Efficiency, equity and equality in health and health care

Anthony J Culyer

Emeritus Professor, Centre for Health Economics, University of York, UK
Adjunct Professor, University of Toronto, Canada

December 2015
Background to series
CHE Discussion Papers (DPs) began publication in 1983 as a means of making current research material more widely available to health economists and other potential users. So as to speed up the dissemination process, papers were originally published by CHE and distributed by post to a worldwide readership.

The CHE Research Paper series takes over that function and provides access to current research output via web-based publication, although hard copy will continue to be available (but subject to charge).

Acknowledgements
My thanks to Ruth Faden and Carleigh Krubiner for valuable comments on a draft.

Further copies
Copies of this paper are freely available to download from the CHE website www.york.ac.uk/che/publications/ Access to downloaded material is provided on the understanding that it is intended for personal use. Copies of downloaded papers may be distributed to third-parties subject to the proviso that the CHE publication source is properly acknowledged and that such distribution is not subject to any payment.

Printed copies are available on request at a charge of £5.00 per copy. Please contact the CHE Publications Office, email che-pub@york.ac.uk, telephone 01904 321405 for further details.

Centre for Health Economics
Alcuin College
University of York
York, UK
www.york.ac.uk/che

© Anthony J Culyer
Abstract

Three common “Es” have high ethical and political content for health policy: efficiency, equity and equality. This article examines the links between the three, with especial attention given to (a) the claimed conflict between efficiency and equity, (b) the equity of inequalities and (c) the conflict between six equity principles: equal health, equal health gain, equal value of additional health, maintaining existing distributions, allocation according to need and equal per capita resources.

Conclusions include:

- Efficiency and equity do not inherently conflict
- an inefficient allocation can be equitable
- an efficient allocation can be inequitable
- an inefficient allocation can become more efficient without increasing inequity
- what is equitable often requires inequality in health and inequality in resource distribution per capita
- equality in health requires inequality in resource allocation
- equality in resource allocation typically leads to inequality in health
- allocation according to need typically leads to inequality in health
Three common ‘Es’ have high ethical and political content for health policy: efficiency, equity and equality. Having an efficient health care system is morally important because health care is one important element that determines people’s health – and good health is a central part of both individual and societal wellbeing, good both for its own sake and because it enables individuals and communities to achieve other good things that contribute to wellbeing. Health is not synonymous with wellbeing, despite a famous World Health Organisation axiom to the contrary\(^1\), but it is usually a necessary condition for it. This is particularly true in poor counties – and the many poor communities that exist in rich and middle-income countries.

There can be no doubt that a principal objective of health care systems is to maximise health. In the UK there has been ministerial authority for that, for example: “The purpose of the NHS [National Health Service] is to secure through the resources available the greatest possible improvement in the physical and mental health of the people of England” (Department of Health 1996). Although it is never expressed quite like this, the underlying axiom is nearly always that ‘other things equal, more health is always to be preferred’. The more interesting, nonfactual, ethical assertion is that efficient systems of health care ought to be preferred.

The ethical underpinnings for maximising population health (or at any rate for increasing it) lie in the importance of good health for people to lead flourishing lives, which I take as an ultimate good. We can all think of individuals with terrible handicaps of ill health who seem to flourish but these are not persuasive counter examples. Such people excite our admiration and are seen as exceptional. In general, I take it that flourishing is an ultimate good and that good health is normally a necessary condition for it. In short, health is needed in the twin senses that it is both necessary and that it also serves an ethically commendable end. This gives an otherwise merely technical relationship between means (health care) and ends (health) its ethically persuasive quality and raises the need for health to high ethical significance (in a way that is not true for my need for a Rolls Royce, even though owning a ‘roller’ would undoubtedly be an effective way of achieving one of my life’s ambitions: impressing my students).

To take the argument further, health care (including medical care) is often a necessary condition for realising better health, though not always sufficient. If so, health care too is needed (that is, is necessary if improved health is to be attained) and it too derives its ethically compelling character from the ethicality of the flourishing that is the ultimate good. So, not only may it be reasonably assumed that individuals want health care; they also need it, with all the moral weight that the word ‘need’ carries in the context of social policy (Culyer 1995, 1998, 2007). If all that is accepted, ‘being efficient’ is a good thing, in the sense that the resources used in health care are to be used to maximise health outcomes.

Distributive justice, or equity, also acquires a high priority and for the same reasons. If health is a necessary condition for human flourishing, health care is too important a resource for it to be accessed and used unfairly. But what, more precisely, is it to be ‘equitable’? And how can we integrate considerations of equity and those of efficiency?

---

\(^1\) “Health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity.” (Preamble to the Constitution of the World Health Organization as adopted by the International Health Conference, New York, 19-22 June, 1946; signed on 22 July 1946 by the representatives of 61 States (Official Records of the World Health Organization, no. 2, p. 100) and entered into force on 7 April 1948.)
Efficiency

Some countries make their commitment to efficiency explicit. Many do not. Nonetheless, I am going to take it for granted that health care efficiency, in the form of using health care resources, like the health care workforce (human knowledge and understanding, human skills), capital goods (buildings, equipment), is a good and ethical thing. Inefficiency, as when more people are used to accomplish something than are necessary, or a more expensive medicine is prescribed when a less expensive one would be no less effective, or more highly skilled staff are used when less highly trained individuals would be at least as good, is a bad and unethical thing. The main reason for taking this view is that the same chain of reasons why health care is instrumental for health and health is instrumental for flourishing apply to being efficient in producing health. Being inefficient implies that population health is less than it could be with the available resources. Given the ethical significance of good health, to tolerate such a state of affairs would require a powerful countervailing argument.

Efficiency provides a useful point of departure, partly, as has just been claimed, because the reasons why efficiency matters ethically are basically the same reasons for why equity also matters ethically. There are many principles of equity that are worthy of exploration and clarification. I shall discuss several candidates. However, I shall not focus here on questions of equitable health care financing, equity in the ownership of health care resources, equity in accessing career opportunities in health care, or equity in policies concerning the non-health care determinants of health, like the equity of the distribution of purchasing power, or good quality housing, or good social care, or clean air, or workplace safety, or political influence. Nor shall I address the many ethical issues relating to processes of health care, for example, issues of accountability (whose and to whom and for what?), transparency of decisions (why were decisions made as they were, by whom, with what evidence in support and what values embodied?) and participation (who may participate in decision making at various levels from the doctor-patient decisions to public policy decisions, who may be consulted, who has a vote or a veto?). Some of these are addressed in Chalkidou et al (2009), Culyer and Lomas (2006), Daniels and Sabin (1998) and Lomas et al. (2005).

\[\text{\textsuperscript{2}}\text{Such arguments undoubtedly exist. For example, if workers in the health system were outrageously (or even modestly) exploited, and better pay and conditions would reduce the amount of health care a given budget could produce, then a reduction in health (even that of the workers in question) might be judged a price worth paying.}\]
Values and equity

It is common to refer to, and applaud, ‘evidence-based’ decision making. In truth, however, few if any decisions can ever be truly evidence-based. Neither can they be truly only ethics-based. Good decisions do indeed need to be informed by evidence and a good appreciation of how variable and contestable evidence can be; and they need also to be informed by values, and good appreciation that values are not always shared, may conflict, and vary in moral worth. Both sets of issues require deliberation, on both moral and political grounds, between those likely to be affected by the decision (commonly referred to as ‘stakeholders’) and prior decisions about the various processes involved.

Health inequalities are of concern in most rich countries. The inequality problem is far more pressing in low- and middle-income countries. In 2012, the under-five mortality rate in low-income countries was 82 deaths per 1000 live births – more than 13 times the average rate in high-income countries. Global life expectancy at birth among men ranged from 75.8 years in high-income countries to a low of 60.2 years in low-income countries – a difference of 15.6 years. For women, a gap of 18.9 years separates the life expectancy figures in high-income countries (82.0 years) and low-income countries (63.1 years). These are averages. Within the low-income countries there are figures far worse than these. These huge inequalities in health appear in all measures and for all age groups and they apply also within countries. There are huge inequalities in the distribution of and access to health care, as well as huge inequalities in health itself. The availability of public and private money for health care is woefully inadequate in the face of the need. All the more important, therefore, that what resources there are for health care are used with maximum effect and to address at least the worst avoidable inequalities.

Yet one of the sad truths about health and health care in low- and middle-income countries is that policy makers are constantly bombarded with claims – many from health economists and public health physicians – for the inclusion in public health plans of interventions whose only virtue is that they are effective. But being effective is not enough. Thousands of interventions are effective but not all – far from all – can be provided. Most of these interventions are simply not effective enough to be included. Nor are they cost-effective. Childhood interventions, particularly vaccinations, often provide greatest value in terms of health gain per unit of expenditure. In sub-Saharan Africa, for example, rotavirus vaccination has been associated with a cost-per-Disability-Adjusted Life-Year averted of $43 (Atherly et al. 2009) and treatment of severe malnutrition costs $53 per DALY averted (Bachmann 2010). Even within HIV, prevention of mother-to-child HIV transmission costs below $150 per DALY-averted using available interventions (Shah et al. 2011), yet a large coverage gap remains across all low- and middle-income countries (UNAIDS 2013, Revill et al. 2015). These highly cost-effective interventions are seriously under-provided, while advocates routinely make recommendations on the basis of absurdly high cost-effectiveness thresholds, aiming to promote access in poor communities to new and more expensive therapies with cost-effectiveness ratios ten times worse than these (Chisholm et al. 2012, Ortegon et al. 2012). Tanzania, whose current list of ‘essential medicines’ contains more than 500 medicines with many controversial drugs on it such as bezacizumab in addition to paclitaxel and carboplatin for treating ovarian cancer and ranibizumab for treating macular eye disease. The first of these is regarded as not cost-effective by the National Institute for Health and Clinical Excellence (NICE) in the context of England and Wales and the second is recommended only if the manufacturer offers substantial discounts. These, in other words, are ‘bad buys’ in one of the world’s richest countries. They stand, nonetheless, on Tanzania’s list of cost-effective essential medicines. These depressing facts reveal a carelessness of attitude to efficiency and equity in countries where they are of the utmost importance.
At this point the reader might throw hands in the air and declare that all is too complicated, too awry, too technical and that it is altogether too difficult to see clearly what should be done. Whereas answering the question ‘what should be done?’ always requires its consideration in the local context, with local history, local values, locally available budgets for health care and other distinguishing marks of local identity, being clear at the level of principle and making clear distinctions is less subject to such local variation. I shall proceed therefore to build as simple a model as possible in order to explore the relationships between efficiency, equity and equality, exploding some myths as we go along and identifying helpful ways of thinking. I shall use a simple model of a health care system to illustrate what I mean by both efficiency and equity.
A simple model of efficiency and equity

We begin with the idea of a health production function that links aggregate health care expenditures to some index of population health. The production function show the capacity of a group (or an average individual) to benefit from changes in some specified determinant of health, here health care as represented by health care expenditures. It is a measure of the effectiveness of health care expenditures in terms of their impact on the health of a population at various rates of expenditure. It is also a measure of the maximum capacity to benefit (in terms of health) of that population at a variety of rates of expenditure. Such a function is depicted in Figure 1. The curve from point E shows how health rises as health care expenditures rise, enabling additional treatments to be deployed and for existing interventions to be scaled up across the population – but at a diminishing incremental rate. Additions to expenditure add to health but at a falling rate per dollar as, for example, less productive interventions are used and harder-to-reach groups are covered. Point E is the endowment, or starting point, of current health: the flipside of the burden of disease. It is determined by historical and recent factors, and measured by an appropriate generic state-of-health indicator such as the quality-adjusted life-year (QALY) or averted disability-adjusted life-year (DALY). The peak of the curve represents the highest of all the maximum abilities to benefit from additional expenditures on health care, beyond which lies a range of iatrogenesis, as further health care expenditures actually reduce average health. Needless to say, we are for practical purposes concerned only with the rising section of the health production function. The curve locates the maximum positive impact on health, or negative impact on the burden of disease, that any given rate of additional expenditure can have, given the available technologies. The production function therefore embodies an assumption of cost-effectiveness: more resources than are necessary for achieving an outcome are not used. At point e”, for example, the system is generating only 50% of its potential health gain and it cannot therefore be a point on the production function. It is inefficient.

![Figure 1 The health production function](image-url)
Let us suppose that such curves exist, at least in principle, for all sections of the community, and that they typically vary from disease to disease, from social and economic context to context, and from one average type of patient to another. We might consider production functions for people with chronic orphan diseases for whom treatments are not very effective and are usually very costly relative to others. Or we might consider the rural poor in a low income country, for whom the most effective treatments are very effective but quite costly to deliver compared with the well-to-do urban population in that country, or a group of multiply deprived people compared to the rest of the population, or just simply the ‘poor’ and the ‘rich’ or the ‘sick’ and the ‘well’. This is, of course, a stylisation of reality. Its purpose is, through simplifying and removing inessential details, to enable us to clarify the kinds of impacts that different allocations of expenditure may have and to compare generic concepts of equity in health and health care.

In Figure 2, the health of two such different groups is measured on the two axes. The point E again indicates current health, in this case, its distribution between the two. The deprived group (B) has on average the health level $H_B$ and the relatively healthier group A has on average the level of health $H_A$. The breaks (/) indicate equal sized gaps in the scales of health and expenditure just to keep the figure on the page. We now suppose that there is additional resource available (say, the budget for public health care rises, or premium income rises).

![Figure 2 The health frontier](image)
This means that the health of both groups can be increased. Given the production function for the As, if all the extra budget were spent on them, the maximum increase in health possible is the distance $E_f^A$. Were it all to be spent on the Bs, the maximum increase is (we assume) much less, at $E_f^B$. Such a differential pattern might be consistent with the As being both relatively deprived of health to begin with and not very responsive to health care (the orphan disease case). Alternatively, the increased funding could be shared between the two groups. The health frontier $f_{A,B}$ shows all the possible combinations of health for the As and the Bs, given a specific increase in funding and the two production functions. A formal derivation of the frontier from the production functions and the budget constraint is in Culyer and Wagstaff (1993).

The health frontier necessarily has the convex shape of $f_{A,B}$ given the assumed shape of the production functions (i.e. they display diminishing marginal returns as expenditure rises). An important characteristic of this curve is that from any point on it, the average health of the As can be increased only at the expense of less health for the Bs (and vice versa)\(^3\). This feature is termed Pareto efficiency. A Pareto-efficient allocation of resources exists when it is not possible to increase one person’s health without causing a reduction in someone else’s. It embraces cost-effectiveness as embodied in the production functions, so there is either no waste in the health care system or no change in its cost-effectiveness, waste included. Any point under the health frontier is by definition inefficient.

If the budget for health care were to rise, the frontier would shift outwards. Conversely, should the budget be reduced, it will shrink towards the origin.

Although every point on the frontier is efficient, every point is not equitable. To select the equitable point, additional criteria are needed. We can now explore some such criteria.

\(^3\) In all cases considered here, we abstract from uncertainty. It may be a more comfortable assumption for some readers to assume that what is measured along each axis is ‘expected health’. 
What is equity?

The treatment of individuals is invariably judged to be inequitable if it is capricious or relates to irrelevant characteristics. This is common to all notions of equity. Commonly cited irrelevant characteristics of this sort include race, religion, gender, ethnicity and gender orientation. These may sometimes become relevant – for example, dietary restrictions on medical grounds or grounds of religion may be regarded as legitimate grounds for patients to be treated differently. These are exceptions, however, to the general rule. What is less frequently perceived is that equality and equity, and inequality and inequity, are not the same, although they are intimately related. At the most general level, equity in health care is held to require that patients who are alike in relevant respects be treated in like fashion and that patients who are unlike in relevant respects be treated in appropriately unlike fashion. In short, there can be equitable inequalities. These requirements are usually referred to as horizontal and vertical equity and derive from Aristotle (1972).

Horizontal equity requires the like treatment of like individuals and vertical equity requires the unlike treatment of unlike individuals in proportion to the differences between them. This is a commonly accepted formal principle of justice. Suppose that need is selected as the only relevant factor that should determine receipt of health care, then the two principles would imply that like needs should receive like attention and resources (horizontal equity) and that greater needs should receive greater attention and resources (vertical equity). Horizontal and vertical equity considerations apply also to entities other than health. One closely related entity is the financial contribution to health care. A horizontal principle here might be, ‘equal contributions from households having an equal ability to pay’ and the corresponding vertical principle is ‘lower/higher contributions from households with a lower/higher ability to pay’. An important implication of these distinctions is that there can be fair inequalities. A fair inequality is one that meets the test of vertical equity. A classic Marxist political slogan combines two vertical criteria: from each according to their ability; to each according to their need (Marx 1875).

Compare three frequently met views of equity in health: equal health, equal health gain and equal value of additional health. These are shown in Figure 3. The 45° line passing through the origin 0 indicates equal average health for each group. Any point on this line would be judged equitable on the criterion of equal health for each individual. Where this line intersects the frontier (at the point marked a) we have an allocation that is both efficient and equitable. This allocation requires the lion’s share of the new money to go to the Bs – the relatively deprived group. The 45° line passing through E indicates equal health gain on average. This intersects the frontier at b, an allocation that is more to the advantage of the As. It is efficient, lying as it does on the frontier, and it is also equitable (according to the criterion of equal health gain). The allocation at point c meets the criterion of equal marginal value. In cost-effectiveness analysis, where health is measured in quality-adjusted life-years (QALYs) this is popularly termed the QALY=QALY=QALY rule, meaning that a QALY is judged to be of equal social value to whomever it accrues. The point on the frontier at which this condition is met is where its slope is equal to -1, i.e. where a small gain to one (in either direction) is equal to the loss of the other. Thus point c indicates an allocation that is again both efficient and equitable (but this time equitable by the principle of equal marginal value – a small gain to the As is exactly compensated by a small loss to the Bs). This equity rule is the most favourable for the As and results from the low productivity of resources when applied to improving the health of the B. Note the awkward clash of principle: at c a small move in favour of the Bs along the frontier involves an equal sacrifice for the As (in terms of, say QALYs) so, in a sense, each is given an equal weight. But to move from c to a involves a much greater loss to the As than is gained by the Bs. Is such a differential justified in pursuit of more equal health overall?
Efficiency, equity and quality in health and health care

Figure 3 Five conflicting equity principles

Suppose the As and Bs are identical in all respects save that the Bs live in a remote mountainous region and the cost of treating them is much greater than that of treating the As. You are reliably informed that the Bs are five time more costly per case than the As. With the available budget you can locate at either a or c. But for every B case you will sacrifice 5 A cases. Efficiency is not at stake – both points are equally efficient – it is entirely a matter of equity (Culyer 2006).

Comparisons between inefficient and equitable allocation can also be made, of course. In figure 3, any point under the frontier is inefficient. If such a point also lies on either of the $45^\circ$ lines, then it would be equitable (according to which ever line it lay on) but also inefficient. But there is still no conflict between efficiency and equity, since efficiency could be improved by moving along the $45^\circ$ line towards the health frontier, thereby gaining efficiency (i.e. more health for both the As and the Bs for a given expenditure) without any loss of equity. The equivalent to point c below the frontier is a point again at which, with prevailing inefficiencies, small additional resources for one group would reduce the average health of the other by the same amount as was gained. Again, such a point would be deemed equitable by the QALY=QALY=QALY criterion but inefficient, and again a move towards the frontier would increase efficiency without necessitating any departure from equity.

Three further criteria remain to be explored: maintaining existing distributions, allocation according to need and equal per capita resources. The implications of a conservative equity rule is illustrated in Figure 3 by the line from the origin (0) through E. Where this line intersects the frontier, as at point d, the distribution of average health is the same as at E. Again the system is efficient. Again, inefficient allocations under the frontier but on the line 0d would be deemed equitable (and
inefficient) and greater efficiency does not require any necessary sacrifice of equity. ‘Need’ is an ambiguous concept (Culyer 1995, 1998, 2007) but is most commonly taken as being proportionate to health itself. In this view, the lower one’s health the greater one’s need. In Figure 3, the line through \( E \) to \( e \) is the inverse of the conservative line \( Ed \): the angle made by it and the vertical line is the same as the angle between \( Ed \) and the horizontal line. As one would expect, it is relatively favourable to the relatively deprived group and, again, is efficient. Points inside the frontier but on \( Ee \) would also be counted as equitable but they would be inefficient. Greater efficiency again does not imply less equity.

Corresponding to each of the allocation solutions discussed above is a distribution of health care resources bought out of the given budget. Only in exceptional circumstances, for example, would the rule of equal health imply equal resources to each group (in fact it requires there to be identical production functions for all). Similarly, a different allocation of resources corresponds to each of the various distributions of health that have been discussed. One can also ask the reverse question: “what distribution of health results from an equal allocation of resources to each?” The answer depends upon the production functions and is likely to produce health distribution results around point \( f \) in Figure 3.

Each of the various distributions of health depends upon the initial distribution of the burden of disease, the production functions for types of patient, classified not only by diagnosis but also by any other factor affecting the impact of health care on their health, the available budget, the extent to which health maximisation or health gain are objectives of the system, and on the equity criteria employed.

In the foregoing, it has been assumed that the least healthy also have the poorest prospects of health improvement. If we assume a different endowment distribution of health, the resultant distributions from applying the various criteria can change dramatically. Assuming instead that people with the less productive health production function have the lower initial burden of disease will change the ordering of points on the frontier. Such a case is shown in Figure 4. Here point \( e \) now represents a different initial distribution of burden. Point \( d \), representing equally distributed health, is now to the left of the other points on the health frontier and allocation at \( e \) is now well to the right.

It is a relatively straightforward matter to show that, even if the initial distribution of health is equal, maintaining its equality will generally require discrimination in favour of those with lower capacities to benefit. This raises an interesting ambiguity in the idea of horizontal equity. Given an equal initial distribution of health and therefore need, the horizontal principle ‘equal expenditure for equal need’ will imply increasing inequality unless the production functions are identical. On the other hand, ‘having an equal right to health’ requires ignoring need and allocating in favour of those with lower capacities to benefit regardless of their need. In each case, equity is associated with a particular form of equality – but it becomes important to enquire ‘equality of what?’ In the special case where there is an equal distribution of the burden of disease and each has on average an equal capacity to benefit (i.e. identical production functions), then the Aristotelian principle of horizontal equity will apply – equal expenditures are implied in every case. Otherwise, inequalities in the endowment and in capacities to benefit will generally require vertical adjustments to achieve any of the desired endpoints.
Figure 4 Impact of alternative endowment distribution
Conclusions

The general points to be made are the following:

- The concept of efficiency we use is that of Pareto efficiency: an allocation of expenditures between groups such that no redistributions can be made to benefit one group without reducing the health of another.
- This concept of efficiency includes that of cost-effectiveness: each health outcome is produced using the technologies requiring the minimum necessary expenditure.
- Efficiency and equity do not inherently conflict.
- An inefficient allocation can be equitable.
- An efficient allocation can be inequitable.
- An inefficient allocation can become more efficient without increasing inequity.
- An inequitable allocation can become more equitable without reducing efficiency.
- What is equitable depends upon the criterion selected but does not negate any of the foregoing implications.
- What is equitable often requires inequality in health and inequality in resource distribution per capita.
- Equality in health generally requires inequality in resource allocation.
- Equality in resource allocation typically leads to inequality in health.
- Allocation according to need typically leads to inequality in health.
- Greater equality in health requires investments in interventions that favour those with the greater burden of disease even if they have a lower capacity to benefit.
- In most cases one is dealing with vertical equity, with horizontal equity applying only to members within the groups being compared; each group being assumed to be in this sense ethically homogeneous, no one having any greater ethical claim on resources than any other.

Of the various distributive principles each may make their own choice. My own preference is for equality of health, mainly on the grounds that, since health is necessary for flourishing, an ultimate human good, and there seems to be no ethically acceptable reason to prefer that some should flourish more than others, all avoidable health inequalities should be removed. The contribution of health care to their removal will vary according to individuals’ initial states of health (their endowment of health) and their production functions (or capacities to benefit). It follows that the resources going to each will be unequal – but determined by the best estimate possible of the contribution they make to health equality. Needless now to say, those resources should also be deployed in a cost-effective way. This will imply that the assumption that a QALY=QALY=QALY be not used. Instead greater weight will be given to the health gains of the less healthy groups in society.

But the main point of all this is that different concepts of what it means to be equitable can have substantially different outcomes in terms of the distribution of health and the consequential distribution of health care resources. This does not represent a conflict between efficiency and equity but between equity and equity. You need to decide where you stand or, if you seek to advise, you need to discover where your client stands.
References


Marx, K. (1875) *Critique of the Gotha Programme*, (checked 22 May 2015 at: https://www.marxists.org/archive/marx/works/1875/gotha/ch01.htm)

