Formula funding of health services:
learning from experience in some developed countries

DISCUSSION PAPER
Number 1 - 2008

Department “Health Systems Financing” (HSF)
Cluster “Health Systems and Services” (HSS)
Acknowledgements: This paper draws heavily on my book *Formula Funding of Public Services* published by Routledge in 2007. Preparation of the book was funded by the United Kingdom Economic and Social Research Council fellowship R000271253. The paper benefited considerably from comments from Guy Carrin and Varatharajan Durairaj, and from participants at a seminar on 15th April 2008 at the World Health Organization (WHO) headquarters in Geneva, Switzerland.

E-mail address: pcs1@york.ac.uk

The views expressed in documents by named authors are solely the responsibility of those authors.
Formula funding of health services:

learning from experience

in

some developed countries

by

Peter C. Smith
1. Introduction

A key requirement for any health system is to ensure that the available public funds are directed to local organizations in line with health system objectives. Such funding seeks to give local institutions - such as local governments, local administrations and health authorities - both the financial capacity and the incentive to fulfill their objectives. This paper focuses on one particular aspect of the financing mechanism: the use of mathematical formulae to determine the magnitude of funds directed towards local health agencies. The use of such formulae has become increasingly widespread in developed health care systems, and offers enormous scope for ensuring that funding is aligned with national health system policies. In particular, a systematic approach towards funding local agencies is an essential prerequisite of successful decentralization policies. However, formula funding also embraces a number of perils that must be addressed if it is to be successful.

This paper summarizes the 'state of the art' of formula funding of publicly funded health care in developed countries, and the lessons that can be learned for implementation in more challenging settings in low and medium income countries. It begins with a brief overview of the funding flows that are addressed by formula funding, and explains what formula funding is. Section 3 examines the various rationales for formula funding under three headings: efficiency, equity and political. Section 4 discusses the main building blocks of any funding formula, and Section 5 then describes in some detail examples of formula funding. The paper concludes with a summary of the steps that need to be taken to introduce formula funding, and a brief discussion of potential future development.

2. The flow of public funds in health services

The World Health Organization (WHO) has argued that national governments play a crucial stewardship role in ensuring that health service funds (from whatever source) are used efficiently and equitably, and that health care agencies are incentivized appropriately. Figure 1 offers a schematic representation of the flow of public funds implicit in the finance of most health systems. An important source of central government funds is taxation, paid in a variety of forms by citizens and businesses. This might in some low-and middle income countries be augmented by a variety of donor funds, creating a pool of revenue available to the central government. The government must then decide on how it will allocate the funds to support locally delivered health services, either wholly or in part.

The central government might pay health care providers directly, as in the United States of America Medicare programme for older people. However, national governments usually devolve health care purchasing powers to local organizations, such as states or various forms of local government, local health authorities, health insurers, or voluntary bodies. I shall usually refer to these devolved institutions as local agencies. They are often financed wholly or in part by grants-in-aid from the central government.

---

1 It is being tried out in some low- and middle-income countries as well. To name a few, examples of some form of formula funding are found in Bangladesh, Chile, Columbia, Ghana, India, Kenya, the Lao People's Democratic Republic, Nigeria, South Africa, Tajikistan, and Zambia.
central government funding mechanism (B)\(^2\). However, note that the central government's financing role might be restricted to redistributing between localities funds that were originally collected by the local agencies, so that all agencies can offer similar benefits packages. This arrangement is often used in systems of local government or social insurance. However, even though local agencies might be responsible for collecting premiums or taxes, from an analytic perspective such redistribution arrangements are equivalent to distributing nationally collected tax revenues between local agencies.

**Figure 1: The flow of funds in the health system**

Local agencies may be solely reliant on national funds, but are often able to augment their revenue with local taxes (C), and sometimes donor funds allocated directly to the local agency (H)\(^3\). They then purchase services from providers (F). In some circumstances, the distinction between purchaser and provider may be unclear (for example, hospitals may be directly provided by local governments). However, even where there is no explicit payment mechanism, local agencies must in practice purchase their services from vertically integrated providers. Finally, the service user might pay a charge to the local administration or (more commonly) directly to the service provider (D). In some countries this payment might take the form of ‘informal’ reimbursement of the provider. Of course the relative magnitude of all of these funding sources varies considerably between countries.

This paper is centrally concerned with just one of the funding flows represented in Figure 1: the mechanisms for funding local agencies, or purchasers of health services,

\(^2\) In some countries, government funds intended for local agencies may flow through intermediate organizations, such as independent insurance agencies. Although sometimes important for institutional reasons, these arrangements do not materially affect the principles of formula funding discussed in this paper.

\(^3\) In addition, in some countries, donor funds are targeted directly at beneficiary individuals without necessarily passing through any government agency.
from national revenues (B), or for reassigning local revenues from one local agency to another. As noted above, these might be local or central government agencies, but could also be social health insurers or other non-governmental bodies. The national government’s problem is to design a payment mechanism that is aligned with and promotes its policy objectives. Flows other than B in the diagram are referred to only when they are material to this policy problem.

Determining the level and distribution of local financial allocations can give rise to serious challenges for national governments. Firstly, it is a difficult technical exercise to determine where limited public funds are best spent. Secondly, any funding mechanism introduces powerful incentives for local organizations, and it is important to ensure they are consistent with health system objectives. Thirdly, the political ramifications of any geographical funding choice can be acute, particularly when representatives in parliamentary democracies are elected on a geographical basis. And finally, the national government requires reassurance that public expenditure is being spent locally in line with intentions, yet monitoring the effectiveness and efficiency of local spending is often a difficult undertaking.

2.1 Traditional approaches to health service funding

There are numerous ways in which the local allocation of health system funds could be determined. At its crudest, the distribution could be based on political patronage, perhaps in representative democracies rewarding localities according to their political support in the past, or their importance for the government in the future. Although few governments would admit openly to engaging in such patronage, there is ample evidence to suggest that to some extent it informs many allocation systems that are supposedly non-partisan.

Another approach in widespread use is to distribute public funds to local health agencies according to historical precedent. Politically, it has the great attraction that it minimizes disruption to existing health services, and avoids potentially large swings from year to year inherent in other allocation mechanisms. Its popularity is manifest in the way that more systematic approaches to distribution are frequently abated by ‘damping’ mechanisms that reduce the magnitude of year-on-year financial losses and gains experienced by localities. However, sole reliance on such methods would leave a government hostage to history, would perpetuate any existing unfairness, and leave a government powerless to react to changed circumstances or to implement new policies.

A third possibility is to allocate funds according to bids submitted by localities, or to make allocations contingent on some measure of local performance. In principle, this approach has much to commend it. If undertaken properly, it could ensure that public funds were spent cost-effectively, in line with national policy intentions. Its major weakness is that it is a demanding technical task to identify how much performance variation is due to the efforts of local agencies. This approach therefore usually entails large transaction costs, in the form of central scrutiny and policing, and the preparation of bids by local organizations. It also makes local budgets dependent on the quality of local management, and so may lead to large geographical inequalities. Moreover, unsuccessful localities may perceive that the allocations have been made according to patronage rather than the quality of bids, leading to further potential for perceived unfairness.
Finally, financial allocations could be made according to how much localities actually spend. In many circumstances, this approach contradicts principles of good public finance, as it is likely to encourage spending in excess of efficient levels. However, it forms an element of many systems of matching grants from central government to local administrations, and can be used to encourage spending when local organizations would otherwise spend below efficient levels.

In practice, most systems of financing local health administrations contain an element of all four types of mechanism. However, a fifth approach – allocation by mathematical formula – is increasingly becoming the favoured approach to determining local financial allocations. It can be defined in broad terms as the use of mechanical rules to determine the level of public funds a devolved local agency should receive for delivering a specified health service.

2.2 What is formula funding?

Under formula funding, mathematical rules that determine the magnitude of the funding received by a local agency are specified in advance. Those rules might be very simple (for example, a fixed amount of per capita funding per annum) or very complex. They might also be to some extent augmented by other funding mechanisms (for example, additional specific grants from the national government, or local taxes). The overarching objective of formula funding is to contribute to the creation of a budget for the local agency with which it is expected to fulfil its duties, in the form of organizing and purchasing local health services.

There are two broad approaches to formula funding. The first reimburses the local entity on the basis of some measure of local activity, for example a count of the number of service users. Such case payment mechanisms are widespread in health care, for example using counts of inpatients, and the highly developed systems of diagnosis-related groups (DRGs) in place in many developed countries [3]. They are especially relevant when an unambiguous indicator of a patient’s need for the service can be established. However, they can be vulnerable to perverse incentives to create unwarranted or inappropriate service utilization. Case payment methods give rise to a variable budget for the local agency, based on recorded activity. Case payment mechanisms may have a useful role to play in funding some elements of local health care, and I make brief reference to them in Section 5. However, because the subject of this paper is on fair funding of entire populations, independent of service utilization, the main focus is on a second approach to formula funding, known as capitation funding.

The capitation approach reimburses local agencies according to the expected level of local activity. Typically, this takes a measure of the size and characteristics of a locality’s population, for example in the form of risk factors such as levels of disease and poverty, and infers the expected level of local service expenditure without reference to actual local health service use. These methods have become known as capitation funding methods because they are based on population counts. They circumvent some of the perverse incentives inherent in case payment, but their effectiveness depends on how successfully the capitation payments are adjusted to account for variations in population characteristics. In general, capitation methods give rise to a fixed budget for the recipient of funds, based on population size.
Three institutional requirements must be in place for formula funding to be relevant. First, the organization and purchasing of health services must be to some extent devolved. At one extreme the devolved entities might still be very large governmental organizations, in the form, say, of the provinces in the People's Republic of China (typical population of 100 million). At the other extreme, the devolved entities might be individual citizens in receipt of vouchers to spend on specified health services. Whatever their form, the devolved entities are then responsible for using the funds they receive to purchase the intended health services.

Second, there must be adequate data, available on a consistent basis across all local organizations, to which can be applied a mechanical formula that determines the level of funding to be allocated to those organizations. The data should of course be verifiable and timely. Also, the formulaic rules should be specified \textit{ex ante}, so that there are no immediate provisions for altering the consequent level of funding. The possibility of \textit{ex post} adjustments to the rules would imply that formula funding budgets are being moderated by considerations of historical spending or political patronage.

Third, there must exist some incentive to adhere to the financial allocation implied by the formula. Formula funding is a mere ritual if the recipients of funds can with impunity ignore the allocations implied by the limits. Sanctions and rewards may take many forms. For example, local agencies that spend in excess of their assumed spending limits may be required to fund the excess from local taxes. Or an agency’s chief executive might face dismissal if its financial allocation is persistently exceeded. There is no requirement that the sanctions should be as ‘hard’ as these examples, but there must be some incentive for recipients of formula finance to take notice of their allocations.

In the developed world, formula funding is becoming the dominant mechanism for devolving health system finances. Outside the United States, capitation methods have become ubiquitous as the means of funding devolved local governments, local health authorities or social health insurance funds [4, 5]. The next section examines the various objectives that a system of formula funding may seek to address.

3. The rationale for formula funding

There are three broad reasons for adopting a formula funding approach, reflecting efficiency, equity and political objectives [6]. This section summarizes the rationale for formula funding under these three headings.

3.1 Efficiency

Economic efficiency has a number of connotations. The two most fundamental notions are allocative efficiency (the extent to which allocations of resources are in line with society’s preferences) and managerial efficiency (the extent to which agencies perform functions at least resource cost). Formula funding is intended to address both aspects of efficiency. It seeks both to align resource allocations with national priorities, and to offer incentives that promote technically efficient practices amongst local agencies. This section briefly elaborates on some of the more important efficiency objectives underlying most formula funding mechanisms.
Securing allocative efficiency

No coherent system of financial transfers can be developed without first establishing clear objectives for the financial regime. Once objectives have been set, the fundamental efficiency argument for implementing formula funding is that – if properly designed – it allows the government to implement an optimal allocation of finance in line with those objectives. Any deviation from such a formula implies a reduction in the effectiveness with which funds are used, and therefore a loss of efficiency. For example, if objectives are set in terms of service outputs – such as maximizing health gain – the formula should in principle be designed to allocate funds so as to secure that objective at a national level, subject to the national budget constraint.

In principle, this implies identifying local budgets that are as well aligned as possible with estimates of the local prevalence of disease that is amenable to cost-effective treatment. Any departure from this criterion leads to a loss in the total health gain secured at the national level within the national budget constraint. Of course the key empirical challenge then becomes one of successfully identifying appropriate data and designing a formula in line with these principles.

Creating appropriate incentives

It is important to note that funding formulae are more than reimbursement mechanisms. Carefully designed, they can allow the government to put in place financial incentives for local organizations to respond in line with national objectives. This may often involve a judicious mix of payment instruments. For example, if a pure capitation system is leading to inadequate supply of certain surgical treatments (because there is no reward for extra activity), then the capitation payment might be augmented by an element of case payment funding to encourage extra provision.

The main incentive introduced by pure capitation funding is to reduce demand for health services, as the local agency receives the same revenue regardless of output. The reduction in output associated with capitation may in some circumstances be desired. In particular, for health services such as communicable diseases for which there are clear prevention opportunities, it may be perverse to reimburse local health agencies for the number of cases treated. Properly designed fixed budgets, created through the use of capitation payments, are more likely to stimulate preventative measures, and a desired reduction in demand for treatments.

However, the reduction in demand may be undesirable. Patients in need may find it difficult to receive treatment, and treatment costs may be shifted to other agencies or patients’ families. Furthermore, the use of fixed budgets also in itself offers no incentive to improve the quality of health services (indeed there may be strong incentives for ‘quality skimping’).

In contrast, the use of case payment methods (such as DRG methods) introduces a different set of incentives. It encourages the treatment of patients whose expected costs are lower than the associated case payment reimbursement. This might be beneficial (if those patients will benefit from treatment) or adverse (if the benefits are questionable). It can also encourage more intensive treatment of patients if such treatment leads to an upgrading in the severity of the DRG, and therefore increased reimbursement. Like capitation methods, pure case payments offer few incentives on
their own to promote the quality of care.

To illustrate the different incentives under the two approaches, consider the funding of accident and emergency services. One possibility is to attach an annual case payment of US$ 40 to each user. The agency’s revenue is then the number of users multiplied by US$ 40. However, if the probability of a member of the ‘at risk’ population requiring the service in any one year is only 0.15, then the equivalent capitation payment would be US$ \( 40 \times 0.15 = US$ 6 \) per head of population. The local agency’s revenue would then be the size of the entire ‘at risk’ population multiplied by US$ 6. The revenue would of course be the same under both payment mechanisms if 15% of the population did in fact use the accident and emergency service. However, the implicit incentives for demand management are very different.

**Dealing with information asymmetry**

A central concern of many governments is the extent to which their methods of financing local organizations rely on data provided those same localities. Under many circumstances, such data may be vulnerable to manipulation or create perverse incentives. For example, a nineteenth century proposal to reimburse local authorities in the United Kingdom of Great Britain and Northern Ireland according to the number of ‘indoor paupers’ – or workhouse residents – was rejected (a) because the count of indoor paupers was vulnerable to fraud and (b) it may have encouraged unnecessary use of the workhouse (in order to raise the proposed basis for payment, the numbers of indoor paupers) [7]. Even in systems in which such responses are rare, a suspicion that they exist may undermine the credibility of the funding mechanism. Many of the most creative systems of formula funding therefore seek deliberately to obviate the direct use of health service data, so reducing the danger of moral hazard in the actions and reporting of local institutions.

The use of formulae can allow the development of estimates of local spending needs that rely only on objective indicators of relative needs, and that are therefore independent of special pleading on the part of the local organization. The government can thereby reduce the need for detailed scrutiny of the case for funding localities, or the accuracy of information provided. The locality does not need to make a case to the government, and is freed to concentrate on the organization and delivery of local health services. Furthermore, and perhaps most importantly, it has no incentive to distort behaviour in order to suggest a need for more funding. Of course, in practice, even under a formula funding system, localities are likely to make representations to the government about the accuracy or fairness of the formula. However, special pleading is likely to be less profitable and more easily rebutted if a well-designed formula is in place.

Implicit in these arguments is the belief that the use of formulae economizes on agency costs, in the form of the analysis, audit and oversight required to construct fair budgets. Almost any other method, such as basing budgets on bids by localities or historical precedent, has great potential for adverse responses on the part of localities, in the form of distorted information or gaming behaviour. Formula funding can therefore offer a practical means of managing the information asymmetry implicit in the relations between national government and locality.
Formal funding can also have important implications for managerial efficiency, as it leads to the creation of clearly defined budgets for local organizations and promotes decentralization. Budgets are a central feature of most systems of health service delivery, and are associated with numerous incentives to local efficiency. In particular, the setting of clear budgets is a prerequisite for the sort of decentralization that offers local agencies the freedom to respond to local circumstances, to innovate, and to seek out economies, in line with the prescriptions of the new public management [8]. Of course, such approaches are predicated on the existence of good quality outcome measures with which to hold organizations to account. However, if such measures are in place, the use of mathematical formulae is the foremost means of enabling a national government to put in place a robust budgetary regime.

3.2 Equity

Many systems of formula funding seek to promote some concept of equity. The pursuit of equity might be valued for its own sake, or it might be valued because it secures acceptance for the government’s funding regime for local agencies. It is usual to divide equity concerns into the two broad principles of horizontal and vertical equity. A concern with horizontal equity suggests equal treatment of equals, whilst a concern with vertical equity suggests that those who are in more need (however that need is defined) should in some sense with higher priority [9]. In practice, many stated equity principles often suggests that national equity objectives are often vague or even misleading. For example, the National Health Services in the United Kingdom of Great Britain and Northern Ireland traditionally sought to offer equal opportunity of access to health care for those in equal need. Yet, if implemented in its pure form, this criterion raises profound questions of how (for example) people living in rural areas are to be guaranteed the same access to all health services enjoyed by those living in conurbations, within limited budgets. In general, health system equity concerns can appear to be a jumble of horizontal and vertical principles, addressing both access to health care and health outcomes achieved by local services. For a full treatment see Hauck et al [10].

When put into operation, most of the high-minded equity principles translate into a more operationally tractable policy objective of enabling local agencies to deliver some ‘standard’ benefit package of health services. The standard benefit package is usually defined in terms of a basket of health care and other services. I am not concerned in this paper how that package is set (although that is a separate, crucial consideration). In practice, most governments avoid making an explicit statement of what the standard package might be. Instead, many systems adopt a conservative approach to specifying the package, implicitly adopting whatever services are – on average – currently delivered in the country as a basis for setting budgets. This interpretation allows for the possibility (for example) that rural areas may have lower levels of service than their urban counterparts. Therefore, implicit in many equity criteria is the understanding that equity will be pursued ‘up to a point’, but that it must to some extent be moderated by considerations of cost and efficiency.

However defined, the chosen benefit package implies a certain level of expected expenditure in each locality, which I term the area’s “spending need”. This will depend on the geographical, demographic, social and economic characteristics of the
area. The characteristics to be taken into account in calculating an area’s expected expenditure will be determined by data availability and the nature of the chosen package. They are discussed in more detail in Section 4. The economist’s usual approach towards inferring expenditure needs is to develop a cost function, which seeks to model the expenditure requirements of delivering a specified package of outputs. Many of the empirical techniques of formula funding discussed in this paper use cost function methodology.

Implicit in any formula funding system is a redistribution of finances between local agencies. Without a national financing function, the costs of delivering the standard package of health services would be borne in their entirety by the local area. Suppose that the sole source of revenue was a local tax base, such as a payroll tax. Then, assuming no revenue from any other source, the local tax rate would be given by the ratio of expenditure needs to the size of the tax base. In general, therefore, the local tax rate required to secure the standard benefit package (and therefore the tax burden for identical citizens living in different areas) would vary substantially depending on (a) the local area needs and (b) the tax revenue base of the area.

If, on the other hand, all local agencies levied a uniform local tax rate, without a national fiscal equalization function, large disparities in the package of care that localities were able to offer would arise. This extreme scenario shows why central redistribution of revenues is almost always needed in order to avoid large inequalities in tax rates, levels of service, and user charges. The exact form of redistribution may vary between nations. For example, it could take the form of a subsidy from national government funds, biased in favour of the more disadvantaged areas, or alternatively a series of direct transfers between ‘overfunded’ and ‘underfunded’ local agencies. The analytic task is however the same whatever redistributive methods are used. It entails the development of measures of local spending need.

Note also that, if the local administration has some discretion over the nature of local health services, the objective is to enable local administrations the opportunity to deliver a standard level of service at a standard local tax rate, whilst charging users a standard charge. Local agencies may then have the freedom to vary some elements of the package of services, the local tax rate, or user charges. It will then usually be the case that - if a local agency chooses to vary the package - the entire burden of any extra expenditure falls on the locality, for example in the form of additional local taxes, higher user charges, or diminution of other public services.

3.3 Politics

The formulaic approach to funding has become popular not just because it can lead to more efficient and equitable outcomes than other methods, but also because it can serve a health system’s more general objectives. Of course, those objectives often include equity and efficiency criteria. But they may also include important political considerations such as: allowing the criteria for funding to be set out explicitly; treating the budget-setting process systematically; promoting accountability; avoiding the need for case by case scrutiny of budgets; binding politicians, bureaucrats and other parties to a set of distributional rules; and more generally economizing on political effort.

McLean [11] discusses at length the political context of formula funding in the United Kingdom public services. He cites a government minister, whose concern was
manifestly not with the technical content of the payment mechanism. Instead, the principal interest was in the ‘exemplifications’ – the impact of alternative technical choices on the finances of individual local governments in the United Kingdom of Great Britain and Northern Ireland (p 127). It would be a very brave minister that made decisions on formulae purely on the basis of technical advice, and did not also acknowledge political considerations. In the extreme, any formula can be supplemented with non-formulaic elements, or additional formulae that target specific classes of localities. Of course this may not always be in a government’s interests, as it suggests special treatment for the beneficiaries. However, the political consequences of an explicit extra-formula payment may outweigh any negative consequences. In this context, Glennerster et al [12] quote a ministry official who claims the government could ‘make music’ with the budgetary allocations.

Analytically and ethically, it is hard to justify some of the devices used by governments to secure a desired distributional outcome, whilst nominally adhering to the use of formulae. Furthermore, if a government’s choice brazenly contradicts technical consensus it runs the risk of exposure by the media or parliamentary scrutiny. However, the complexity of many funding formulae means that there is rarely vigorous public debate of the government’s choices. The reason there is so much flexibility is that there are usually numerous judgements to be made in implementing a funding mechanism that transcend purely technical considerations. The dimensions of freedom include:

- The choice of analytic methods;
- The choice of what constitutes a ‘legitimate’ influence on spending need;
- The extent to which formula simplicity should be pursued at the expense of sensitivity to local needs;
- The extent to which known inadequacies in data are ignored, or the use of potentially useful data sources ruled out;
- The speed of implementation of a new formula.

A persistent feature of the political debates about funding formulae is the tension between the technical accuracy of formulae (intended to promote efficiency and equity) and a desire for simplicity (intended to promote political accountability) [13, 14]. Politicians and the media frequently, and reasonably, bemoan the complexity of many funding mechanisms, and its adverse impact on accountability. Indeed, the ability of the government to ‘make music’ with funding mechanisms attests to the accountability problem. At the same time, interest groups, politicians and localities also frequently complain that local ‘special circumstances’ are not accommodated within a funding mechanism. This lack of sensitivity to local spending needs usually implies a wish to search for more intricate formulae which would reflect the special circumstances more satisfactorily.

Many systems of formula funding have intricate damping mechanisms that reduce the year-on-year changes in an organization’s budget, notwithstanding the budgetary targets implied by the formula. Because of such mechanisms, the original recommendations of the NHS Resource Allocation Working Party (RAWP) [15] took 15 years to take full effect in the United Kingdom of Great Britain and Northern
Ireland. Similarly, Louis et al [16] describe widespread use of ‘hold harmless’ mechanisms in the United States of America federal formula funding for public services, under which states are guaranteed a certain minimum budgetary allocation, regardless of the outcome of the funding formulae. Such damping serves two principal purposes: reducing political turbulence and reducing local organizational problems associated with big changes in funding.

Although an all-powerful government could secure virtually any desired outcome from formula funding, it is more often the case that political consensus and compromise will be a central feature of the development of funding formulae. In many countries the formulaic outcome is determined by the national government (or some delegated authority) in the light of the representations made by many interest groups. There is often not an ineluctable scientific logic to the choice of funding mechanism. Rather, the development of formulae is a bargaining process in which the constituent parties seek a compromise on issues such as the nature of the funding formulae, the sources of data, the analytic methods to be used, and the legitimate indicators of spending needs to be used in the formula.

Another attraction of funding formulae to a government may be that they constrain either politicians or bureaucrats in the budgeting process. For example, if a national government makes a commitment to allocate expenditure on a large programme of health care only through a formula, it precludes politicians from seeking to influence the budgeting process by directing targeted expenditure at particular localities or groups of localities. The only mechanism for such influence is through the choice of funding formula, which is a more difficult process for politicians and lobbyists to subvert, particularly if the formula is intended to be scientifically based. In the same way, a government might circumscribe the freedom of its bureaucrats to manipulate the budgetary process, by insisting on the use of a specific formula, or at least requiring a particular process for the development of the formula.

4. The elements of capitation funding

This section outlines the rudimentary practical considerations that must be taken into account when developing a funding formula for health services. It first considers choices about the agencies in receipt of funds, and the scope of the health services to be covered by the funding mechanism. It goes on to discuss the basic data issues associated with counting the population, and varying the capitation payments in line with citizens’ health care spending needs (so-called risk adjustment). I briefly discuss the measure of costs to be used in any analysis. The section then concludes with some observations on the statistical issues that arise when seeking to develop empirical funding formulae.

4.1 Entities in receipt of funds

The first, most basic consideration in formula funding must be: to what local agencies are funds to be distributed? Often this is self-evident. However, there are circumstances when there are choices to be made about what organizations should be the target of funding, and to what level of responsibility to devolve funding. For example, in a federal state, there may exist municipalities, within larger local authorities, within provinces. A simple per capita funding mechanism may be adequate for distributing funds from central to provincial government, because the provinces have broadly similar demographic profiles, and because they have access to
a range of other funding sources with which to abate any inaccuracies in estimates of spending need. However, the same formula may be completely inadequate for use at the municipal level, because of the much greater heterogeneity of social and demographic circumstances amongst those organizations, and the more serious implications of any inaccuracy in the formula.

Financial risk is central to any consideration of the administrative level at which the finance is to be directed is the issue of. Broadly speaking, larger entities are more able to absorb such risk than smaller entities, because any inaccuracy in the formula can be absorbed across a large population. In the same way, agencies such as general local governments, responsible for a broad range of public services, can more readily accommodate inaccuracies in the health services formula by spreading the risk across a larger number of public services, or adjusting local taxes. In contrast, small, single purpose health agencies, reliant for all their income on formula funds, are placed at much greater risk, and so any formula will have to be correspondingly more accurate if serious inequities are to be avoided. In practice, of course, the choice of the entity is likely to be constrained and guided both by existing administrative structures, and by the practical data constraints set out below.

4.2 Services covered by the formula

A closely related issue is the set of health services for which funding is to be distributed. At one extreme, a formula might embrace all health services, broadly defined to include public health and other functions beyond direct health care. At the other extreme, it might relate only to one specific aspect of health services, such as HIV/AIDS. Again, this choice depends on both administrative and technical considerations.

In practice, many systems of formula funding have been disaggregated into a number of discrete services, for which different formulae are developed. For example, many health systems distinguish between the formulae used for distributing funds for mental illness, and those used for non-psychiatric care, because the ‘needs’ drivers for these two groups of services are quite distinct. A decision must first be made as to how much finance to allocate (at the national level) to each sector of care receiving separate formulaic consideration. The funds for each sector are then allocated using the sector-specific formula. The formulae that are developed in each sector should wherever possible use data that are relevant specifically to that sector.

The local agency’s total budget is the sum of its sector-specific budgets. Whether the local agency must adhere to its budgets within each sector is matter for national policy. In general, imposing such sectoral budgetary constraints is not good practice, as the freedom to vary spending between sectors is one of the ways that local agencies can manage the risk that any inaccuracies in the individual sector formulae might give rise to. Of course, the local agency should nevertheless be required to adhere to its aggregate budget.

In some cases, there may be a wish to ‘carve out’ certain health services from a general funding formula. For example, it may wish to omit (say) mental health services from local budgetary allocations, so that it can use other payment mechanisms for the sector that is omitted. The technical requirements for such carve-outs are broadly similar to those that are arise when developing sector-specific formulae. In effect, the government must know how much to subtract from each local
agency’s budget to reflect the omitted services, so must implicitly develop a formula for those services.

4.3 Counting the population

Capitation methods first require a verifiable count of population, disaggregated where necessary into demographic groups. Although the population count used in a capitation system is often uncontentious, it can give rise to difficulties when it relies on local reporting, as there are obvious incentives for local organizations to maximize the population on which their revenues are based. For example, for many years general medical practitioners in the United Kingdom of Great Britain and Northern Ireland received a large part of their income on the basis of unreliable estimates of the size of the population registered with the practice. There was widespread acknowledgement that the registered population sizes were inflated to very different extents in different practices, arising from factors such as delay in removing patients from the register when they died or changed provider, transient populations, and fraud [17].

Disaggregation of the population into demographic groups is a fundamental requirement of most capitation methods, because age and sex are often important predictors of expected health care spending that are readily available from routine data sources. Indeed for some services (such as maternity services) an accurate estimation of the relevant demographic group may be the single most important element of the formula. Ideally, the basic demographic disaggregation would be supplemented by epidemiological indicators (such as the presence of diabetes) yielding richer and more refined predictions of spending need. However, such epidemiological refinement is rarely feasible, as it is both administratively demanding (to collect universally) and vulnerable to misrepresentation by local agencies.

A final consideration is that, if a proportion of citizens ‘opt out’ of public health services and instead use private care, then efforts should be made to exclude them from the formula calculations. This principle can be rather hard to follow if citizens can exercise some choice as to whether or not to use public services, depending on the treatment they require and their personal preferences. In this case, some estimate must be made of the expected costs falling solely on public services. In effect, the possibility that an individual may use private care reduces the expected incidence of health care spending needs that are relevant to the public sector.

4.4 Risk adjustment

Although counts of the population can be problematic, it is usually the process known as risk adjustment that leads to most technical debate. The purpose of risk adjustment is to reflect variations in an individual’s needs for health service expenditure, if that individual is to secure access to the ‘standard’ package in line with entitlement. In principle, this entails modelling the determinants of (a) the probability of requiring services and (b) the intensity of use associated with that need. In practice these two issues are usually concatenated into a single estimate of expected expenditure on an individual.

Key issues to address are the choice of characteristics to include as ‘risk adjusters’, and the relative weight to attach to each factor. Different technical choices can lead to major changes in payment rates, and there is often little methodological guidance for
those seeking to design risk adjustment schemes. The methods described in the following section demonstrate a range of approaches to handling the risk adjustment process.

It is important to note that in many circumstances the range of satisfactory data available for risk adjustment purposes may be highly circumscribed. The first criterion in the design of risk adjustment will always be feasibility. For example, some obvious risk adjusters in health care would be measures of chronic health status, such as established hypertension. However, such information can rarely be reliably collected at reasonable cost, although some imaginative schemes have been tested, such as the use of routine prescribing data as a proxy for some chronic health condition [18].

Even where potentially useful data do exist, there is frequently a tension in the design of formulae between a desire to model expected expenditure accurately, and a desire to avoid perverse incentives. For example, the best predictor of an individual’s current health care expenditure is her previous history of expenditure and utilization. Such variables are often used in systems of competitive health insurance in order to model individual expenditure accurately (see Box 1) [19]. However, policy makers in other countries have sought to avoid the use of such data in the design of health service formulae, on the grounds that they may offer a perverse incentive for providers to increase provision in order to secure an increased capitation payment for the individual in the future.

**Box 1: Competitive Social Health Insurance**

Some countries that have traditionally relied on Social Health Insurance as the main source of finance have implemented systems of competition between insurers [5, 20]. Coverage is universal, and mandatory, and there is a national benefits package that all insurers must cover. However, citizens are able to enrol in the insurer of their choice. Examples include Belgium, Germany, Israel, the Netherlands, and Switzerland. A prerequisite for such systems is an accurate estimate of the expected annual costs of health care for the citizen. The role of the risk adjustment is then to ensure that insurers are fairly reimbursed for the risk profile of their insured population. In principle, this should ensure that all insurers are able to offer the statutory basic benefits package, thereby assuring equity between insurees and creating a ‘level playing field’ for the insurance market. In practice, competitive social insurance schemes have found it challenging to develop sufficiently accurate risk adjustment mechanisms. As well as compromising the efficient and fair operation of the market, any lack of precision gives insurers an incentive to ‘cream skim’ patients whom they judge to have lower expected expenditure than implied by the capitation formula, and to discourage enrolment by patients with high spending needs. This report is not directly concerned with systems of competitive social health insurance. However, their experience with developing individual level capitation payments is highly relevant to formula funding.

Any proposed risk adjuster must be reliably and consistently recorded across all recipients of funds. There will often be a need for a strong audit function to reassure all localities that payments are fair. A suggestion that some localities are manipulating information may be seriously corrosive. For this reason, a national government may often feel unable to use some otherwise suitable metrics as risk adjusters because they
cannot be satisfactorily verified. This is an important reason for the use in many funding mechanisms of area-wide data, such as that collected from periodic censuses of population, that obviate reliance on data provided by local agencies. Fraud is an ever-present danger when funding systems are based on data provided by the recipients of funds and has been a persistent concern in the USA Medicare scheme [21]. A fundamental constraint hampering many analytic endeavours is the extent to which the scope for misrepresentation may rule out the use of certain types of risk adjustment.

Risk adjusters should also be plausible, in the sense of being manifest drivers of expenditure. Moreover, a government will wish to use only factors that are legitimate drivers of expenditure, and will seek to avoid use of illegitimate factors [22]. Loosely speaking, legitimate drivers of expenditure are influences on the costs of delivering the required service that lie entirely outside the control of local organizations, and so can be used as risk adjusters. Examples might be some (but not necessarily all) user characteristics, and variations in local input prices. Illegitimate drivers of expenditure are influences on costs function that arise from the organizations’ own policy choices, and so should not be used as risk adjusters.

As noted in section 3.3, a persistent theme in the literature is the tension between parsimony in the use of risk adjusters and the need to model spending needs sensitively. Generally speaking, simple funding mechanisms are often preferred because they can be more readily understood and therefore promote accountability. However, there will often be an element of rough justice in a simple funding formula, so those local organizations that feel they are adversely affected by the choice of a simple mechanism will press for ‘refinement’, in the form of an increased number of risk adjusters and added complexity. Balancing simplicity and sensitivity of the funding mechanism is a key role for the national government.

In summary, numerous criteria for selecting risk adjustment characteristics have been indicated. For example, they should be:

- feasible, with low administrative cost;
- consistently, reliably, verifiably and universally recorded;
- not vulnerable to manipulation or fraud;
- legitimate predictors of expected health service expenditure;
- encourage efficient delivery of health services, and be free from perverse incentives;
- respect confidentiality requirements;
- parsimonious and plausible, thereby promoting transparency and accountability.

In practice, this often severely limits the choice of variables, as in many situations there exists only very restricted information on the characteristics of individuals or areas that conforms to such criteria.
4.5 Measuring costs

Capitation formulae seek to model the expected costs associated with each individual, assuming he or she has access to the chosen standard benefit package. Loosely speaking, the estimation of those costs poses three broad modelling challenges: estimating the probability of the need for services; estimating the ‘intensity’ of service use associated with that need; and estimating the costs of providing those services. As discussed above, much of the methodological debate surrounds the first two of those challenges. However, it is also important to model the third element – the costs to the local agency – with some care.

The choice of ‘dependent variable’ in any funding formula should reflect as closely as possible the expenditure requirements of the services for which funds are being distributed. In principle, this should seek to reflect the costs of providing the chosen standard benefits package. In practice, it is often possible to use only an approximation to this variable. For example, the initial RAWP report (see 5.2) used ‘bed days’ in each disease category as a proxy for costs.

The costs variable should reflect only the true costs to local agencies of delivering the service in question. For example, if agencies have to pay hospitals a standard tariff for certain procedures (such as a DRG payment) the dependent variable should reflect that tariff, and not any variations in local hospital prices. Likewise, if patients bear a proportion of the costs, the dependent variable should reflect costs net of any patient payments.

The dependent variable should also reflect any other cost sharing arrangements. For example, if the national government agrees to reimburse separately the local agency for costs on individual patients in excess of some threshold, the dependent variable should be constructed so as to reflect the expenditure limit. In short, the measure of costs used in any modelling work should reflect only relevant costs falling on the local agencies.

In the same vein, if there are inescapable variations in the input prices paid by localities, then it is important that the formula reflects this accurately. For example, labour and capital prices vary greatly within the United Kingdom, and ‘area cost adjustments’ have therefore been designed to reflect such variations [23]. They have resulted in adjustments to capitation payments that vary from 11% below the national average (in Cornwall) to 29% above the national average (in Westminster). It is important to note that the methodology to derive these adjustments should avoid use of specific public sector input prices. In particular, to use local public service wage rates as a basis for an area cost adjustment would allow localities to increase pay with impunity above efficient rates implied by local labour market conditions. Instead, estimates of local pay variations should be based on wages in comparable parts of the competitive local economy.

4.6 Statistical analysis

Fundamental to any type of formula funding is the information to be used as the basis for reimbursing local organizations. It is very rare to have available independent ‘engineering’ evidence of what levels of reimbursement each devolved entity should receive. Instead, in order to infer optimal reimbursement levels, analysts are highly dependent on historical expenditure patterns amongst the intended recipients of funds.
The reliance on past spending as a basis for current funding can give rise to profound philosophical and practical difficulties for the government.

These difficulties can be illustrated in their most extreme form when reimbursement of each organization is based solely on the historical spending of that organization alone – for example, by basing the organization’s current budget on expenditure last year. Such budget-setting rules are endemic within many bureaucracies, and were also a central feature of the Russian Federation's planning system, which tended to set production levels and budgets ‘from the achieved level’ [24]. They lead to many adverse consequences, and contain few incentives for efficiency or effectiveness.

As a result, health systems have sought out funding mechanisms that reduce the reliance on an organization’s own past expenditure in setting its future budgets. Instead, they have relied on some form of statistical analysis of general patterns of expenditure amongst all (or at least a substantial proportion) of the local agencies that will be in receipt of funds, along the lines described in the Section 5. The usual approach is to develop a statistical model that seeks to explain legitimate spending variations amongst individuals or localities.

One final issue of great importance relates to the variation in expenditure that is not captured by the chosen funding model. Even on those rare occasions when there are good data and the statistical model explains a high proportion of variation in health care expenditure, a considerable element of unexplained variation remains. The unexplained variation can be ascribed to two broad sources: omitted explanatory variables and random fluctuation.

Clearly, given the paucity of data usually available, there is an ever-present danger of variable omission, especially as some important but illegitimate determinants of expenditure (such as variations in clinical practice) may have to be omitted from the funding model. However, it is also important to recognize that there is a large element of unpredictable variation in the use of most health care that will always defy systematic modelling. This is particularly true under capitation funding, when both the probability of using a service and the expected costs of that use are subject to large random fluctuation. For example, a statistical analysis might indicate that – for a given package of care – a male aged 40-44 requires on average an expenditure of US$ 65 per annum. However, it would be absurd to expect every such individual to require that level of expenditure. Rather, a capitation payment of US$ 65 would offer an expected annual level of expenditure, around which there might exist substantial variation.

5. Country experiences with capitation funding

Capitation methods play the central role in most formula funding methodology found in the developed world. Numerous examples can be found in centrally planned health systems, devolved public sector systems, social insurance systems, and market systems. They include the methods used to allocate national funds to local health authorities in the United Kingdom (as initiated in the 1974 report of the RAWP) and the increasingly sophisticated methods used to equalize revenues in systems of competitive health insurance [5]. Properly designed, capitation methods give each local agency the means to deliver the national desired levels of output (in the form of the standard benefit package). This section describes how countries have in practice responded to the challenge of designing satisfactory funding formulae.
There are two broad approaches to risk adjustment in capitation methods: basing funding on the characteristics of the individuals for which the agency is responsible, and basing funding on the characteristics of the areas in which they live. The first offers a much more accurate indication of individual spending levels. However, it is very demanding of data. Therefore, area based capitation methods are often a more practical alternative. In practice the two methods are often blended, but we consider each in turn. The section concludes with a brief discussion of the role that case payment methods can sometimes play in formula funding, usually to supplement capitation methods.

5.1 Individual based capitation formulae

The most rudimentary formula for paying the local organization under capitation is to reimburse local agencies an equal amount per head of population, without regard to variations in the personal characteristics of that population. This approach can be justified when differences in expected service use between citizens in different social circumstances are not substantial, or when the local agencies in receipt of funds have only small differences in socio-demographic profile. It may also be justified if there is no further information with which to refine payments. However, for most health services such methods are manifestly crude, and will not satisfy national efficiency or equity objectives when there are substantial differences between the expected service use of different types of citizen.

A modest refinement to the crude capitation payment may be to confine the population ‘at risk’ to any obvious demographic stratum from which service users will be drawn. For example, financing for maternity services is clearly better distributed according to the estimated number of women of childbearing age than the total population.

More generally, however, different citizens exhibit very large variations in their propensity to use health services. For example, it is clearly the case that the expected expenditure on a patient aged 63 with diabetes is likely to be higher than that on a healthy male aged 24. If such variations in spending needs exist, and the payment system does not adequately compensate local agencies for the variations, clear inequalities will arise in the package of care that local agencies are able to deliver.

Capitation payments are therefore often further adjusted to account for more subtle variations in expected spending needs, using the principles of risk adjustment. For example, the expected expenditure requirement for health services is known to increase with age, so the estimates of expected expenditure needs should therefore usually be disaggregated by age. A typical schedule of risk adjustment according to age is shown in Figure 2. This age-based capitation payment rates are derived from a simple statistical analysis of actual utilization, and indicate the expected per annum health care expenditure on a citizen at certain ages. Note that – as in most capitation methods – the capitation payment represents the expected health care expenditure: effectively the product of the probability of the need for care arising, the intensity of that need, and the costs of meeting that need.

In principle, numerous approaches to risk adjustment of capitation schemes can be envisaged, ranging from such rudimentary age adjustments, to the extraordinarily ambitious schemes found in some health care systems [4]. Ideally, additional risk adjusters would include indicators of an individual’s morbidity status, especially the
presence of chronic disease or even smoking status. However, comprehensive and reliable collection of such data is usually infeasible, so capitation methods must rely on less direct indicators of the expected expenditure on an individual.

**Figure 2: Age-based capitation payments, English hospital and community services 2006-2007**

As new risk adjusters are added, so the capitation payments can be presented in the form of a contingency table. For example, most systems disaggregate by sex as well as age, leading to a two-dimensional matrix of capitation payments. A particularly rich example of individual risk adjustment is a matrix of capitation payments developed for Stockholm health care. This took advantage of a comprehensive register of Swedish citizens that records both personal characteristics and health care utilization in some detail. In such circumstances, the number of potential capitation payments to be estimated could be very large. For example, with indicators of age (eight groups), sex (2), social class (5), employment status (3), housing tenure (2) and marital status (2) the number of distinct capitation payments to be estimated in Stockholm might in principle have been $8 \times 2 \times 5 \times 3 \times 2 \times 2 = 2160$.

An important challenge is to reduce the potentially massive matrix of capitation payments to manageable proportions, by minimizing the number of risk adjusters used [25]. The abridged matrix of capitation payments recommended for Stockholm acute care, giving estimates of the expected monthly expenditure on health care, is summarized in Table 1. This shows, for acute services, the expected monthly expenditure on individual citizens, depending on their personal characteristics (age, sex, social class, employment status, housing tenure and marital status). It is derived from a quite rudimentary cross-tabulation of actual expenditure by those characteristics. However, to reduce the number of separate capitation payments, many of the cross-tabulated cells were amalgamated with neighbouring cells if they exhibited similar levels of spending, or were sparsely populated. A health authority’s budget was then calculated by multiplying the number of citizens in each cell by the associated capitation payment, and summing across the whole population.

Alternative approaches to collapsing the potentially vast number of capitation payment categories has been adopted by the USA Medicare system. This has acknowledged the unmanageable complexity of a full capitation matrix, and has
instead sought simply to quantify the additional contribution to expected expenditure brought about by a citizen’s risk factors. The starting point has been the citizen’s age and sex, and successive refinements to the capitation methods have seen the introduction of additional risk adjustment factors as data permit.

Table 1: The abridged Stockholm matrix of capitation payments for medical and surgical health care, Swedish krona per month, 1994

<table>
<thead>
<tr>
<th>Age</th>
<th>Owner occupier</th>
<th>Rented</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age &lt;1</td>
<td>7,200</td>
<td></td>
</tr>
<tr>
<td>Age 1-24</td>
<td>1,900</td>
<td>2,100</td>
</tr>
<tr>
<td>Age 25-64 cohabiting</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Higher non-manual</td>
<td>3,100</td>
<td>3,600</td>
</tr>
<tr>
<td>Other non-manual</td>
<td>3,700</td>
<td>4,300</td>
</tr>
<tr>
<td>Manual</td>
<td>4,000</td>
<td>4,400</td>
</tr>
<tr>
<td>Not employed</td>
<td>5,300</td>
<td>6,400</td>
</tr>
<tr>
<td>Age 25-64 living alone</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Higher non-manual</td>
<td>3,600</td>
<td>3,900</td>
</tr>
<tr>
<td>Other non-manual</td>
<td>3,600</td>
<td>4,200</td>
</tr>
<tr>
<td>Manual</td>
<td>3,900</td>
<td>4,600</td>
</tr>
<tr>
<td>Not employed</td>
<td>5,100</td>
<td>5,400</td>
</tr>
<tr>
<td>Age 65-84</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cohabiting</td>
<td>13,500</td>
<td>16,500</td>
</tr>
<tr>
<td>Living alone</td>
<td>15,400</td>
<td>18,200</td>
</tr>
<tr>
<td>Age &gt;84</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cohabiting</td>
<td>27,600</td>
<td>29,800</td>
</tr>
<tr>
<td>Living alone</td>
<td>24,200</td>
<td>29,400</td>
</tr>
</tbody>
</table>

Source: [26]

In the early manifestations of the Medicare approach, the only additional factors introduced were welfare status and whether or not the person live in a nursing home [27]. A Medicare matrix for hospital care for elderly people is shown in Table 2, with the average Medicare beneficiary given a weight of 1.0. The Medicaid category indicates that the individual qualifies for Medicaid support, the health care scheme for low income citizens. The weights reported in Table 2 were derived statistically from actual expenditure patterns amongst traditional fee-for-service Medicare beneficiaries. They show the amount of expenditure to be assigned to a citizen in each category, relative to a national average of 1.0. For example, for a non-nursing home resident, a working woman aged 68 would be assigned a capitation payment of 0.35, less than a quarter of that for a male Medicaid beneficiary aged 79, who attracted a weight of 1.50.

These capitation weights were used as the basis for payments to competitive health maintenance organizations. However, the methods could be used in any setting where there is a need to construct a total health care budget for defined populations, such as health authorities, local governments or social insurers. The key requirement is of course the universal availability of the necessary data.
The early Medicare methodology explained only about 1% of the variation in expenditure on individual Medicare enrollees, and was clearly inadequate in failing to adjust for the sickness level of the beneficiary. It also gave competing insurers a considerable incentive to cream skim healthier patients. The federal Balanced Budget Act of 1997 therefore introduced a much more aggressive risk adjustment mechanism, and from 2000 a scheme known as Medicare + Choice was introduced that included previous inpatient experience as a risk adjuster, in an attempt to model the beneficiary’s health status [28].

Table 2: Medicare 1999 demographic capitation weights

<table>
<thead>
<tr>
<th>Sex</th>
<th>Age</th>
<th>Nursing home</th>
<th>Medicaid</th>
<th>Non-Nursing home</th>
<th>Non-Medicaid</th>
<th>Working</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>65-69</td>
<td>1.75</td>
<td>1.15</td>
<td>0.65</td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>70-74</td>
<td>2.25</td>
<td>1.50</td>
<td>0.85</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td></td>
<td>75-79</td>
<td>2.25</td>
<td>1.95</td>
<td>1.05</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td></td>
<td>80-84</td>
<td>2.25</td>
<td>2.35</td>
<td>1.20</td>
<td>0.80</td>
<td></td>
</tr>
<tr>
<td></td>
<td>85+</td>
<td>2.25</td>
<td>2.60</td>
<td>1.35</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>65-69</td>
<td>1.45</td>
<td>0.80</td>
<td>0.55</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td></td>
<td>70-74</td>
<td>1.80</td>
<td>1.05</td>
<td>0.70</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td></td>
<td>75-79</td>
<td>2.10</td>
<td>1.45</td>
<td>0.85</td>
<td>0.55</td>
<td></td>
</tr>
<tr>
<td></td>
<td>80-84</td>
<td>2.10</td>
<td>1.70</td>
<td>1.05</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td></td>
<td>85+</td>
<td>2.10</td>
<td>2.10</td>
<td>1.20</td>
<td>0.80</td>
<td></td>
</tr>
</tbody>
</table>

The basis of the revised capitation payments was a risk factor table that included what was known as the Principal Inpatient Diagnostic Cost Group (PIP-DCG) of the patient [29]. This indicated the most severe category of inpatient diagnosis experienced by the citizen over a previous one-year period. Each citizen was allocated to one of the sixteen PIP-DCG categories of increasing severity, and capitations adjusted accordingly. Table 3 contains illustrative data, which are based on empirical analysis of expenditure on a 5 per cent sample of Medicare fee-for-service enrollees in 1996.

The national average factor was 1.0. An individual’s factor was built up additively. The base was determined by age and sex. To this was added a disability or Medicaid factor if appropriate. Then, if the beneficiary underwent hospital inpatient treatment in the base year, a further factor was added based on the highest PIP-DCG score of all qualifying inpatient spells. Thus a male aged 78 who qualifies for Medicaid and had an inpatient spell with PIP-DCG score 11 would have a total factor of 0.907 + 0.461 + 1.271 = 2.639. That is, the provider would receive 2.639 times the average capitation for accepting that beneficiary.

The PIP-DCG model was a clear improvement on its predecessor, in that it acknowledged some aspects of an enrollee’s sickness level. It offered a dramatic improvement in predictive power, explaining about 6.2% of the costs of traditional Medicare beneficiaries [30]. However, its reliance on inpatient diagnosis was a highly selective indicator of health status, and introduced potentially serious adverse incentives, by encouraging inpatient care in preference to potentially more cost-effective care in other settings. It was always seen as a transitional instrument, and the USA policy makers continued to seek out more suitable risk adjustment mechanisms.

The outcome was a scheme known as the hierarchical condition categories (HCC) model. The essence of the HCC approach was unchanged from the preceding PIP-
DCG model. Capitation payments were adjusted for the severity of a beneficiary’s sickness level, as indicated by previous health care diagnosis. However, the new diagnosis cost groups were based on both ambulatory and inpatient diagnoses, and are very much more refined than the PIP categories. The exposition below summarizes the comprehensive description given by Pope et al [31].

Table 3: Medicare + Choice 2000 risk adjustment capitation weights

<table>
<thead>
<tr>
<th>Age</th>
<th>Base</th>
<th>Previous disability</th>
<th>Medicaid</th>
<th>PIP-DCG Score</th>
<th>Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>0-34</td>
<td>0.367</td>
<td>0.125</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>35-44</td>
<td>0.380</td>
<td>0.283</td>
<td>5</td>
<td>0.375</td>
</tr>
<tr>
<td></td>
<td>45-54</td>
<td>0.487</td>
<td>0.370</td>
<td>6</td>
<td>0.458</td>
</tr>
<tr>
<td></td>
<td>55-59</td>
<td>0.615</td>
<td>0.397</td>
<td>7</td>
<td>0.697</td>
</tr>
<tr>
<td></td>
<td>60-64</td>
<td>0.760</td>
<td>0.418</td>
<td>8</td>
<td>0.822</td>
</tr>
<tr>
<td></td>
<td>65-69</td>
<td>0.541</td>
<td>0.415</td>
<td>9</td>
<td>0.915</td>
</tr>
<tr>
<td></td>
<td>70-74</td>
<td>0.705</td>
<td>0.398</td>
<td>10</td>
<td>1.170</td>
</tr>
<tr>
<td></td>
<td>75-79</td>
<td>0.907</td>
<td>0.334</td>
<td>11</td>
<td>1.271</td>
</tr>
<tr>
<td></td>
<td>80-84</td>
<td>1.077</td>
<td>0.287</td>
<td>12</td>
<td>1.662</td>
</tr>
<tr>
<td></td>
<td>85-89</td>
<td>1.258</td>
<td>0.237</td>
<td>14</td>
<td>2.000</td>
</tr>
<tr>
<td></td>
<td>90-94</td>
<td>1.376</td>
<td>0.189</td>
<td>16</td>
<td>2.438</td>
</tr>
<tr>
<td></td>
<td>95+</td>
<td>1.357</td>
<td>0.141</td>
<td>18</td>
<td>2.656</td>
</tr>
</tbody>
</table>

Female

<table>
<thead>
<tr>
<th>Age</th>
<th>Base</th>
<th>Previous disability</th>
<th>Medicaid</th>
<th>PIP-DCG Score</th>
<th>Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-34</td>
<td>0.362</td>
<td>0.192</td>
<td>-</td>
<td>23</td>
<td>3.823</td>
</tr>
<tr>
<td>35-44</td>
<td>0.403</td>
<td>-</td>
<td>0.312</td>
<td>26</td>
<td>4.375</td>
</tr>
<tr>
<td>45-54</td>
<td>0.526</td>
<td>-</td>
<td>0.367</td>
<td>29</td>
<td>5.189</td>
</tr>
<tr>
<td>55-59</td>
<td>0.643</td>
<td>-</td>
<td>0.397</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60-64</td>
<td>0.891</td>
<td>-</td>
<td>0.412</td>
<td></td>
<td></td>
</tr>
<tr>
<td>65-69</td>
<td>0.453</td>
<td>0.605</td>
<td>0.433</td>
<td></td>
<td></td>
</tr>
<tr>
<td>70-74</td>
<td>0.588</td>
<td>0.576</td>
<td>0.440</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75-79</td>
<td>0.747</td>
<td>0.519</td>
<td>0.454</td>
<td></td>
<td></td>
</tr>
<tr>
<td>80-84</td>
<td>0.918</td>
<td>0.415</td>
<td>0.423</td>
<td></td>
<td></td>
</tr>
<tr>
<td>85-89</td>
<td>1.096</td>
<td>0.313</td>
<td>0.327</td>
<td></td>
<td></td>
</tr>
<tr>
<td>90-94</td>
<td>1.162</td>
<td>0.232</td>
<td>0.231</td>
<td></td>
<td></td>
</tr>
<tr>
<td>95+</td>
<td>1.128</td>
<td>0.152</td>
<td>0.168</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The HCC researchers developed a hierarchy of diagnoses, so that for related diagnoses a patient was assigned to only the most serious category (so avoiding double counting). On the basis of previous research, the researchers first collapsed over 15,000 health care intervention codes into 70 HCCs that reflected clinically meaningful categories of diagnosis. The predictive power of the model diminishes as the number of HCCs is reduced, so a careful balance had to be struck between a manageable number of HCCs and good predictive power. Although related diagnoses within a HCC are counted only once towards the capitation payment, if a patient has more than one unrelated diagnosis, then both can contribute to the payment calculation.

The initial HCC model was estimated using ordinary least squares regression of preceding year (1999) diagnoses on individual costs of Medicare fee-for-service patients in 2000. The explanatory variables were 24 age/sex categories, four Medicaid status variables, a disability marker and the 70 HCC categories. Restrictions were imposed to ensure there were no perverse signs on coefficients. Statistical tests were undertaken to determine whether certain diagnoses interacted to create expenditure needs in excess of the simple sum of the associated adjustments. In the event, only six
interaction terms (such as the simultaneous presence of congestive heart failure and chronic obstructive pulmonary disease) were found to be necessary. Measured on a consistent basis, the use of the HCC model increased model $R^2$ from 6.2% under PIP to 11.2%.

The development of the HCC model sought to apply a number of criteria to guide model building [31]. These included the requirements that diagnostic categories should:

- be clinically meaningful;
- predict medical expenditure;
- be based on adequate sample sizes;
- be based on hierarchies that – within a disease process – record only the most serious diagnosis;
- encourage specific (rather than vague) clinical coding;
- not reward proliferation of diagnostic coding;
- not penalize the recording of additional diagnoses;
- be internally consistent;
- be exhaustive;
- exclude discretionary diagnoses.

These criteria are often in conflict with each other, and not always attainable. Moreover, other criteria (such as model parsimony and data collection costs) were also important considerations in the development of HCCs. The Medicare agency therefore had to come to a judgement on how to balance the criteria.

5.2 Area based capitation formulae

The individual based capitation methods discussed above are undoubtedly the most methodologically sound predictors of individual health care expenditure. However, in many health systems, individual capitation methods are infeasible because of their demands on information. Also, in order to secure accurate expenditure predictions, most individual based capitation schemes use some indicator of the individual’s previous health service utilization or expenditure. There is a concern that this approach may perversely lead to supplier-induced demand, as increased health care use may feed through into increased capitation payments for the affected populations in the future. Furthermore, schemes based solely on previous inpatient use (such as PIP-DCG) may distort some treatment inappropriately towards hospital settings.

Many health systems have therefore adopted ‘area-based’ empirical methods to act as the basis for capitation payments. In effect, these use the characteristics of the area in which a citizen lives as the basis for the capitation payment, rather than the individual’s own circumstances. Many such methods are implemented alongside a rudimentary individual-based capitation method (such as an age and sex risk
adjustment for the individual), and mirror the mixed individual and area based methods used by many commercial insurers of houses and vehicles.

In the United Kingdom, the celebrated report of the RAWP was one of the first to propose such methods, to allocate NHS funds to the regions of the United Kingdom [15]. The methods of the RAWP are summarized in Box 2. It is worth examining in some detail the steps followed by RAWP, in order to see how they correspond to the building blocks of capitation methods described in Section 4. The targets of the RAWP approach were the 14 regional health authorities, covering populations of about 4 million, and the services in question included hospital inpatient and outpatient care, and some community care, but not primary care or prescribing.

First, the health services were disaggregated into a small number of disease categories, corresponding to specialities, based loosely on International Classification of Diseases (ICD) chapter headings. The relevant national pool of expenditure to be distributed by the RAWP formula was disaggregated between specialities in proportion to the historical number of inpatient bed days used by the speciality (a proxy for the relative expenditure on the speciality). Capitation rates for each age and sex group (based on bed days used) were then constructed for each speciality.

In each speciality, the rudimentary age and sex capitation rates were then adjusted according to the region’s disease-specific Standardized Mortality Rate (SMR) for that speciality. So, if a region’s all age SMR for cancer was 113.7, it received 13.7% more funding per head than the national average for all cancer services. In effect, the SMR was used to raise (or lower) each age/sex capitation rate by a fixed proportion. RAWP therefore assumed there was a one-to-one relationship between a region’s disease specific SMR and its relative need for expenditure in that disease. These ‘needs’ calculations for each region resulted in numbers based on ‘bed days’ and were not budgets. Budgets were derived by dividing the total national budget available between regions in proportion to this notional bed days calculation. The total health services budget for each region was therefore the sum of its disease-specific budgets. There was no requirement that regions adhered to the disease-specific budgets.

Furthermore, there have historically been large variations between regions in the costs of delivering services (predominantly wage rates, but also capital costs). The final stage of the calculation was therefore to adjust each region’s budget by an index of relative local input costs. It is important to note that the calculation did not use NHS pay rates as the basis for this adjustment (this would have given each region an incentive to increase pay). Instead, the input price adjustment was based on the pay rates found in comparable services and industries.

The RAWP recommendations were implemented in 1976, and phased in gradually over a 15 year period. They took formula funding in the United Kingdom to a new level of intellectual coherence and sophistication, and have been highly influential internationally. However, they were hampered by a shortage of adequate data, and there was no empirical justification for the assumption of a one-to-one relationship between mortality rates and the need for health care expenditure. Therefore, when the opportunity emerged to develop a more empirically based formula, the RAWP system was superseded in 1991 by the small area methods described below [32]. The major innovation was to test a larger number of area-based measures of spending need, and
to estimate from empirical data the strength of the link between those measures and health care spending.

**Box 2: Summary of the Resource Allocation Working Party Recommendations**

The Resource Allocation Working Party first broke down health care into a small number of broad categories of specialities. The national level of spending in each speciality was assumed to be proportionate to the number of inpatient bed days used *per annum* by each speciality.

Within each speciality, the starting point for capitation estimates was the expected costs *per capita* after disaggregating the population by age and sex. The different expected health care costs of each demographic group were approximated by the national average *per capita* hospital bed utilization within the speciality under scrutiny.

These were in turn adjusted by the all ages SMR relevant to the speciality. The SMR is defined as the number of *observed* deaths in an area as a percentage of the *expected* deaths in the area, given its demographic profile. It was used by RAWP as an index of an area’s relative morbidity, and therefore as a proxy for medical need over and above demographic considerations.

Within a speciality, the index of relative need for care for each demographic category was therefore determined by applying the condition-specific SMR to the population of an area. This process generated a notional total use of inpatient bed days by the population in an area, assuming utilization conformed to the national average, after adjusting for local need, as indicated by the SMRs.

The total national pool of NHS finance was then allocated in proportion to these notional bed days. Algebraically, the equation can be represented as follows:

\[
RA_i = T \times \sum_j SMR_{ij} \times \left( \sum_k BEDS_{jk} \times POP_{ik} \right)
\]

where \(RA_i\) is the financial allocation to area \(i\); \(SMR_{ij}\) is the SMR of condition \(j\) in area \(i\); \(BEDS_{jk}\) is the average national number of bed days for condition \(j\) required by someone in age/sex group \(k\); \(POP_{ik}\) is the population in area \(i\) in age/sex group \(k\); and \(T\) is a constant, calculated such that the total local budgets summed to the total national NHS budget available.

The final stage of the calculation was to apply an ‘area cost adjustment’ to all budgets to reflect the large variations in input prices, especially pay, amongst the regions.

However, the move to a more empirical basis introduced new challenges. Rice and Smith [22] explain that a fundamental problem when using empirical methods is that variations in observed costs are likely to reflect a jumble of ‘legitimate’ influences (loosely speaking, variations in health care needs) and ‘illegitimate’ influences (variations in local policies, local clinical practice, etc.). This is a potential problem using the individual based empirical methods described above. However, it becomes especially acute if areas are used as the basis of empirical analysis. In particular, if the unit of observations used are the same administrative areas that will be in receipt of budgets, it becomes impossible to distinguish statistically between variations in
expenditure due to ‘true’ health needs and those arising from variations in local policies. As a result, researchers have strongly advocated the use of small areas (within larger administrative areas) as the unit of observation. Used carefully, such methodology offers a methodologically adequate statistical approach when individual methods are not possible. It can offer a compromise between the infeasible data demands of individual risk adjustment and the unsatisfactory analysis of data at the administrative area level.

A full exposition of the statistical methods underlying small area risk adjustment can be found elsewhere [33]. A relatively recent example is the Allocation of Resources to English Areas (AREA) study that has in recent years informed allocations to about 300 local health authorities in the United Kingdom, currently known as Primary Care Trusts (PCTs) [34]. The units of statistical analysis were about 8400 small areas covering the whole of the United Kingdom.

The dependent variable was calculated from the inpatient health care expenditure of the small area populations. A cost was attached to each episode, and the first stage was to develop a national set of age/sex capitation rates based on these data. These rates served two purposes. First, they could be used as basis for the first stage for allocating resources to health authorities, as described in the RAWP example above. Second, they became the basis for creating the measure of a small area’s utilization, standardized for age and sex. The standardization procedure calculated a small area’s utilization as the ratio of actual local costs to expected local costs, given the small area’s demographic profile. This standardized measure was then modelled as a function of supply and needs, using two stage least squares regression methods. Health authority dummies were included to capture health authority effects. Separate models were estimated for acute and psychiatric specialty groups.

Regression methods were then used to estimate a logarithmic model of health care utilization by the small areas, over and above any age and sex effects. Explanatory variables included indicators of local population health care needs (such as mortality rates, epidemiological indicators, and measures of deprivation) and other indicators of local health care supply (such as distance to hospital, waiting time). The econometric methods used were quite intricate, and beyond the scope of this paper. Note that multilevel modelling methods were used in order to allow for variations in spending and policies of the local health authority in which the small areas were located. The favoured model is reported in Table 4 (standard errors are omitted, all variables are significant at 5% level).

The chosen variables were selected on the basis of statistical significance, from the much larger set of potential explanatory variables. They can be considered in three groups. Group 1 variables are considered legitimate indicators of health care needs, and form the core of the formula. Although measured at a small area level, they reflect epidemiological or social factors that are widely acknowledged to be associated with individual health care spending needs.

Group 2 are ‘illegitimate’ supply variables, reflecting influences of local health services on utilization. These are included in the model in order to model small area costs satisfactorily, but they cannot be used in the funding formula (or the formula would perpetuate existing supply inequalities). When making expenditure predictions,
the formula therefore assigns all areas the national average for these variables, effectively ‘freezing’ their influence on allocations.

Group 3 were initially considered to be potential needs variables, of the sort selected in Group 1. However the statistical analysis yielded perverse signs – decreased expenditure associated with increased proportion of ethnic minorities and increased unemployment, other things equal. The researchers interpreted these results as indicating ‘unmet need’ amongst the associated population groups, and so also froze these variables in the funding formula, again to avoid perpetuating the inequity. The active part of the chosen formula therefore reflected the left hand side of Table 4, and is referred to as the relative ‘needs index’. When applied to the observed Group 1 variables for a small area in Table 4, it yields an estimate of the area’s expected expenditure levels, relative to the national average (keeping supply constant, and adjusting for unmet need).

**Table 4: Favoured AREA model, English resource allocation formula, hospital and community services.**

<table>
<thead>
<tr>
<th>Constant</th>
<th>-0.152</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Legitimate needs variables</strong></td>
<td>2. Supply variables</td>
</tr>
<tr>
<td>SMR under 75 years</td>
<td>0.070</td>
</tr>
<tr>
<td>Proportion of LBW babies born</td>
<td>0.013</td>
</tr>
<tr>
<td>Standardized birth ratio</td>
<td>0.108</td>
</tr>
<tr>
<td>Index of education deprivation</td>
<td>0.008</td>
</tr>
<tr>
<td>Aged 75+ living alone</td>
<td>0.026</td>
</tr>
<tr>
<td>Index of income deprivation</td>
<td>0.103</td>
</tr>
<tr>
<td>Nervous system morbidity index</td>
<td>0.225</td>
</tr>
<tr>
<td>Circulatory morbidity index</td>
<td>0.548</td>
</tr>
<tr>
<td>Musculoskeletal morbidity index</td>
<td>0.375</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Sutton *et al* [34]

In aggregation, these predictions were used to distribute funds to the 303 PCTs. Financial allocations to PCTs are calculated by first applying a conventional age adjustment, based on the data in Figure 1. The ‘needs index’ derived from Table 4 was then applied to the age-adjusted allocations. In effect, the age capitations for a PCT are raised or lowered by a fixed percentage, determined by the PCT’s needs index.
An example of the calculation for a single PCT (Bebington and Wirral) in 2004 is given in Table 5. The national age-specific capitation rates are shown in column (a). The funding formula yielded a needs index for the area of 0.8956 (compared to the national average of 1.0). Therefore, each capitation rate is multiple by that factor before multiplying by the local population in each age group (column (b)). The PCT budget is the sum of these needs-adjusted age-specific calculations (column (d)).

### Table 5: Calculation of budget for Bebington and Wirral PCT, 2004

<table>
<thead>
<tr>
<th>Age</th>
<th>National Age Capitation Payment (£)</th>
<th>Population</th>
<th>Age Adjusted Allocation (£)</th>
<th>Needs index = 0.8956</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-4</td>
<td>591.43</td>
<td>5,062</td>
<td>2,993,850</td>
<td>2,681,365</td>
</tr>
<tr>
<td>5-14</td>
<td>225.02</td>
<td>12,670</td>
<td>2,850,954</td>
<td>2,553,383</td>
</tr>
<tr>
<td>15-44</td>
<td>444.87</td>
<td>34,270</td>
<td>15,245,542</td>
<td>13,654,277</td>
</tr>
<tr>
<td>45-64</td>
<td>531.76</td>
<td>27,803</td>
<td>14,784,635</td>
<td>13,241,478</td>
</tr>
<tr>
<td>65-74</td>
<td>966.18</td>
<td>11,200</td>
<td>10,821,163</td>
<td>9,691,696</td>
</tr>
<tr>
<td>75-84</td>
<td>1,583.87</td>
<td>7,640</td>
<td>12,100,901</td>
<td>10,837,861</td>
</tr>
<tr>
<td>85+</td>
<td>2,357.64</td>
<td>2,768</td>
<td>6,526,778</td>
<td>5,845,541</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>101,414</strong></td>
<td><strong>65,323,823</strong></td>
<td><strong>58,505,602</strong></td>
<td></td>
</tr>
</tbody>
</table>


The consequences of the AREA formula are summarized in Figure 3. The first chart ranks the 303 PCTs in order of gains (and losses) from the initial age adjustment. It indicates at the extremes a *per capita* gain of 25% (Bexhill and Rother PCT) and a *per capita* loss of 15% (Tower Hamlets PCT), relative to an allocation with no age adjustment. Of the total budget of £49.8 billion, the age adjustment redirects £1.2 billion from PCTs with younger populations to those with older populations. The needs adjustment is more redistributive, giving rise to a maximum *per capita* gain of 52% (North Manchester PCT) and a maximum *per capita* loss of 29% relative to an allocation with no needs adjustment. The needs adjustment redirects £4.1 billion from PCTs with low needs populations to those with high needs.

**Figure 3:** Percentage redistribution of finance between English Primary Care Trusts 2005/2006 caused by (a) age adjustment (b) needs index.
5.3 Case payments: reimbursing according to outputs

Although capitation methods form the core of most formula funding systems, there are occasions when it is appropriate to augment or replace a capitation formula with a case payment approach. The usual objectives of case payment methods are (a) to reimburse local agencies fairly in line with the number of service users (b) to encourage provision of more services, and (c) to encourage efficiency improvements so as to reduce unit costs. For example, Norwegian health authorities are funded 40% by capitation and 60% by DRG payments. The reason for introducing the case payment element was to stimulate inpatient activity, and thereby address inpatient waiting times.

The usual approach to case payment is to reimburse according to some measure of the expected expenditure on a service user. For example, the case payment might be based on an estimate of the national average cost per case, perhaps with some reduction for assumed efficiency gains. Or it might be based on (say) the lower 25% percentile of costs, assuming these represent a challenging but feasible cost target for the majority of service providers. Any surplus of the case payment over the local cost is usually retained by the local agency. When users are reasonably homogenous, such as births without complications, it may be feasible to use a single case payment for all users. However, for most health services, service users are usually not homogeneous and therefore risk adjustment will be needed. This seeks to vary payments according to the dependency of users, analogous to the methods used in capitation methods.

For example, it is clearly the case that the expected expenditure on a patient requiring an appendectomy is much in excess of that on a patient requiring treatment for minor abrasions, so a higher case payment should be made.

The most celebrated case payment approach is the system of DRGs used in many health systems, under which the case payment varies according to the diagnosis of the patient [35]. DRGs were introduced in 1983 as the basis for reimbursing providers under the USA Medicare system. Their use led to manifest cost savings and efficiency improvements compared to the previous ‘fee-for-service’ reimbursement system [36]. Medicare’s DRG system was the starting point for other DRG systems inside the United States of America. It also secured great international interest, and health systems throughout the developed world have adopted their own versions of the DRG payment mechanism [3].

The first purpose of DRGs is to offer an accurate assessment of the costs of treating a given patient, in the light of observable and measurable patient characteristics, especially the diagnosis – but, to a varying degree, also the interventions chosen. Indeed, the antecedents of today’s DRG systems were developed as a method of adjusting measures of the performance of providers for the different mixes of patients they treat, rather than a mechanism for reimbursing providers [35].

In this role, the main challenge is a technical one of ensuring that the risk adjustment process is unbiased and accurate. This requires careful decisions on the design of the DRG system, such as the hierarchy and algorithms used to classify patients into a limited number of groups. In principle, patients within one group should have homogeneous costs. Moreover, cases allocated to one group should form a clinically distinguishable entity based on main diagnosis, severity, co-morbidity and/or treatment performed. But what exactly is homogeneous? When are costs or clinical
diagnosis so different that a group should be split? In practice, most DRG-type systems use several hundred different categories of patient, based on criteria such as clinical diagnoses, procedures undertaken, length of hospital stay and age.

In addition to the intended incentives of stimulating activity and reducing unit costs, case payment systems can introduce unintended and adverse incentives. If the risk adjustment mechanism is inadequate, and a case payment system does not reimburse providers fairly for the higher costs of more intensive users, providers may seek to ‘cream skim’ only low cost users. More intensive users may then find it difficult to secure access to services because their expected costs exceed reimbursement levels. The potential for cream-skimming has been the prime motivation for progressive refinements to risk adjustment methods.

Note that there are often considerable variations in patient costs even within a DRG, so there may still be an incentive to cream skim within the group. Generally, the local provider always has a clearer picture of the likely costs of a specific user than any organizational purchaser, whether local or national. This information asymmetry sometimes leads to pressures to define a finer gradation of user types, for example by increasing the number of DRGs. However – even if this is technically feasible – as the gradation becomes finer, so the payment becomes closer to reimbursing providers according to actual costs. The incentive to reduce costs becomes diluted. There is therefore a fine balance to be struck between the coarseness of the case payment categories, the incentives for cost-reduction, and the incentives for cream-skimming.

Under case payments, local organizations usually have an incentive to reduce unit costs, subject to satisfying any quality requirements. Yet they also have an incentive to ensure that they secure the maximum possible case payment for each service user. Where risk adjustment is used, it is often the case that the level of case payment depends to some extent on the type of service delivered. This may encourage providers to ‘gold plate’ services in order to secure a higher case payment (where the extra case payment outweighs the additional costs of service delivery).

For example, under many DRG systems, a patient’s case payment category may increase if there are ‘complications’ associated with the treatment, in some cases offering physicians an incentive to over-treat patients in order to secure a higher DRG payment. The phenomenon of ‘DRG creep’ has been a source of widespread concern. Even if provider treatment patterns do not change, the DRG system offers incentives to ensure that all possible indicators of case severity are properly recorded, and software packages have been developed to help medical coders maximize DRG revenue, giving rise to the phenomenon known as ‘upcoding’.

Case payment methods rely fundamentally on quantification of the volume of outputs, in the form of the number of service users and (where appropriate) their risk characteristics. Governments should always be alert to the dangers of reliance on counts of service users as the basis for reimbursement for a number of reasons, such as:

- The counts often rely on the local organization’s own information sources and so may be vulnerable to fraud and difficult to verify.
• Using the actual number of service users as a basis for reimbursement may encourage the local organization to stimulate demand for health care rather than invest in preventative measures or otherwise manage local demand.

• The purchaser of services does not know in advance the total number of service users, and so may not be able to control aggregate expenditure satisfactorily.

The use of case payments therefore generates fundamental requirements to check on the validity of the local data, and the appropriateness of the services being delivered. Such audit requirements will in some circumstances be costly and administratively demanding.

Furthermore, although encouraging increases in outputs, case payment methods rarely consider the ultimate outcomes to patients, and local agencies therefore often have an incentive to stint on quality if no quality constraints are placed on them. For example, rudimentary DRG systems take no account of patient outcome (perhaps in the form of mortality). There are now experiments under way to see whether and how case payments can be altered according to some measure of quality, but these raise numerous technical and incentives issues, and are so far in their infancy.

The principal reasons why capitation methods prove attractive are (a) they encourage appropriate preventative measures for services where this is an important objective, (b) they remove the incentive for supplier induced demand, which can be a problem under case payment (c) they reduce reliance on output data provided by local providers, and (d) they can be aligned with national equity objectives. Case payments can nevertheless be a useful adjunct to certain aspects of the formula funding regime. They remove from local agencies the risk associated with the incidence of disease, which can be an important consideration for rare and costly conditions, or where local agencies have high levels of financial risk, perhaps because they are small in size. However, if used uncritically can give rise to important unintended consequences, and should therefore be integrated with care into any formula funding mechanism.

6. Conclusions

Under the influence of what has become known as the ‘new’ public management, health systems are seeking to nurture improved responsiveness to users’ needs, heightened attention to the outcomes secured, and improved efficiency. In pursuing this agenda, some governments have implemented fundamental reforms to their health services [8], such as:

• decentralizing powers and authority, alongside reduced reliance on direct central command;

• viewing local agencies as purchasers and regulators of health services, rather than direct providers;

• introducing increased diversity and competition into the provision of services;

• seeking to empower patients by improving information and enhancing provider choice;

• increasing efforts to measure and report performance.
Satisfactory methods of funding are an essential requirement for the success of many of these innovations, and the methods described in this paper should therefore play a central role in the reform process. Yet hitherto there has often been little discussion of how funding mechanisms should be designed in order to support health system reforms. This paper has examined experience in high income countries. However, it is important to emphasize that there have been concerted efforts to implement formula funding in low and middle income settings, such as Chile, Columbia, Ghana, South Africa, and Zambia [37, 38].

This section briefly examines the practical steps that are required for successful implementation of formula funding. The paper first rehearses the institutional arrangements that must be in place before formula funding can be considered, and the main tasks that are needed to make formula funding a reality. I then conclude by discussing the priorities for further improvement, once formula funding has been implemented.

6.1 Making formula funding a reality

Section 2.2 highlighted three institutional requirements for the use of formula funding to be remotely feasible: the need for devolution of responsibility for health services to local agencies; the need for adequate data; and the need for incentives to adhere to budgets. Here I consider each in turn.

First, the organization and purchasing of health services must be to some extent devolved, with appropriate systems of governance and accountability in place. The national government must be able to set clear objectives for the health system, and decide on aggregate levels of funding and the level of services that citizens can expect to be delivered. The agencies in receipt of devolved funding must then be responsible for a clearly defined population, and be managerially capable of managing the funds they receive. Furthermore, they must understand the scope of health services that they are expected to deliver, and be capable of holding to account the local providers from whom they purchase services.

These are demanding requirements. In particular, there is limited international experience of successfully implementing “strategic purchasing” of health services by local agencies. In the short term, it may be that all that can be expected is that local agencies make contracts with providers that are explicit and enforceable, that all services in the specified health basket are delivered, and that there are checks on the efficiency and probity with which funds are used. Devolved agencies may need considerable guidance and support in developing these skills. In short, if the benefits of decentralization are to be secured, it is not enough merely to devolve funding: the devolved entities also need to have the power and capacity to use the funds in the intended manner.

Second, there must be adequate data, available on a consistent basis across all local agencies, to act as the basis for the formula funding. As discussed in section 4.4, the data that serve as the basis for distributing funds must satisfy numerous criteria. These include being:

- feasible, with low administrative cost;
- consistently, reliably, verifiably and universally recorded across all agencies;
not vulnerable to manipulation or fraud;

- legitimate predictors of expected health service expenditure;

- encourage efficient delivery of health services, and be free from perverse incentives;

- respect confidentiality requirements;

- parsimonious and plausible, thereby promoting transparency and accountability.

These too are challenging requirements. Yet without consistent, reliable data it will never be feasible to implement formula funding, because the devolved agencies cannot be assured that funding choices are systematic and even-handed. In many situations it may be possible to rely on quite rudimentary counts of population associated with each agency, possibly disaggregated by demographic characteristics. Yet even in the absence of any other data, the use of such population counts as a basis for capitation payments will represent a starting point on which future refinements of the formula can build. The RAWP experience in the United Kingdom shows that with even modest additional information resources (standardized mortality rates) it is possible progressively to refine the funding formulae. Then, as new data become available, the experience with limited data can lead to more ambitious schemes.

There are a number of technical requirements at the national level for formula funding to be successful. Of course, the development of formulae under severe data constraints is a demanding analytic task, and there will usually be considerable scope for sharing experience with countries that have successfully implemented formulae under similar conditions. However, development of formulae cannot be left just to the technical domain. The national funder must also make important policy choices, such as the speed with which budgets will be expected to move towards the spending targets implied by any formulae. The funder must also be prepared to announce the chosen budgets for devolved agencies in advance of the funding period, with no provision for changes over that period. Furthermore, there will be a continual need to review the functioning of the formula, and to revise when appropriate.

Third, there must exist an incentive for local agencies to adhere to the budgets implied by the formula. Such financial discipline has traditionally been hard to impose within health systems. However, it is a prerequisite of decentralization that local agencies take responsibility for their budgets, and bear the consequences of overspending. Such budgetary discipline is essential if principles of equal treatment and fairness are to be respected. In the early years of implementation it may be appropriate to build some leeway into the budgetary regime, in the form of financial risk sharing arrangements. For example, the sanctions associated with early experience of ‘fundholding’ in the United Kingdom were generally very soft. However, as experience grows, so the sanctions can be made progressively harder.

This paper has sought to set out the practical tasks involved in implementing formula funding. Many of the refinements I have described will in many countries be infeasible in the short run. However, Box 3 summarizes the key practical steps that are needed to make even rudimentary and effective formula funding a reality. This extensive nature of the list indicates that implementation will always require
considerable commitment in terms of data sources, analytic capacity and political leadership.

Box 3: Key steps to making formula funding a reality

1. **Identifying the objectives of the formula funding regime.** These are likely to include efficiency, equity and political considerations. However, underlying most systems is the desire to secure a fair distribution of limited health care resources.

2. **Identifying the entities in receipt of funds.** These can vary from huge provinces to small medical practices. However, the discussion has suggested that the technical challenges become more severe as the entities become smaller in size and scope, because smaller entities will generally be exposed to more financial risk than their larger counterparts.

3. **Identifying the health services covered by the formula.** In principle this stage might entail setting out an explicit health basket, or standard level of services, that the entities are expected to provide. However, most systems of formula funding implicitly assume some ‘national average’ level of service.

4. **Counting the population.** All capitation systems require a reasonably reliable count of population, preferably disaggregated by demographic stratum, so that variations in health care expenditure needs can be more accurately modelled.

5. **Adjusting the population for variations in relative expenditure needs.** If disaggregated demographic data are available, or there exist other individual indicators of individual needs (such as disability status), then these might be used as a starting point for risk adjustment. However, in the early stages of formula funding it may be necessary to adjust demographic using quite crude population data, such as mortality rates. Any data used will have to be consistently and reliably collected across all localities in receipt of funds.

6. **Attaching costs to the risk adjustment process.** Cost estimates should of course be realistic, but in aggregate must adhere to the national spending budget. Also, any major variations in input prices between parts of the country should be modelled.

7. **Deciding how fast local agencies should be expected to converge towards spending targets.** In practice, in the early years of implementation actual levels of spending of local agencies are likely to vary greatly from the amounts implied by the funding formula. How long they should be expected to take to adjust to the new spending targets is mainly a political judgement.

8. **Putting in place relevant incentives and cost sharing mechanisms.** Sanctions for overspending should be carefully designed, and any relevant cost sharing mechanisms for individual patients or treatments put in place. These might take the form of user charges, cost sharing with other local agencies, or subsidies from the central risk pool.

9. **Measuring performance.** Arrangements should be put in place to ensure that local agencies are providing the required health services, at an acceptable level of quality.

10. **Monitoring and refining the operation of the funding formula.** The funding mechanism should be kept under constant review. Data should be updated as appropriate, and where necessary new data sources introduced and incorporated.
6.2 Developing formula funding in the future

The preceding section indicates the minimal prerequisites for any formula funding regime to be made operational. However, the national funder should also seek to enhance the funding system in order to make the formula funding more effective. Here I highlight just three actions that will promote such improvements: linking the funding mechanism to the performance standards required of the devolved agencies; understanding more fully the incentives inherent in formula funding; and correcting the information weaknesses in the funding mechanism.

Many health systems continue to design payment mechanisms in isolation from performance standards. Yet, as explained above, there is an inextricable link between the performance standards agencies are expected to achieve and the funding they receive. A major benefit of formula funding should be that it enables performance comparisons to be made between local agencies on a level playing field. In effect, as discussed above, each agency should in principle be given the funds needed to secure a specified performance standard. If the formula is accurate, performance reporting can then operate effectively, because any variations in achieved performance should be directly attributable to the actions of local management, rather than a shortage of funding or adverse environmental circumstances.

There are some very strong incentives inherent in pure capitation funding systems. These include reducing health care activity, increased preventative measures, cream skimming healthier patients (within a risk group) and quality skimping. Some of these are intended and virtuous. Others are unintended and adverse. In general, the health system will wish to reinforce the intended incentives and abate some of the stronger adverse incentives. To some extent this can be achieved by adjusting the payment mechanism. Possibilities include refining the risk adjustment mechanism (to reduce the gains from cream skimming) and using some activity-based funding alongside pure capitation payments (in order to moderate the stark incentive to reduce activity under capitation).

However, the payment mechanism will usually have to be augmented with other regulatory instruments in order to ensure that desired outcomes are secured. These might include periodic inspection of local services, performance reporting, promoting patient choice of provider, and implementing managerial incentives alongside formula funding.

Finally, it is self-evident that improved information plays a central role in all funding mechanisms, both to serve as a basis for calculating funding levels and as a means of checking on performance standards. It can also serve many other purposes for improving health system performance. However, for many health systems, a key issue is the limited capacity to collect and verify data, and it may often be politically difficult to devote significant resources to data enhancement. Yet successful decentralization depends on improved data collection and coordination efforts. A major role for the national government is therefore to ensure that an appropriate information strategy is in place data, setting realistic priorities for the specification, collection, analysis and audit of data sources.

By aligning funding mechanisms with central objectives, formula funding offers a powerful tool for making strategic objectives operational. It is in any case almost certainly an improvement on alternative funding mechanisms in most circumstances.
However, implementation of formula funding has many perils, and there is a great need for attention to detail if it is to be achieved successfully. There are many challenges, in the form *inter alia* of:

- Articulating objectives for the health service under consideration;
- Identifying appropriate data sources;
- Identifying what influences on expenditure can be considered legitimate;
- Determining the preferred systems of formula funding, or mix of systems;
- Estimating the parameters of the funding formula;
- Putting in place appropriate audit and inspection regimes;
- Monitoring and reviewing the operation of the funding formula.

Furthermore, formula funding cannot on its own secure adherence to central policy objectives. Numerous other health system instruments must be aligned with funding mechanisms if they are to be effective. Examples include policies on regulation, provider markets, performance reporting and good practice guidelines. However, if these instruments are deployed carefully, and are made to articulate with each other, they offer the potential for enormous gains in the efficiency and effectiveness and equity of the health system.

**References**


