit choice, presumably by appeal to some

underlying theory of horizontal equity. Equal treatment in the sense of equal expenditure is consistent with productive efficiency and may be equitable when needs are the same, but equal treatment in the sense of equal receipt of additional health, which may also be equitable when needs are the same, requires selection of a different point on the contract curve, where the two isoQALY curves have an equal value. This will also involve a different configuration of inputs in the health care system and different input prices. Since, at e, A receives less additional health (given the production functions as drawn in Figure 2), equal treatment in the sense of equal additional health requires the selection of a point on the contract curve northeast of e and a relatively higher equilibrium price of input L.

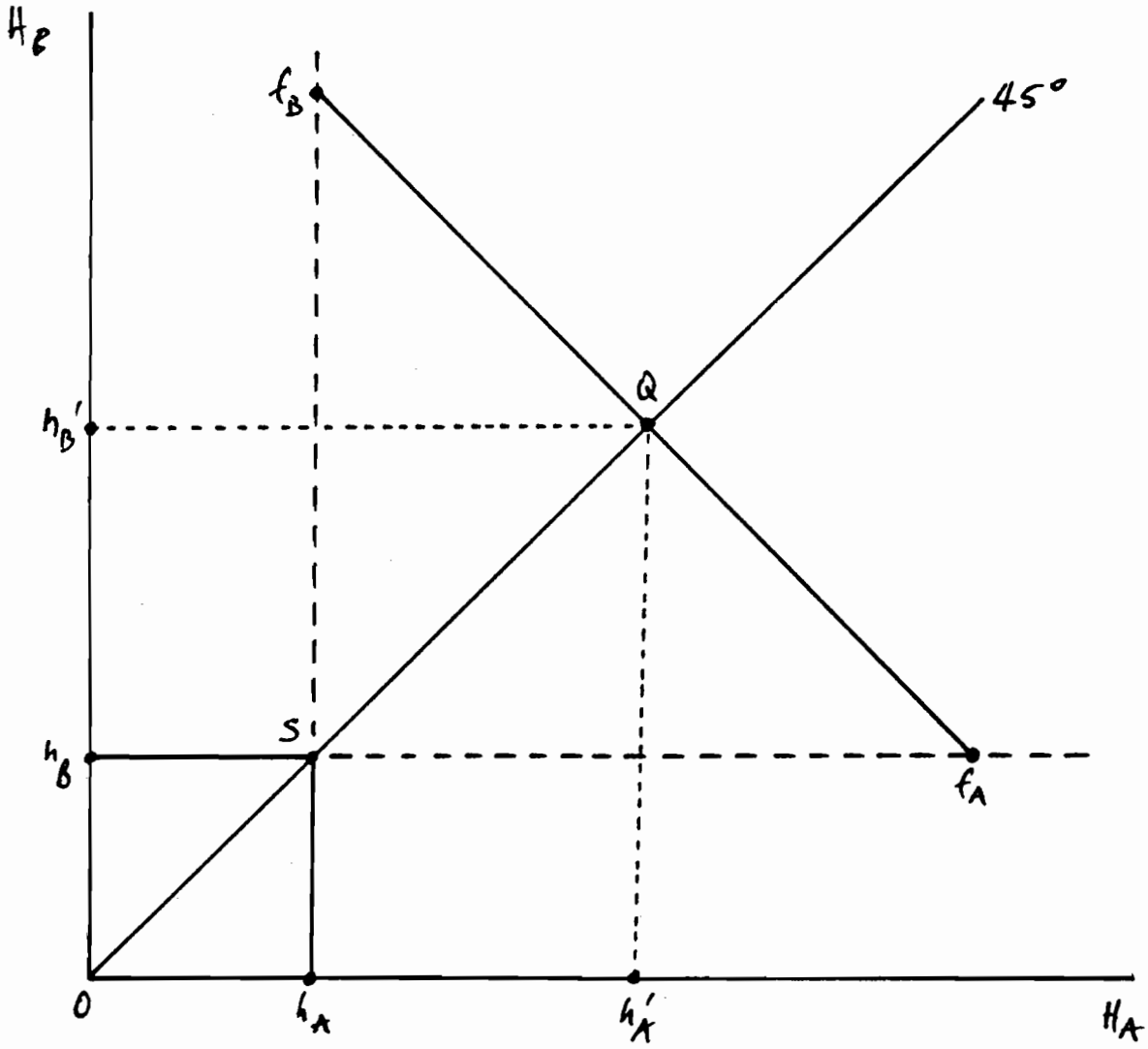
The equity requirement of equal treatment in the sense of equal expenditure for equal need is also in general inconsistent with full allocative efficiency (under QALY egalitarianism) since this requires equal additional health for each. Relaxing QALY egalitarianism might produce a tangency in Figure 1 that corresponds to the QALY distribution implied by equal expenditure, but since the factors determining the social tradeoffs between A's and B's health are not the same as those determining their shares under equal expenditure, such an outcome can only be coincidental.

Thus, equal treatment in the sense of equal expenditure on health care implies different health outcomes and conflicts with both allocative efficiency and final health equality. Equal treatment in the sense of equal additional health is consistent with allocative efficiency and also with final health equality, but will involve different expenditures for equal needs. Both forms of "treatment" are consistent with productive efficiency if input prices are free to adjust optimally.

The inconsistency between the two versions of "equal treatment" arises out of the different productivities of health care for A and B (even under constant returns). Retaining this assumption, so that equal proportionate increases in inputs produce the same proportionate increase in QALYs within each activity, but assuming production homogeneity in the sense of identical input-intensities in each activity[7], produces a special case where the conflict vanishes. Under these conditions, the contract curve coincides with the diagonal in Figure 2 and a given increase in QALYs for one is exactly compensated by the QALY loss for the other as one moves along the diagonal. The health frontier becomes linear as shown in Figure 3 and, for the case of equal need (equal capacity to benefit), is symmetrical about the 45° line through the origin. Given QALY egalitarianism, society is indifferent between points on the new health frontier[8]. Under these conditions, the equitable point Q is not only equitable under all the horizontal equity principles (provided that A and B are equal in all relevant respects) but the system is efficient in both productive and allocative senses, and expenditure on A and B is equal[9].

In this special case[10], provided that one can be confident about the production homogeneity of the As and Bs taken as a whole, equal expenditure for equal health, equal need, or equal final health, are all consistent and, moreover, consistent with QALY egalitarianism and productive and allocative efficiency. A relevant inequality of persons justifying a possibly equitable unequal treatment will, of course, arise where the homogeneity assumption does not hold. But, in horizontal equity, like expenditure for like cases becomes a principle consistent with the outcome oriented approach.

FIGURE 3



Trading health off against other goods

The entire discussion so far has treated equity and efficiency in health care and health within a given resource availability to the health sector. It will be apparent that the optimal size of the health care sector cannot be considered independently of, first, the productivity of other elements amongst those determining the health of individuals and groups relative to that of health care itself and, second, the value of health relative to the value of other produced goods and services. Although I have treated health as the efficiency maximand in this paper, and as the central matter of distributional concern, it is not synonymous with "welfare" (notwithstanding a classic WHO early definition) and health is not the only moral pursuit or candidate for specific egalitarianism. One therefore needs also to consider the optimal size of the health care sector in terms of the valuation of health relative to the other good things of life and to evaluate the overall equity of a particular society by taking account of the equity of distributions of entities other than health care and health. It may be that relatively inequitable distributions of some entities may be compensated in such overall judgments by more equitable distributions of other entities. The question of the "separability" and "additivity" of degrees of inequality in specific entities of concern is, however, a difficult question - unresearched so far as I know - not further considered here.

Summary

I have tried to provide some reasons why it may be interesting to examine the distribution of health, health care, and health care payments, the most important of which hinge on the notion of the existence of

"ultimate" entities that are of specific distributional concern which implies that other, less ultimate entities, are of less (or even no) distributional concern. I have argued that the ultimate entity of concern here is "health", for which health care and health care expenditures are only instrumental. It is argued that the nature of health implies that it is an ultimate entity whose distribution across the whole range is of concern (unlike some others, where equity may be satisfied by the meeting of some minimum standard). A concept of need for health care was developed, analogous to the general equilibrium individual demand curve in that the constraint is taken as the economy's production possibilities but differing from that construct in that social values replace (at least in part) consumer values. Horizontal equity in health was considered in terms of three alternative principles, whose consistency with one another and with the efficient production of health was explored. It was shown that equality of treatment in the sense of either improved health (outcome view) or amount of health care expenditure (input view) were mutually consistent given appropriate patient classifications and also consistent with economic efficiency. A diagrammatic technique was introduced that enables the simultaneous consideration of these various equity principles, efficiency, and the interpersonal comparison of health. This technique can also be used to explore the consistencies and inconsistencies of vertical equity principles with one another and with efficiency, and to explore the difficulties that arise when individuals are like in some relevant respects but unlike in other relevant respects, though these issues are not pursued here. The comparisons made do not depend upon there being an explicit tradeoff between health and other arguments of the social welfare function, though the optimal size of the health care sector will depend on such tradeoffs, as well as the relative productivity of the health sector and other environmental influences in the promotion of health.

Notes

1. Others might include protection from the elements, opportunities to enter careers suited to talents, having one's legal rights protected - to which the corresponding less ultimate entities might be housing, schooling, and access to the courts.

2. Assuming that there is likely to exist some means by which the actual distribution, if found inequitable, can be altered for the better, or compensated by relatively favourable distributions of other entities, and that the means adopted for making changes in distributions satisfy any criteria relating to procedural equity.

3. For further discussion of these tiers see Culyer 1990b.

4. It is worth emphasising that the value judgments do not have to be those of the analyst. So far as possible, scholarly analysis should be used to explore the consequences of making value judgments rather than itself be intrinsically ideological.

5. It is assumed that A and B are of the same age but not that they have an equal expectation of life. Age standardisation is required in order to avoid arbitrary (and inequitable) bias in favour of the young who, ceteris paribus, have a greater life expectancy at all ages in developed coutries. Age is therefore another respect in which individuals are to be reckoned equal or unequal for equity purposes. In efficiency analysis, this assumption is not required, and if it is not made an inconsistency may arise between the demands of efficiency

and those of equity. In this paper allocative efficiency is considered only with respect to persons of like age.

6. If input prices cannot adjust optimally there will be an excess demand for one input and an excess supply of the other, producing an interior solution in Figure 1. Note also that, although expenditure includes all input costs, it does not include "fixed" costs (if any).

7. This may be more likely to apply when A and B (or the groups of patients for whom A and B are representative individuals) belong to the same diagnostic group - suggesting another "respect" in which to consider equality or inequality. However, this standardisation may not always, or even usually, imply homogeneity on the production side; for example, duodenal ulcers may be treated either medically or surgically, with different input combinations and different costs per case.

8. For any other constant weighting of QALYs, the social optimum is a corner solution.

9. I assume no interregional variation in input prices.

10. "Special" in the sense of requiring particular empirical conditions to apply for its valid application. Whether in practice they do commonly hold is another (empirical) matter.

References

BUCHANAN, J M (1968) What kind of redistribution do we want? Economica, 35, 185-190.

CULYER, A J (1971) The nature of the commodity 'health care' and its efficient allocation, Oxford Economic Papers, 23, 189-211.

CULYER, A J (1976) Need and the National Health Service: Economics and Social Choice, London, Martin Robertson.

CULYER, A J (1978) Need, values and health status measurement, in A J CULYER and K G WRIGHT (eds.) Economic Aspects of Health Services, London, Martin Robertson, 9-31.

CULYER, A J (1990a) Health, equity and efficiency: a simple diagrammatic approach, Osaka Economic Papers, forthcoming.

CULYER, A J (1990b) Ethics and efficiency in health care: some plain economic truths, 1990 Perey Lecture, McMaster University, Canada.

CULYER, A J, LAVERS, R J, and WILLIAMS, A H (1971) Social indicators: health, Social Trends, 2, 31-42.

FRIEDMAN, M (1957) A Theory of the Consumption Function, Princeton, Princeton University Press.

LE GRAND, J (1978) The distribution of of public expenditure: the case of health care, Economica, 45, 125-142

MCKEOWN, T (1976) The Modern Rise of Population, London, Arnold.

MEHREZ, A and GAFNI, A (1989) Quality-adjusted life years, utility theory, and healthy years equivalents, Medical Decision Making, 9, 142-149.

PAULY, M V (1971) Medical Care at Public Expense, New York, Praeger.

SIMONS, H C (1938) Personal Income Taxation, Chicago, University of Chicago Press.

TOBIN, J (1970) On limiting the domain of inequality, Journal of Law and Economics, 13, 263-278.

TORRANCE, G W (1986) Measurement of health state utilities for economic appraisal: a review, Journal of Health Economics, 5, 1- 30.

WAGSTAFF, A (1990) QALYs and the equity-efficiency trade-off, Journal of Health Economics, forthcoming.

WIGGINS, D and DIRMEN, S (1987) Needs, need, needing, Journal of Medical Ethics, 13, 63-68.

WILLIAMS, A H (1974) 'Need' as a demand concept (with special reference to health) in A J CULYER (ed.) Economic Policies and Social Goals: Aspects of Public Choice, London, Martin Robertson.

WILLIAMS, A H (1978) 'Need' - an economic exegesis, in A J CULYER and K G WRIGHT (eds.) Economic Aspects of Health Services, London, Martin Robertson.

WILLIAMS, A H (1981) Welfare economics and health status measurement, in van der GAAG, J and PERLMAN, M (eds) Health, Economics, and Health Economics, Amsterdam, North-Holland.

WILLIAMS, A H (1985) Economics of coronary artery bypass grafting, British Medical Journal, 291, 326-329.