Risk and the GP budget holder

Stephen Martin, Nigel Rice, Peter C. Smith

DISCUSSION PAPER 153
Risk and the GP Budget Holder

Stephen Martin
Department of Economics & Related Studies, University of York

Nigel Rice
Centre for Health Economics, University of York

Peter C. Smith
Department of Economics & Related Studies, University of York

April 1997
CENTRE FOR HEALTH ECONOMICS DISCUSSION PAPER SERIES

The Discussion Paper (DP) Series was established in 1984 to make research available quickly to a broad readership drawn from the academic community, professional health care staff and policy makers, both within the UK and internationally. Material published in DP form ranges from work in progress, to edited versions of completed project reports, from papers intended ultimately for journal publication, to those which seek to stimulate immediate debate. Hence the variation in length and complexity of DPs.

In the past, DPs have been subject to a minimum of editorial revision, it being principally seen as the responsibility of individual authors to ensure that appropriate standards were met. Since June 1996, a more formal system of internal peer-review has been introduced to consolidate the high levels of quality that have been achieved in the past. That system involves the refereeing of DPs in draft form, with any recommended revisions being verified independently by a senior member of staff.

DPs are not a substitute for publication in peer-reviewed journals. It is to be hoped that journal editors will recognise the nature of the DP series, and accept that this does not compromise the status of those papers published by the Centre which may later be submitted for consideration in more formal publications. Comments and suggestions on DPs encouraged and should be sent directly to the authors.

ACKNOWLEDGEMENTS

The study was in part funded by the NHS Executive. Thanks are due to Mark Groom and Bill Barron. However, the views expressed are those of the authors. The paper was written while Smith was at the Department of Management, University of St Andrews.

THE AUTHORS

Stephen Martin is a Research Fellow based in the Centre for Defence Economics which is located within the Department of Economics & Related Studies, University of York. Nigel Rice is a Research Fellow in the Centre for Health Economics, University of York. Peter C. Smith is a Professor in the Department of Economics & Related Studies, University of York.

FURTHER COPIES

Further copies of this document are available (at price £5.00 to cover the cost of publication, postage and packing) from:

The Publications Office
Centre for Health Economics
University of York
York YO1 5DD

Please make cheques payable to the University of York. Details of other papers can be obtained from the same address or telephone York (01904) 433648 or 433718.
ABSTRACT

For most individuals, the use made of health care in a given year is determined principally by unpredictable random incidents. Of course, some individuals have a predictably higher predisposition to illness than others. However, the general consensus is that only a fraction of individual variability in health care costs can be predicted. The purpose of this paper is to explore the implications of this inherent randomness for budget setting for GP purchasers. The paper argues that variability in utilization in the NHS is very high; that no formula will ever completely capture that variability, even for large populations; that the problem of variability is likely to be very acute for a GP practice; and that health authorities and GP budget holders will therefore need to adopt a range of strategies to manage the variability.
INTRODUCTION

The 1991 NHS reforms introduced a system of GP fundholding, whereby part of the Hospital and Community Health Services (HCHS) budget was devolved from Health Authorities to those general practices which choose to become fundholders. Fundholders use their budget to purchase health care from a range of providers within the NHS “internal market”. Since its inception in 1991, the scope of the standard fundholding scheme has been expanded to include most elective procedures as well as prescribing, and experiments are underway to assess the feasibility of “total” fundholding, under which virtually all health care is encompassed.1

Before the advent of fundholding, general practitioners referred patients to hospitals and prescribed medicines and other forms of health care without any clear idea of the cost implications. The central purpose of fundholding is to make general practitioners aware of all of the cost implications of their clinical decisions. The hope is that the receipt of a budget will enable GPs to make efficient decisions about (a) whether to treat a patient and (b) how to treat that patient. The Conservative government claims it forms a crucial element in the move towards a primary care-led NHS.2

The principle of fundholding has not enjoyed universal political approval, and it is not clear whether in its original form the scheme would survive a change of government. However, what is almost certainly the case is that, in one form or another, general practitioners can expect to continue to receive budgets within which they are expected to operate. The political uncertainty relates merely to the breadth of services encompassed by the budgets and the sanctions available if budgets are exceeded.

Fundholding has not received a great deal of research attention. Indeed, this neglect mirrors the general management literature, in which the impact of budgets on organizational performance has been the subject of only a modest research effort.3,4 The first major study of budgets in businesses found "astounding" hostility between budget supervisors and budget holders.5 This led early commentators to advocate more dialogue between the two parties,
and more participation in the budget-setting process on the part of budget holders. However, Macintosh argues that it is naive to assume that such dialogue will in itself improve relations. Much more important, in his view, is that budget supervisors should investigate and act on variations from budgets in an intelligent and flexible manner. In general, researchers have unhelpfully tended to report conflicting findings on the role of supervisory style on corporate budgetary performance. This has given rise to the development of a "contingency" approach towards budgetary policy. However, one clear finding is that a very sensitive style is needed when environmental uncertainty is high.

In the NHS, definitive evaluation of the consequences of the fundholding scheme remains elusive. However, the number of general practices included has steadily increased, so that in 1996 about half the population is registered with a fundholding general practice. Fundholding now accounts for 15% of HCHS expenditure. Furthermore, some Health Authorities are now creating “indicative” budgets for non-fundholders, along the lines suggested by the National Audit Office, suggesting that the principle of setting budgets for general practice purchasing has become securely established within the NHS. Thus, whatever the future of the fundholding scheme, it is likely that most, if not all, GP practices will eventually receive a budget which indicates their “fair share” of local HCHS resources.

How are those local resources determined? At present the NHS Executive makes HCHS resource allocations to Health Authorities using a weighted capitation formula. This uses data on age and various social characteristics to model expected predictable variations in utilization around the country. Thus Health Authority purchasers receive a national average capitation for each citizen resident within their boundaries, adjusted to account for local demography, needs and costs. The Executive now recommends that Health Authorities in turn should use similar methods to set GP fundholding budgets. The arguments for using a formula are compelling. In summary, they are:
first and foremost it promotes equity: only by giving GPs equal budgets for equal need will there be any chance of securing the equity objectives of the NHS;

• a formula makes the construction of GP budgets systematic and transparent;

• an appropriately designed formula can make allocations independent of past GP behaviour, and therefore remove incentives for perverse referral behaviour;

• the use of a generally accepted formula will reduce political pressures;

• the consequent reduction in levels of conflict might thereby reduce managerial costs.

A formula must use expected levels of utilization as a basis for setting budgets. However, the use of averages implicit in a formula disguises wide variability in actual health care utilization between individuals. In any one year, the incidence of illness and the associated costs of health care vary considerably, even within a single risk class (as perhaps defined by age and needs). While many individuals may suffer no episodes of illness of any significance, some will experience serious illnesses requiring substantial NHS expenditure. Thus giving a national average NHS "budget" for each citizen would clearly be nonsensical. Some citizens would make no use of such budgets, while others would require funds vastly in excess of their budgets. It is only when individual capitations are aggregated into populations that budgets begin to make some sense.

Newhouse et al present the following model of individual annual health care expenditure:

\[
\text{Expenditure} = b.X_i + u_i + e_{it}
\]

where \(b\) is a universal response to known risk factors, \(X_i\) is the vector of risk factors exhibited by individual \(i\), \(u_i\) is a time-invariant component of variance associated with individual \(i\) which is independent of the risk factors, and \(e_{it}\) is a person-specific time-varying component of variance.\(^{17}\) The component \(b.X_i\) can be thought of as the part of the total variance in expenditure which can be predicted by a capitation formula. Component \(u_i\) represents the
additional part of variance which could in principle be predicted if it were possible to incorporate all potential risk factors of an individual into a formula. The component $e_{it}$ is the random part, which is independent of any patient risk factors, known or unknown. This model implicitly subsumes variations due to clinical practice into the random element $e_{it}$. Newhouse *et al* estimated that this random component is likely to account for at least 85.5% of the variance in expenditure on any one individual.

This insight can be adapted therefore to partition variability in health care expenditure into four components:

1. an element that is due to individual characteristics - such as age - which are captured by whatever formula is used to allocate funds to the GP;
2. an element that is due to patient characteristics which are *not* captured by the formula, such as - for example - the presence of diabetes;
3. an element that is due to clinical practice;
4. an element that is totally random, caused by the unpredictable incidence and severity of illness.

Components (1) and (2) can be considered the *systematic* element of health care utilization, in the sense that they could in principle be predicted. In practice, the chosen formula compensates GP practices for the first aspect of this systematic element, but not for the second. The quality of any capitation formula should in principle be judged by the extent to which it maximizes (1) and minimizes (2). By contrast, the formula should not seek to capture variations in utilization which are purely due to variations in clinical practice (3). Instead, it should be based on some "standard" set of practices (for example, the current national formula uses national average costs as its basis). And of course no formula can compensate for the random element (4). The hope is that - as the population to which the formula is applied becomes larger - so the "uncompensated" element (2)+(3)+(4) becomes smaller.
This analysis implies there are likely to be large unpredictable deviations from any individual capitation formula caused by unsystematic variation in expenditure. However, if - as in the NHS - individual budgets are aggregated into populations, then the variability in average spending requirements diminishes. As the size of the population group increases, so "low" spenders are combined with "high" spenders, and the percentage variation in actual per capita spending from budget steadily declines. When the populations are very large, average spending requirements can be predicted with some confidence, and percentage variations from per capita annual budgets become trivial.

THE YORK ACUTE SECTOR INDEX

Although urged to use some sort of formula for allocating resources to GP fundholders, Health Authorities are free to adapt any system which is suitable for local circumstances. Nevertheless, many are making use of an amended version of the national acute sector index developed by the University of York. The York Index was based on an empirical analysis of all acute inpatient episodes in years 1990/91 and 1991/92 which sought to identify health and socio-economic factors associated with utilization, after adjusting for variations in health care supply. The analysis used data from 4,985 small areas in England with average populations of about 10,000 to identify the link between health care needs and NHS inpatient spending, and resulted in the development of the needs index shown in Table 1. Note the model is multiplicative, so that the reported coefficients represent the exponent on each variable. The index is now used to distribute the acute element of Hospital and Community Health Service funds to health authorities in England.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standardized limiting long standing illness (under 75)</td>
<td>0.2528</td>
<td>0.0183</td>
</tr>
<tr>
<td>Standardized mortality ratio (under 75)</td>
<td>0.1619</td>
<td>0.0131</td>
</tr>
<tr>
<td>Proportion of economically active unemployed</td>
<td>0.0287</td>
<td>0.0092</td>
</tr>
<tr>
<td>Proportion of pensionable age living alone</td>
<td>0.0765</td>
<td>0.0130</td>
</tr>
<tr>
<td>Proportion of dependants in single carer households</td>
<td>0.0436</td>
<td>0.0122</td>
</tr>
</tbody>
</table>

Table 1: The York acute sector index.

In examining the implications of variability for GP budget setting, we use the York analysis as the basis of our analysis. It should be borne in mind that the data refer to all acute admissions (including emergencies) and do not include prescribing. Therefore they do not correspond exactly to the current standard fundholding scheme. Furthermore, the study refers to a two year period. Nevertheless, the principles discussed in the paper are likely to remain the same whatever formula is adopted and whatever broad area of health care is addressed.

**ANALYSIS**

The York analysis yielded a satisfactory statistical model which accounts for about 55% of the variability in NHS acute inpatient utilization within small areas, after taking account of systematic health authority effects and demography. However, by the same token, about 45% of the variability remained unexplained. The above discussion suggests that this should not necessarily be construed as a failure on the part of the analysts, who were seeking to capture component (1) above. Some of the unexplained variation may indeed be due to a lack of data on certain needs factors - component (2). To this extent the acute sector formula is amenable to improvement. However, much of the 45% is likely to be caused by variations
in clinical practice - component (3) - and the "random" element in health care costs - component (4) - a phenomenon that cannot be captured in any capitation formula.

The acute sector index provides an unbiased estimate of an area's relative need for acute inpatient care. The focus of attention in this paper is however the aspect of variation that has not been captured by the formula. This comprises the omitted part of systematic variation (2), the variability due to clinical practice (3) and the random variation (4). It is captured by the level one residual in the small area multilevel model used as the basis of the acute sector index.\(^\text{21}\) This residual represents the difference between actual utilization and the utilization predicted by the formula.

The index was estimated in logarithmic form, so deviations from the multiplicative model must be calculated by exponentiation of the residuals. We take the logarithmic model first. The residuals are assumed to be normally distributed. Their average (weighted for population size) is zero, reflecting the fact that the acute sector index is an unbiased indicator of utilization. In aggregate, therefore, variations from the logarithmic model should cancel each other out (assuming there are no behavioural responses amongst purchasers to the budgets they are set!). The weighted variance of residuals in the logarithmic model is calculated as 0.0113, standard deviation 0.1062.

Residuals from the multiplicative model - the basis of the resource allocation formula - can be calculated by exponentiation. For example, assuming normality, two standard deviations above the mean yields a multiplicative variation from the multiplicative model of \(\exp(2 \times 0.1062) = 1.237\), or +23.7%. The percentage deviation from the budget implied by the multiplicative acute index can therefore be calculated by taking the exponent of each residual and subtracting one. The weighted average of the absolute values of these deviations is approximately 8%, an indication that - amongst the small areas - the unexplained component of utilization is substantial. It should be noted that the average small area population size of 9,648 is roughly equivalent to the list size of a typical fundholding general practice.
The interest in this paper is in the extent to which aggregation of small areas leads to a diminution in the percentage variability of utilization. In accordance with the above results, we assume that the logarithmic residual for each small area is drawn from a normal distribution, mean zero, variance 0.0113, standard deviation $s = 0.1062$. For a sample of $n$ small areas, the deviation from the predicted logarithm of per capita expenditure will be distributed normally with mean zero and standard error $s/\sqrt{n}$. That is, as the number of small areas increases from 1 to $n$, so the variability in per capita expenditure - as measured by the standard error - decreases by a factor of $1/\sqrt{n}$. This result enables us to construct a variety of ways of illustrating the variability in budgets for different population sizes. Those that follow are merely illustrative of the sort that can be carried out, and are not intended to be exhaustive.

Figure 1 shows the 95% confidence limits for variations in actual annual expenditure from the per capita target implied by the acute sector model as population size increases - that is, the limits within which actual expenditure will fall in 95% of cases. These are constructed in the logarithmic model by computing $\pm 1.96s/\sqrt{n}$ for each value of $n$, the number of small areas aggregated. (Note that one small area is equivalent to a population of 9,648 in the following charts.) The Figure converts the results into multiplicative form, and shows the need for a very large tolerance (of about $\pm 20\%$) for a population size of 10,000. This decreases rapidly as population size increases.

An alternative way of examining variability is to explore the probability that actual expenditure will be more than (say) $\pm 10\%$ away from target in any one year as population size increases. This is calculated as follows. First the multiplicative deviations 1.1 and 0.9 (equivalent to $\pm 10\%$) are converted to logarithmic form: namely, $\ln(1.1) = 0.09531$ and $\ln(0.9) = -0.10536$. For each $n$, that part of the probability distribution lying outside those limits is then calculated for a normal distribution with mean zero, standard deviation $s/\sqrt{n}$. While the probability of a 10% deviation is 1 in 3 for populations of 10,000, for populations of 100,000 the probability of such a large deviations from target becomes very small.
Figure 1: 95% confidence intervals for actual expenditure in relation to budget

Figure 2: Probability of more than an x% variation from annual budget
becomes 1 in 400, and for a typical Health Authority with a population of 450,000 it is microscopic. By way of contrast, the probability of 5% deviation for a Health Authority is 0.0013, and of a 1% deviation 0.52.

Finally, Figure 2 offers another perspective on the analysis. For three population sizes (10,000, 100,000 and 500,000) the chart shows the probability that the absolute percentage deviation from target will exceed certain levels. For each population size P the number of small areas n_p required is calculated as P/9,648. Then the relevant standard error is s/√n_p. Notice that an annual deviation of more than 4% is very unlikely for the largest population group. However the chart implies that large deviations from target will be the norm amongst populations of 10,000. These empirical findings are consistent with results from simulation studies in which the allocation "formula" is implicitly known precisely.22

DISCUSSION

The above analysis indicates that any budgeting system for GP practices will be subject to considerable random variability that cannot be captured by any allocation formula. Clearly such variability may have a number of consequences, most especially in the way that GPs respond to any variation in expenditure from their budgets. In examining these consequences it is important to bear in mind that in the first instance neither the GP practice nor the health authority is likely to have a clear idea as to whether the recorded variations in expenditure are due to variations in clinical practice, variations in contract prices or variations in the incidence of illness. An algebraic formulation of the problem might be as follows:

\[
\text{Expenditure} = \text{Budget} \times (1 + C) \times (1 + P) \times (1 + I)
\]

where C is the variation in clinical practice from the local norm; P is the variation in prices from the local norm; I is the unpredicted variation in incidence and severity of illness experienced within the practice. The units of C, P and I can be rescaled so that their averages
are zero, but the above analysis suggests that within any one practice there may be big variations from zero. These variations have potentially serious consequences for NHS purchasing policy, particularly as it relates to GP fundholding. In particular, they may have important consequences relating to the way in which GP budget holders respond to any variation in expenditure from their budgets.

Some budget holders will find that their outturn expenditure is considerably lower than their budgets. The underspend could therefore be used in a number of ways:

- to fund other services and developments for the practice patients;
- to refer more patients than they otherwise might have done;
- to purchase a higher quality of care than is the norm for NHS procedures;
- to contribute to the practice's reserves as a contingency for future overspends;
- to contribute to overspends in other budget-holding practices.

Furthermore, low spenders who are fundholders might adopt a relatively relaxed attitude in negotiating prices with providers.

In contrast, some budget holders will experience substantial overspends. Such "high spenders" may respond in a number of ways. On the clinical side they might reduce the number of referrals and the quality of care (as expressed, say, in waiting time) offered to patients for fundholding procedures. Fundholders in this position are likely to become relatively price sensitive in contract negotiations with providers. Behaviourally, they may react with hostility to the budget setter (the Health Authority). In the extreme, high spending practices may resort to crisis management, giving rise to a wide range of potentially adverse consequences.

A most important result of the variations in financial pressures experienced by GPs could be that patients with identical conditions will be treated differently depending on the circumstances of their GP - that is, the budgeting scheme will give rise to inequities. Such inequity may take two forms: inequity within GP practices and inequity between practices.
Inequity might arise within a practice because GPs may experience differing pressures depending on the time of year.\textsuperscript{23,24} For example, at the start of the year a practice might feel relatively relaxed about its budget position. However, as events unfold, the practice might perceive that its current pattern of spending is projected to lead to an overspend in the budget. One response could be to amend its criteria for referring patients, either by adopting a more severe threshold for referral or by referring under a contract with a lower price or lower quality of care. In other words, patients who present later in the year might experience different referral practice.

In the same way, inequity could arise between practices because of the differences in budgetary pressures. As noted above, such variations might allow different practices to adopt different standards of care, in terms of both quantity and quality of treatment. Identical patients could therefore experience different health care depending on the practice they belong to.

The high levels of risk inherent in meeting expenditure targets might induce GPs to adopt a number of defensive tactics. For example, fundholders might negotiate block contracts with providers. Such tactics, if adopted on a widespread basis, would undermine many of the objectives of budget holding, as they remove the incentive for GPs to adopt innovative approaches to health care. As an alternative defensive devise, practices might establish large contingency reserves by reducing the amount and quality of care offered to their patients. Moreover, GPs might seek to insulate themselves against overspends by "cream skimming" healthier patients.\textsuperscript{25} They might even take out insurance with the private sector, resulting in an unproductive outflow of funds out of the NHS.

Thus, under a budgeting regime, the unpredictable variability inherent in the demand for health care gives rise to a fundamental tension between efficiency and equity. If budgets influence GP behaviour, they are likely to lead to variations in clinical practice which seek to
countervail unpredictable variations in disease incidence. Clearly variations in referral practice compromise the principle of equity - one of the cornerstones of the NHS. Therefore, if Health Authorities do not handle variations from budgets with care and understanding, the whole budgeting system could be discredited. And, unless measures are taken to moderate the pressures towards inequitable treatment, the budgeting scheme could give rise to hostility and alienation amongst GPs and patients.

What, then, can be done to avoid this potentially damaging situation? Given below are some strategies that could be adopted to manage the variability we have described, and to moderate the risk faced by general practitioners.

One of the obvious mechanisms for ameliorating the potentially adverse effects noted above would be to increase the size of patient groups for which budgets are set. This principle would suggest that a number of practices might be pooled, effectively to become a single fundholder. The members of such a grouping would in effect offer each other a degree of insurance against random variability. An additional benefit of such arrangements may be some reduction in managerial costs associated with negotiating and monitoring contracts with providers. These arguments would support the move towards multifunds and locality purchasing. However, pooling practices for the purposes of risk sharing requires a degree of mutual trust and cooperation. In choosing the appropriate level of aggregation, therefore, there is a need to balance the reduction in risk against coordination costs, and there may be an argument for basing the choice of partners on factors such as similarity of patient characteristics as well as geographical proximity.

In the same way, the problem of unpredictable variability can be alleviated by setting budgets for a period of more than one year. Under such arrangements, practices can to some extent offset underspends in one year against overspends in another. Longer budget periods would enable Health Authorities to become aware of serious budgetary problems at an early stage and to take appropriate action without inducing a crisis in the affected practice. Health Authorities, in practice, inevitably find themselves adopting a time horizon longer than one
year in their dealings with fundholders, and a move towards "rolling" budgets of (say) three years may be a helpful way of formalizing this longer term view.

To some extent the problem of large unpredictable variability is caused by the random incidence of a small number of very expensive procedures. There is therefore a strong argument for excluding certain expensive procedures from the budgeting scheme. The existing standard fundholding scheme acknowledges this difficulty by excluding certain procedures and costs in excess of £6,000 incurred by any one patient in any one year, and there may be an argument for local discretion in such matters.27

Similarly, a practice's expenditure might be highly influenced by a small number of highly dependent patients. There is therefore an argument for excluding predictably expensive patients from the budgeting scheme. The Health Authority would cover all or part of the costs of procedures required by such patients. By definition, such patients will suffer from some chronic condition which necessitates predictably high health care expenditure. The principal difficulty with such arrangements would be determining criteria for excluding a patient, and ensuring that any such criteria did not offer GPs perverse incentives to exaggerate the seriousness of a patient's condition. A beneficial by-product of such arrangements is that they would reduce the incentive for GP practices to cream skim healthier patients.

Although presented above as a potentially dysfunctional outcome, there might nevertheless be some merit in experimenting with contractual form on the part of those GPs who are fundholders. Clearly a move to block contracts - under which all the risk is transferred to the provider - would effectively remove many of the intended incentives underlying fundholding. However, if it is believed that the use of cost per case contracts (which is implicit in our analysis) exposes the budget holder to unacceptable risk, then there may be an argument for judicious use of cost and volume contracts, under which providers are guaranteed a block fee
and incremental patients then attract less than full costs. Such contracts effectively enable the risk to be shared between purchaser and provider.

A consequence of the above analysis is that all Health Authorities will need to establish contingency reserves to accommodate "overspends". Such reserves may be treated in a number of ways. For example, the total Health Authority budget might be top-sliced to create a type of insurance fund. Or "underspenders" could be given incentives to return some of their unused funds, perhaps by guaranteeing that they will receive a credit for any overspends in the future.

Finally, and most importantly, variations from budgets should be explored carefully. It is essential that every effort should be made to establish the extent to which such variations can be attributed to differences in disease incidence, variations in contract prices or variations in clinical practice. Some element of peer review may be appropriate, and engaging in a dialogue with budget holders will be essential if GPs' confidence in the budgeting system is to be retained.

CONCLUSION

All of the measures described above are likely to have some role to play in easing the tensions generated by a GP budgeting regime. However, if the setting of budgets is to continue to influence GP behaviour, no managerial measures in themselves can completely eliminate the potentially damaging consequences of unpredictable variability in demand for health care. Instead, all the evidence suggests that Health Authorities will have to adopt a very sensitive managerial style in how they react to variations from budgets, and be prepared to engage in constructive dialogue with fundholding practices. In spite of the difficulties raised above, the increased devolution of budgets within the NHS is likely to continue, regardless of political leadership. The challenge to both Health Authorities and general practitioners is to make the system work.

REFERENCES


