Journal of Epidemiology and Community Health, 1978, **32**, 3-9

Principles of allocation of health care resources

E. G. KNOX

From the Health Services Research Centre, Department of Social Medicine, University of Birmingham

SUMMARY The methods and principles of allocating centrally provided health care resources to regions and areas are reviewed using the report of the Resource Allocation Working Party (RAWP) (Department of Health and Social Security, 1976b) and the consultative document (Department of Health and Social Security, 1976a) as a basis. A range of practical problems arising from these papers (especially the report of the RAWP) is described and traced to the terms of reference. It is concluded that the RAWP misinterpreted aspects of social and administrative reality, and it failed to recognise clearly that the several principles on which it had to work conflicted with each other and demanded decisions of priority. The consequential errors led to *(a)* an injudicious imposition of 'objectivity' at all levels of allocation, *(b)* an unjustified insistence that the same method should be used at each administrative level in an additive and transitive manner, *(c)* the exclusion of general practitioner services from their considerations, *(d)* a failure to delineate those decisions which are in fact political decisions, thus to concatenate them, inappropriately, with technical and professional issues. The main requirement in a revised system is for a mechanism which allocates different priorities to different principles at each appropriate administrative and distributive level, and adapts the working methods of each tier to meet separately defined objectives.

Any centrally supported health care system requires a complex process for dividing a total resource into a number of component parts. Cash, manpower, buildings, and equipment must be allocated to various administrative levels, geographical regions, client groups, and institutions. The problems attendant upon formulating this process cannot be avoided; if they have been resolved badly on one occasion they must be resolved better on the next, but they will not go away.

The process responds, rationally, to three main constraints. These are: (1) the availability and transferability of the resources themselves, (2) the demands (or needs) which the service elects to meet, and (3) the standards of provision at which it elects to meet them. Health care planning is concerned essentially with achieving an accommodation between these three components, and with the exercise of priority decisions in those situations where needs cannot be met at acceptable standards within the available resource. It must be said that in the past many planning exercises have been limited to a consideration of only one or two of the three components, and experience has shown that an unconsidered element will always fall victim to a conflict between the other two. In the National Health Service (NHS) at the present time, for example, the pressures of demand and restricted resources have in many areas pre-empted any commitment to explicit standards of care.

The latest public intervention in health care planning in the United Kingdom is the report of the Resource Allocation Working Party (RAWP) (Department of Health and Social Security, 1976b). Its terms of reference were: 'to review the arrangements for distributing NHS capital and revenue to area health authorities and districts respectively, with a view to establishing a pattern of distribution responsive objectively, equitably and efficiently to relative need . . . and to make recommendations'.

The RAWP undertook one of a long series of studies carried out on behalf of the Department of Health and Social Security (DHSS) on matters of general policy. As with its immediate predecessor—which resulted in a discussion document entitled *Priorities for Health* (DHSS, 1976a)—the RAWP clearly felt and responded to the contemporary national economic disaster. The appearance of these reports coincided with a transition from a long period of slow but steady growth in health care investment, to a period of standstill and retrenchment, and they bear the marks of the political and

economic emergencies during which they were formulated. *Priorities for Health* was frankly a jumble of ideas (Knox, 1976, commentary), and its primary purpose was probably as 'cover' for the introduction of the new system of 'programmed budgeting' (pages 82 and 83), developed at the DHSS during a period of some years. The report by the RAWP, by contrast, is directly interventionist, so far as the NHS is concerned, and seems intended as the basis for future policy changes.

The RAWP proposals are, however, highly complex and, as subsequent discussions clearly show, provide grounds for debate and dispute on many points of detail. It now seems likely that after one or two cycles of application, and after some initial movement towards an envisaged redistribution of resources, a tide of legitimate and less legitimate caveats will overwhelm the operation and bring it to a halt. A new clarity of reformulation will then be required if momentum is to be regained.

The first purpose of the present paper is to extract and display, from among the many questions which have already arisen, those defects that appear to be the more important.

Its second purpose is to trace the origins of these defects and identify the confusions of premise or understanding from which they appear to spring. Its final purpose is to clarify the principles on which a subsequent formulation may be based.

Some technical problems

The intricacies of the RAWP proposals, and of the various objections already voiced, make the identification of the principal issues a difficult and an arbitrary task. The issues identified here, and set out below, are justified chiefly because they spring directly from the manner in which the terms of reference were set, and because clarification might determine the terms under which the next round of iteration proceeds.

THE ADMINISTRATIVE LIMITS

The terms of reference of the RAWP excluded consideration of health-related social services. Furthermore, although they purported to cover wide questions of 'distributing NHS capital and revenue', the subsequent specification of particular NHS authorities effectively excluded any consideration of the general practitioner and other general medical services. The subsequent limitation to the hospital services overshadows much of what follows. Because these various services interact extensively with each other, and because they may in different circumstances and in different geographical regions be 'traded off' against each other to different degrees, it is difficult, subsequently, to handle in any exact (or objective) sense the adopted criteria of equity, efficiency, or response to relative need. In addition, a resource allocation mechanism effectively limited to the hospital services creates (or consolidates and perpetuates) the principle of a bipartite service, which NHS reorganisation took such pains to avoid. The more transparent expediencies of the RAWP proposals spring largely from this single source.

EQUITY, NEED, AND OBJECTIVITY

The call for objectivity is central to the RAWP approach, as was manifest both in the terms of reference and the spirit of their interpretation. The instruments of objectivity adopted by the RAWP were a series of formulae, of which that covering 'revenue' (page 110) is the paradigm.

Objectivity is the negation of judgement, and they are opposites in a quite exact sense. A requirement for one amounts to a denial or circumvention of the other. It must be assumed that the opposition was recognised and was deliberate and that it was explicitly intended that managerial judgement should be curtailed on the grounds, presumably, that it had been tried and had failed. It was clearly intended that this managerial curtailment should pass through all three administrative layers of the NHS, right down to district level. That is, the instrument of objectivity was to be regarded as additive and transitive.

The criterion of objectivity marries well with the call for equity, although there is a residual choice of the basis on which the equity shall be based which will be examined later. But it contrasts sharply with the requirement of responsiveness to need, since need can never be determined objectively and is always perceptual in nature. At the level of the individual, too, it is clear that a responsiveness to need conflicts with a requirement for equity. At intermediate administrative levels, where effective management depends jointly upon an appreciation of subjectively determined value systems, projections and predictions, and good intuitive judgement there is a conflict between objectivity and efficiency. It is clear, therefore, that the structure on which the RAWP proposals are based is patterned with strains and incompatibilities.

The cracks show in the superstructure too as detailed examination of the various formulae shows. The full formula for regional revenue allocations, as given in the report (DHSS, 1976b), is too intimidating to reproduce, but may be represented more simply as follows:

$$\frac{x}{\Sigma_{x_r}} \cdot \text{Total}$$

where

$$x = \Sigma_k \left[\Sigma\Sigma_{ij} \frac{RP_i}{NP_j} \cdot NB_{ij} \cdot SMR_i \right]_k$$

and where NP = national population
RP = regional population
NB = national daily bed occupancy
i = condition (disease)
j = age group
k = sex
r = region
SMR = standardised mortality ratio

The essential element (x) is a moderately complex summation with two national and two local elements. The local elements are the numbers of the population in different age and sex groups, and a series of disease-specific sex-specific SMRs. The national elements are (again) the numbers in the different age and sex groups, together with the numbers of bed-days used for each age and sex group with respect to each disease. It is instructive also to expand the SMR values to their elements to provide the following:

$$x = \Sigma_k \left[\Sigma\Sigma_{ij} \frac{RP_i}{NP_i} \cdot NB_{ij} \cdot \left[\frac{RD_i}{\Sigma\Sigma_{ij} \frac{RP_i}{NP_j} \cdot ND_{ji}} \right] \right]_k$$

It can be seen firstly that most of the elements are obtained from national rather than local data, secondly, that to a first approximation, national and regional population elements cancel out, and, thirdly, that the regional revenue depends essentially upon absolute numbers of regional deaths (RD) with differential weights according to cause.

From one point of view this is a reasonable system. A major part of hospital expenditure relates to the few years leading up to death, and some ways of dying are more expensive than others. It is also equitable, in the sense that every member of the population with perturbations ultimately attracts a nearly fixed sum.

From other points of view, however, it is less justifiable, particularly when medical need, as opposed to equity, is considered. For example, a region with a low death rate has a population with a relatively extended life span, and excessive numbers of people in age groups which demand large expenditures. Health care planning teams concerned with care of the aged might see more sense in dividing by the SMR in the RAWP's formula, rather than multiplying. On the issue of equity, too, it is questionable whether the attracted benefit should be a standard amount per death, modified according to cause and age rather than be related to the duration of life and sojourn.

There are of course many examples where health care expenditures bear no relationship to the numbers of deaths, or even display an inverse relationship, particularly when non-hospital services are considered (spina bifida, traumatic paraplegia, chronic nephritis, hernia, epilepsy, asthma, mental retardation, mental illness). In addition the RAWP's methods provide no basis whatever for exercising centrally formulated preferences between competing areas of medical care, as proposed in the consultative document (DHSS, 1976a).

GEOGRAPHICAL BOUNDARY PROBLEMS

Administrative responsibilities within the NHS are defined geographically but the terms of the responsibility differ for different parts of the service. Responsibilities for general practitioner services (and personal social services) are defined in terms of the residence of the patients receiving care, as far as the area level. For hospital services, by contrast, the responsibilities are defined in terms of the location at which the care is provided. (This definition is modified in that some area authorities are responsible for hospitals outside their boundaries in order that they may maintain a reasonable balance of accommodation; they are in the main 'secondary-referral' units, such as hospitals for chronic mental illness or mental subnormality). There has never been any stated requirement to limit hospital services to the populations within their area boundaries and nothing in the 'five principles' of the reorganised NHS, to suggest that this was intended. With respect to the RAWP's terms of reference, and the reference to district allocations, it should be noted that statutory administrative responsibilities, of whatever kind, reach only as far as the area divisions. None of the financial allocations for any of the main branches of the service depends upon residential definitions at district level.

The dilemma facing the RAWP can be imagined. Their brief was essentially equity-based, and therefore population-based, yet it was limited to the only part of the health care or personal social services where responsibility is *not* at any level defined in terms of the recipients' residence. They were forced into inventing new terms of responsibility for the NHS authorities, in notional terms at least, and thence into considering compensatory cross-boundary flows.

Where area administrative authorities are reimbursed from a central source according to the bed-days of service provided—essentially the system which has operated in the NHS hospital service in the past—allowances made for cross-boundary flows lead to a well-known absurdity. Transfer of charges for patients receiving care outside the area of their residence, leads simply to a financial system based directly on the total number of bed-days provided, no matter for whom. The inertial properties of such a system are intrinsic, and cross-boundary corrections, at any level of the service, do nothing to assist a redistribution; indeed, they contribute to the inertia.

The RAWP's formulation, however, proposes that the remuneration of the health authorities should not be related either to the scale or the standard of the service, so that this immobility may be broken. In these circumstances cross-boundary corrections will still provide an inertial element but will no longer result in total standstill. It should be noted, however, that the manner in which cross-boundary allowances are made is an arbitrary function of the way in which boundaries themselves were drafted, and that since they were drawn for purposes other than the administration of the hospital services, a great part of the need for charge-transfers may arise directly from their having been put (for this purpose) in the wrong places. At least part of the problem is artificial in the sense that it would disappear if the boundaries were notionally redrawn. For most of England subregional allocation problems would probably disappear entirely in the wake of such a procedure, and only the inter-regional problems would be seen to be real.

The RAWP itself recognised the arbitrary nature of the boundary problem in its treatment of capital expenditures, and did not pursue its consideration, there, below regional level.

OPERATIONAL OBJECTIVES AND ADMINISTRATIVE EFFICIENCY

Target-setting techniques such as those adopted by the RAWP are not in themselves a sufficient basis for action and must be translated into truly operational objectives. They are a prerequisite of efficiency, one of the criteria set out in the terms of reference. It is necessary to set dates as well as targets and, for a complex operation, a series of intermediate targets with intermediate dates. The complexities of actively transferring resources between different administrative authorities were recognised by the RAWP, and enforced acceptance of a flexible approach to the process: a requirement only that progress should be in the right direction. However, it was not explicitly recognised that this amounts to an erosion, an abandonment even, of the principle of objectivity so much emphasised in earlier sections, and of the principle of 'transitivity'—the principle that 'formula methods' should operate additively through districts, areas, and regions. The contradiction was not squarely faced and no reconciliation was offered. As a result of this, NHS administrators have been left with a difficult problem of interpretation. Some have apparently accepted the report of the RAWP, and subsequent DHSS communications, as an instruction that the formulae be applied right down to district levels; others believe that below the level of inter-regional allocations, there is no need to take the report of the RAWP literally.

The separate treatments of capital and revenue targets in the report of the RAWP are justified in terms of equity; it was intended that regions and areas under-capitalised in the past must be allowed some special compensatory capacity in their budgets in order to catch up. The retention of the distinction between capital and revenue, and of central control of their overall balance, is supported by tradition; in addition, logical justifications can be made. However, it proved impractical to carry the process of capital-deficit assessments below regional level.

Furthermore, the methods used for evaluating capital short-fall were based upon simple amortisation techniques and were entirely unrelated to any assessments of running costs. The difficulty is accentuated in that in some parts of the report the notion of a balance between capital and revenue has been interpreted in the sense of a 'trade-off', so that a region or area which opts for more of one may do so at the expense of the other. In practice, of course, relatively few capital plans *save* revenue; on the contrary they more often *stimulate* it, and capital plans may often be adopted explicitly for this purpose.

Next time round

The list of problems discussed above is far from comprehensive. The service increments for teaching (SIFT) were not discussed for example, although they are considered by Snaith (1978); these calculations are not clearly related to any of the criteria given in the terms of reference and they raise new issues, especially concerning standards. Enough has been said to show that the RAWP's proposals are logically untidy, contain many contradictions and inconsistencies, and engender serious difficulties of interpretation and implementation at practical levels.

The report itself suggests that its main difficulties have arisen from the urgency of the task and from

the inadequacy of suitable data (on need) from which to work. If this were so, time and industry would be all that were needed to put matters right. However, there is much in the above analysis to suggest that the problems are deeper, and spring from the manner in which the principles of the exercise were declared and interpreted.

A CONFLICT OF PRINCIPLES

It appears upon analysis that the basic problem is an unusual one. It springs not so much from an initial failure to declare the principles upon which a distribution method was to be devised, as from the declaration of too many. It was to be based jointly upon criteria of objectivity, need, equity, and efficiency, and one of the failings of the RAWP was to recognise and to act upon the fact that these principles in practice conflict with each other. Professional workers with technical, scientific, or mathematical backgrounds may be unfamiliar with the notion that sound arguments based upon legitimate premises and accurate data can lead to inconsistencies. They tend to assume that inconsistencies arise through errors in the data, or mistakes in the logic or arithmetic, and they seek solutions through their technical review and revision. They may be unwilling to accept (at first) that in those fields of study subsumed under the titles 'political science' and 'social justice', conflicts and inconsistencies are normal phenomena and may not be susceptible to correction through attention to logic.

An example of this kind of contradiction is developed by Rawls (1972) in relation to a discussion of the theory of law, and concerns the conflict between actions designed to further the interests of individuals, and actions designed in the interests of society as a whole. The first is associated with the notion of personal justice; the second with the 'utilitarian' principle of the greatest good for the greatest number. It must not be assumed that the one is simply an aggregate of the other; indeed this is demonstrably false. Any redistribution of rights (to goods, services, etc.) designed for the larger purpose necessarily infringes upon the (otherwise) rights of *some* individuals. Rawls (1972) examines a compromise principle—that of minimising the sum of individual injustices—but he recognises the need to attach subjective values to the various kinds of injustice suffered. If his analysis is accepted, then a realistic attempt at resource allocation must recognise that the criteria of efficiency (utility) and of response to need (social justice) are ultimately irreconcilable and that they may be accommodated only through the introduction of value judgements and at the expense of objectivity.

Miller (1976) goes beyond the bounds of the theory of law and analyses the basis and the operations of social justice. He identifies three main principles upon which social justice may variously be founded: the principle of desert, the principle of right, and the principle of need. As in the analysis of Rawls (1972) the different approaches are shown to be mutually antagonistic; and even within them he demonstrates the existence of further contradictory subsets. For example, the notion of need is resolvable into several distinct and conflicting interpretations. The principle of social justice based upon desert scarcely enters the report of the RAWP, but the contradictory nature of the alternative criteria of rights (equity) and need is starkly clear. Practical circumstances will often enforce a *choice* between them rather than permit a reconciliation within a single formula. Miller (1976) also has a lesson for the basis of that choice; none of the various interpretations of social justice can lay claim to be correct. The choice is *necessarily* arbitrary and imposed, arising legitimately (for want of an alternative) from the purposes of the imposer and his image of the society in which he operates.

A third constellation of related ideas was put forward by Arrow (1963). His approach is supported by mathematical proofs and he describes in formal terms the processes through which individual preferences and value judgements are assembled into corporate preferences and judgements to form the basis of policy. For example, how are individual needs, assessed by individual patients or by their medical attendants (or both jointly), aggregated into statements of corporate priorities at various administrative levels, so that resources can be allocated accordingly? Arrow demonstrates that this is far more complex than a process of simple addition (cf., the summation processes of the RAWP's formulae); indeed, that it cannot be achieved through any formal process whatever. The inherent conflicts contained within these processes are such that consistent policies can be achieved only through their imposition. The point is not simply that corporate policies necessarily over-ride the wishes of at least some individuals—this is Rawls's point—but that formal (that is, objective) methods for achieving a compromise are incapable of leading automatically to a rational plan. Without the imposition of consistency, mechanisms such as majority vote, proportional representation, or the use of mathematical formulae, will *always* produce absurdities such as circular preference-orders for choosing between alternative actions. Arrow (1963) has provided infinite consolation for many experienced committee workers

through demonstrating that this is *inevitable* and is not a question of simple incompetence. In a multi-layered structure such as the NHS, where the centripetal assembly of needs into a common policy must pass through several strata, these constraints must apply at each level and it is legitimate, or even necessary, that different ways of imposing consistency will be appropriate at each step.

PRIORITY OF PRINCIPLE

It should be clear that there is no hope of finding a correct solution, in the sense that somewhere it exists and is only waiting to be found out. The situation contains a complex pattern of conflicts between the several principles on which it is based, and the problem is one of apportioning priorities between these principles. If consistency is desired then these priorities will have to be imposed and it is totally diversionary to engage upon a consideration of arithmetical questions—such as, the use of SMRs, London weightings, use of beds, and the SIFT—until the conflicts are faced and the priorities decided.

The report of the RAWP is defective on this count. A second fundamental defect arises from the attempted concatenation of two quite separate processes within the terms of a single formula and mechanism. The two processes are *(a)* the assessment of individual medical needs and their aggregation into group needs, and *(b)* the disaggregative process of allocation. Needs-based planning systems are intrinsically cyclical (cybernetic) and the afferent and efferent streams of activity are necessarily separate. Without such a structure the process becomes rigid, centrally doctrinaire rather than responsive to assessed needs, and in the present context simply a rationing system. The RAWP concatenation leads also to the conceptual error of 'transitivity' and the false 'principle' (for example, sections 1.4, 3.3.1) that the same distributional criteria must be used consistently at each administrative level. There is in fact no basis for this assertion either in experience or logic, and elementary analyses of social processes (Arrow, 1963; Rawls, 1972; Miller, 1976) confirm its invalidity.

A further guideline for the RAWP's successor relates to the criterion of objectivity. Objectivity is a valuable property in its proper place, but it is an operational contrivance rather than an over-riding principle and must not be allowed to bar the introductions of value judgements—in *their* proper place. It is indeed strange that one of the two recent DHSS papers on resource distribution (DHSS, 1976a) should so explicitly demand the introduction of politically and professionally determined value systems, while the other (DHSS, 1976b) should so explicitly deny them.

There is only one way of fitting all these requirements into a single working system. This requires that each level of administration or planning has to arrange its guiding principles in a different order of priority. It is possible to argue, even, that the very development of stratified administrative arrangements represents a general response to this requirement, and that health services are not alone in this respect.

A STRATIFICATION OF PRIORITIES

Equity, as a principle, must be awarded its chief priorities at the centre rather than at the periphery. That is, unless equity were established between regions it would be quite impossible to establish it at area and district levels, whereas if equity *were* established between regions, then serious disparities of access at the level of the consumer would be less likely to occur. At the other extreme, equity at the individual level is a nonsense; it would imply entitlement to equal provision of service regardless of the individual's needs or state of health. At the individual level it is the principle of need that has priority and its determination is primarily a professional matter, or a matter of joint concern between the patient and his professional attendants. At the national level, with the emphasis on equity, the issues are chiefly political.

At intermediate administrative levels the main concerns are efficiency, effectiveness, the maintenance of standards, and in achieving an accommodation between the constraints of equity (coming from above) and of assessments of need (coming from below). Objectivity is more readily associated with the achievement of equity at national levels than with the functions associated with clinical judgement or with efficient management.

Once the spectres of transitivity of method and universal objectivity have been exorcised, and once the inherent conflict of principles has been recognised, and their priorities separately decided for each administrative level, a simpler, more practical, and more credible basis for long-term resource allocation begins to emerge. For example, equity might be made almost the sole basis of resource allocation targets down to any level where cross-boundary transfers of demand are negligible (say, less than 5%). All inter-regional distributions might converge upon equity-based targets and in many parts of the country (although not all) this could be carried to the level of areas. Major cross-boundary transfers are not then a factor to enter into the calculations themselves, but are taken to indicate the level at which other

considerations—such as efficiency and need—must begin to take precedence. Nor is there any need, now, to ignore expenditures on general practice and other general medical services, or even the health care-related personal social services. Inclusion of these expenditures within calculations appropriate to equity-based levels might well generate a desire for better administrative control of the balance between them; this would be a highly desirable outcome which the report of the RAWP, in its present form, effectively abandons.

The basis of equity between population groups must be chosen from among several alternatives. They include the size of the population, the number of deaths a year, the number of births, some combination of these measures, and a number of possible elaborations. Secondly, the allocations may be based upon cash expended, or upon levels of service provided, taking due record of the various costs of providing that service in different regions. The report of the RAWP, in its present form, incorporates some of these factors but it is necessary to be clear about two points. Firstly, the decisions are political decisions, taken in order to implement a political principle (that is, equity), and data analyses, particularly elaborate analyses, have a relatively limited supporting role. That is, the targets may be determined as £x per life + £y per death + £z per birth in a more or less arbitrary manner, perhaps with a regional cost-of-provision weighting. The actual calculation of the targets may then be totally objective, although it would be misleading to pretend that this applied to the manner in which the formula was compiled, or to decisions about the rate at which the targets were approached.

A corollary of the last point is the existence of a continuing managerial role at the centre (DHSS) for conducting the steering process. Furthermore, since both capital expenditure and service increments for teaching are major instruments in the control of this process, they might both, themselves, be withdrawn from the equity-based allocations. The national responsibility for medical manpower planning, and the possibility of a regional preoccupation with capital 'games' as a means of pre-empting expenditures, both add force to this suggestion.

Once it is understood clearly that the method for setting targets on the macro-scale is arbitrary and political, and that appearances of objectivity here are illusory, it can be seen that assessments of regional variations of medical need are likely to impose only marginal perturbations. In the early periods, when practical rates of movement towards the targets are a far more powerful constraint upon progress than the setting of the targets themselves, these variations may probably be entirely ignored. The development of appropriate methods, appropriate data, appropriate arguments, and appropriate consensus-based value-judgements could probably with advantage take place in the context of the evaluative mechanisms set up within NHS management specifically for these purposes. The requirement might enforce a clearer recognition of the fact that differences in the functions of regional and area administrations are not simply questions of scale, that their differing objectives and differing priorities of principle need to be more explicitly differentiated, and their structures and methods of working must be separately adapted to meet their differing functions.

Reprints from E. G. Knox, Health Services Research Centre, Department of Social Medicine, University of Birmingham, Edgbaston, Birmingham B15 2TJ.

References

Arrow, K. J. (1963). *Social Choice and Individual Values.* Wiley: New York.

Department of Health and Social Security (1976a). *Priorities for Health and Personal Social Services in England: A Consultative Document.* HMSO: London.

Department of Health and Social Security (1976b). *Sharing Resources for Health in England. Report of the Resource Allocation Working Party.* HMSO: London.

Knox, E. G. (1976). Priorities for health: a manipulative document? *Lancet*, **2**, 789-792.

Miller, D. (1976). *Social Justice.* Oxford University Press: London.

Rawls, J. (1972). *A Theory of Justice.* Oxford University Press: London.

Snaith, A. H. (1978). Subregional resource allocations in the National Health Service. *Journal of Epidemiology and Community Health*, **32**, 16-21.