BUDGETS IN THE NEW NHS

Kate Baxter, Lecturer in Health Economics, University of Bristol

Marjorie Weiss, Lecturer in Primary Care, University of Bristol

Paper presented at HESG, Nottingham, July 10th-12th 2000

Kate Baxter
Department of Social Medicine
University of Bristol
Canynge Hall
Whiteladies Road
Bristol BS8 2PR
Email: c.baxter@bris.ac.uk

Acknowledgements: Thanks to Gwyn Bevan, LSE, for comments on an earlier draft.

Work in progress. Please do not quote.

This paper comprises early thoughts about the purpose and consequences of the new budgetary arrangements in the NHS in England. The ideas put forward form the basis of a PhD thesis and are in a formative stage. The paper is structured as follows. Section 1 gives a brief history of the organisational changes that have taken place in the NHS in the last decade. Section 2 illustrates the importance of the principal-agent problem in health care, with the associated theories of insurance and transactions costs, in an attempt to lay the foundations for understanding some aspects of these changes. Section 3 illustrates some related findings from the national evaluation of total purchasing pilots and suggests where these issues might occur with primary care groups and trusts. Section 4 summarises our proposed research plan, and Section 5 concludes with areas for further consideration and debate.

SECTION 1: CHANGES IN THE STRUCTURE OF NHS PURCHASING

Hospital services comprise emergency (unplanned) care and elective (planned) care. Pre-1991, GPs were free to refer their patients (as emergency or planned cases) to any secondary care provider for specialist outpatient review or inpatient admission. GPs therefore managed some of the demand for secondary care and their decisions had a direct impact on costs. Although GPs acted as gatekeepers to secondary care, there was no direct monetary link between the GPs making the referral and the payment for the care subsequently received. Not only did GPs refer where and in what quantity they chose, they also varied widely from each other in the extent to which they made referrals (Hutchinson, 1993; Fertig et al. 1993; Wright and Wilkinson, 1996).

Therefore, GPs in their traditional gatekeeper role managed the demand for care, either by treating patients themselves, or by referring them for specialist care. (Once in the secondary care system, the management of care transfers to the hospital doctors, who then make treatment decisions on the patient's behalf.) On the opposite side of the equation, the health authorities (HAs) funded hospitals to provide services. However, HAs were funded through a complex capitation formula based on their catchment populations, and there was no real relationship between the services hospitals provided and the funding they received from HAs. In effect, HAs were acting as independent insurers, funding care but having no influence over the level of demand.

In 1989, the government white paper *Working for patients* (Secretaries of State, 1989) changed these roles. The internal market was introduced, purchaser and provider roles were split, and a new scheme called GP fundholding began. This scheme gave large general practices the option to manage some budgets, including cash limited GMS, prescribing and a proportion of elective secondary care. In the first instance, the services were restricted (to reduce financial risk from random fluctuations in demand) and budgets were small relative to total hospital and community health services (HCHS) budgets. Nonetheless, this was a first step towards integrating the two sides of the NHS equation – bringing together the management of patient demand and payment for care. Simultaneously, the purchaser/provider split and internal market created a system dependent on contracts for the provision of secondary care services; HA and GP fundholders agreed contracts with hospital trusts to provide services to their patients. Over the following decade, the fundholding scheme was expanded, with the types of services for which budgets could be held increasing and the sizes of practices eligible to hold a budget decreasing.

Other formal and informal schemes, such as locality commissioning, multi-funds and GP commissioning groups, developed over the years, each taking a slightly different stance on the basic premise of GPs managing their own budgets for secondary care services.

In 1994, an NHS Executive letter (NHSE, 1994) entitled *Developing NHS purchasing and GP fundholding* announced the expansion of GP fundholding and the start of a new pilot scheme – total purchasing (TP). TP gave single or multiple groups of general practices the opportunity to hold a budget for and purchase *all* HCHS for their patients. This meant that total purchasing pilots (TPPs) could buy elective care that was outside the GP fundholding scheme and buy emergency care. This scheme was another major step in integrating the GP demand manager/gatekeeper role with the traditional HA role of insurer. In fact, TPPs did not choose to manage all HCHS, but instead managed only the budgets for those services to which they wanted to make changes (Mays et al. 1997).

With the change of government came another white paper on the NHS in England in 1997 (Secretary of State for Health, 1997). This paper, *The new NHS: modern, dependable*, announced a further restructure, building upon the experiences of the previous decade. The aim was to keep the systems that appeared to work (for example the purchaser provider split), but abolish those that were perceived as not working or not aligned with New Labour's philosophies (for example, fundholding, which was perceived as creating a two tier service within the NHS). Instead of voluntary fundholding, the 'new' NHS would comprise large groups of general practices, between them holding budgets for the majority of HCHS, and they would be known as primary care groups (PCGs) or primary care trusts (PCTs). These new bodies could choose to function at one of four levels:

(1) to advise the HA in its commissioning of care;

(2) to take responsibility for a delegated budget, as a sub-committee of the HA;

(3) to become a free-standing body for commissioning health care, accountable to the HA; and

(4) to become a free-standing body for commissioning health care, accountable to the HA and in addition, to be responsible for the provision of community health services.

The first two levels are known as PCGs, the latter two as PCTs. This new NHS was introduced in April 1999.

Over the last decade, we have therefore moved from a system in which there was no relationship between demand management and costs, to one where every GP in England is part of a relatively small budget managing group. Each group has direct control of the budget from which services are provided.

SECTION 2: THEORETICAL BACKGROUND

This section describes the principal-agent problem in health care and two subsequent economic theories that appear to have some relevance to how PCG/Ts may operate. The first is the theory of insurance and moral hazard - relevant because GPs have taken on the role of insurer by managing demand within budgets, thus removing the

principal-agent relationship at that level, and taking responsibility for controlling the effects of moral hazard. The second is that of market failures, transaction costs and contracting - relevant because the purchaser/provider split (and consequent agency relationship) is still operational, with long term contracts needing to be agreed between PCG/Ts and hospital trusts. These issues are illustrated in the diagram on the following page.

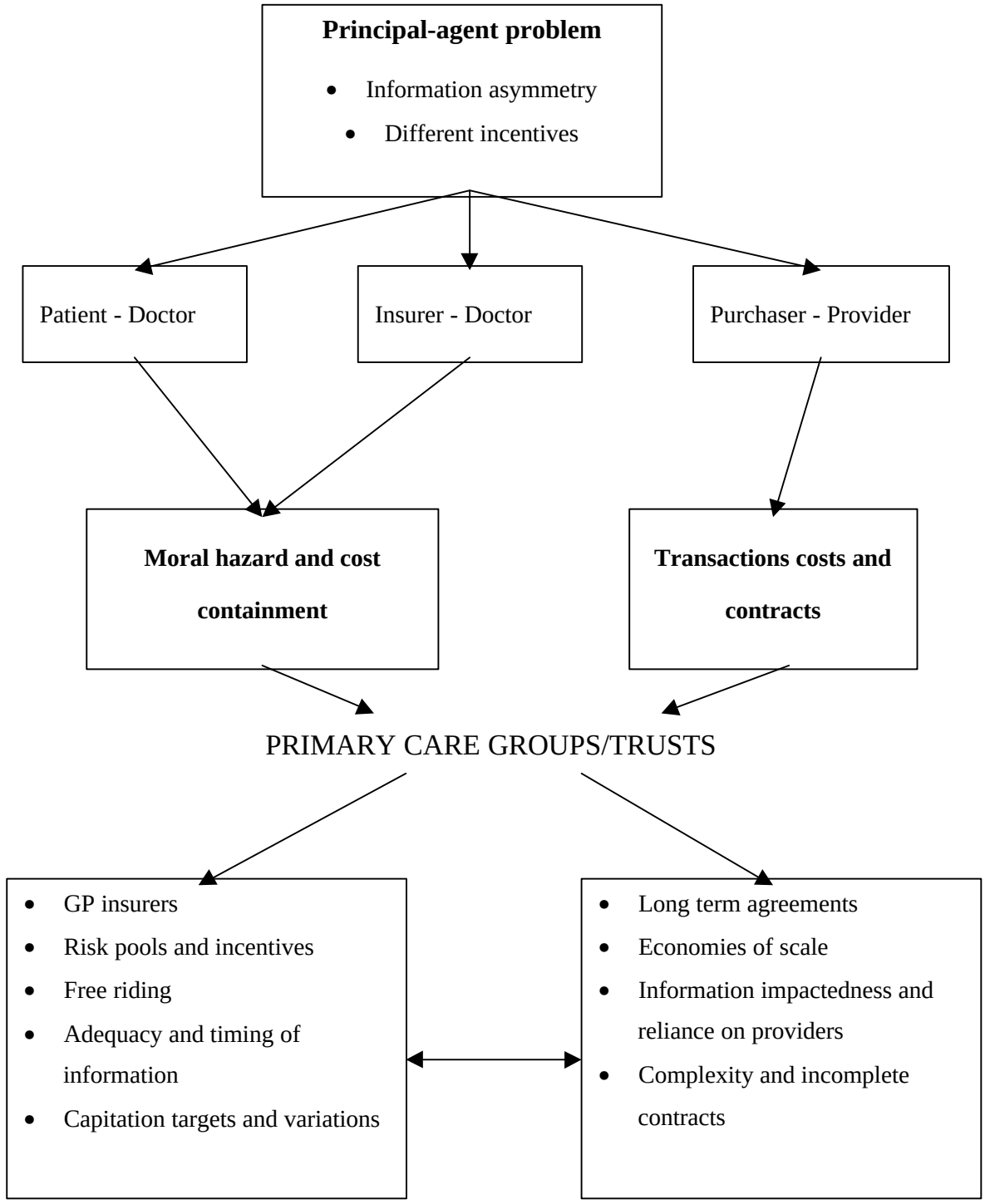
Principal-agent relationships in health care

The principal-agent problem arises in health care as in many other areas of economic activity. It occurs when a person (or group of people) hires an agent to perform tasks on their behalf, but cannot guarantee that the agent performs the task in exactly the way in which the principal would like (Bannock et al. 1998). This happens as a result of (1) information asymmetry between principal and agent – the agent is better informed, and (2) different incentives – the incentives faced by the agent are not those that maximise the principal's utility. Both these conditions must be met for the relationship to become a potential problem. The efforts of the agent are impossible or very expensive to monitor, and the outcome is similarly difficult for the principal to judge due to information asymmetry. If a complete and enforceable contract can be drawn up to specify all the duties of the agent, problems do not arise. However, drawing up a contract that specifies all possible contingencies is also very expensive or impossible.

The principal-agent problem is inherent in health care at a number of levels. The most frequently cited level is that of the doctor acting as agent for the patient. The patient has little knowledge about his/her medical condition and turns to the GP (who is better informed) for advice on treatment. Despite this information asymmetry, so long as the incentives faced by the GP result in him/her making treatment decisions that maximise the patient's utility, the principal-agent relationship is not a problem. If the doctor faces incentives that encourage decisions that do not maximise patient's utility (for example, financial incentives), and the patient cannot judge the resulting benefits, the principal-agent relationship can become a problem.

The doctor also acts as agent for the insurer. In this case it is the insurer that wishes to provide 'good health' to its clients, but lacks adequate knowledge to do this. The insurer therefore relies on the better informed doctor to supply the appropriate amount of health care to the insured population, within the budgetary constraints of the

THEORY INTO PRACTICE?



insurance company. However, although the doctor is better informed about the population's health care needs, s/he is not as knowledgeable about the financial constraints of the insurer, nor is the doctor facing incentives that encourage maximisation of the insurer's utility within these constraints. The principal-agent relationship is therefore not perfect.

A further situation in which the principal-agent relationship plays a role is between the provider and the purchaser; the provider acts as agent for the purchaser. This is a similar case to the insurance situation, in that the purchaser commissions health care from a provider, but is less well informed than the provider about the amount and quality of care delivered. Indeed, the purchaser may also be less knowledgeable about the type and amount of treatment needed. The provider faces incentives to maximise its own utility (which may or may not include profit), and thus does not necessarily maximise purchaser utility. Again, the relationship is therefore imperfect.

At each level, where an individual or group of individuals is reliant upon another to perform tasks on their behalf, problems attributable to the principal-agent situation arise. The cost to the principal of increasing their knowledge to that of the agent (that is, reducing the information asymmetry) in each case is unrealistically high. Similarly, the cost of developing complete contracts to ensure the maximisation of the principals' utility (allowing for all possible contingencies) is also unattainably high. The remainder of this section explores issues arising from the complex principal-agent relationships between the patient/doctor/insurer, and that between the purchaser and provider.

Insurance and moral hazard

Individuals, when faced with a risk of an undesirable event occurring, the result of which is a potentially large loss (monetary or otherwise), purchase insurance. Insurance provides protection against these potentially large losses incurred as a result of unpredictable events. Risk averse individuals prefer the certainty of regular predictable and relatively small premiums, rather than the uncertainty of large losses. Typical examples are household, car and holiday insurance.

However, although an insurance scheme protects the individual, the incentives that the individual then faces can induce a change in behaviour. In a full insurance system, individuals pay nothing at the point of consumption, therefore consuming the product in question more often, and in greater quantities, than if they paid the full cost. For example, take the case of car insurance. After minor damage to a car, an individual without comprehensive cover (a) will not make a claim and (b) may not repair the damage. If the damage is repaired, this will be undertaken at minimum cost to the individual. An individual with comprehensive cover (a) will make a claim and (b) will expect a 'no expenses spared' repair by the insurance company. This phenomenon is what is often termed an 'insurance job'. It occurs because of incentives that encourage individuals to act in ways which incur costs that they do not bear themselves. In economic terms, this is labelled *moral hazard*.

Health and health care are characterised by uncertainty, information asymmetry and potentially high costs (Arrow, 1963). These factors explain why health care is funded through insurance, with (in most countries) governments

being the dominant insurer. However, a consequence of insurers being third party payers, protecting both consumers and suppliers from the costs of services, creates the problem of moral hazard, with neither supplier nor consumer having an incentive to be concerned about costs.

Views about the importance of moral hazard and the effects of incentives differ. One dominant view is that, if the cost at the point of consumption is reduced or removed, then the consumer has an incentive to move down his/her demand curve, and consume more. Thus the welfare gains from insurance in health care (wider access to care, unlimited by ability to pay) may be, at least in part, offset by increased consumption (Arrow, 1963).

An opposing argument is that rational consumers can see that their over consumption of health care leads to higher overall costs of care and therefore increased premiums (Pauly, 1968). This then acts as a disincentive to consume more (i.e. reduces moral hazard). In large insurance risk pools however, the impact of each individual on the whole is small, thus negating the incentive to contain consumption. In fact, the incentives can be reversed. Consumption may be increased, in the belief that the effects of an individual's actions on the whole insured group are so small as to be negligible. However, if all those insured hold this belief, total consumption will increase, and premiums will subsequently be raised. This can then result in even greater consumption as individuals try to recoup the cost of higher premiums through even higher claims.

Solutions to the moral hazard problem can either be through institutional control - regulations to stop over consumption, for example, financial restrictions, or through the market - based upon some degree of cost sharing by consumers, in the form of either deductibles or co-payments, reintroducing incentives to be concerned about costs.

Insurance, moral hazard and health care

Experiences of controlling moral hazard in health care have included introducing regulation on the supply side. One example is to regulate efficiency, where efficiency is defined as reduction in unit costs per inpatient admission. In the UK in the 1990s, purchasers were regulated through the efficiency index, requiring a reduction each year in the average cost per admission. Due to large fixed costs, the easiest way to decrease costs per case is to increase throughput. This, however, can then result in increased total costs. The UK controls this in part through cash limits on expenditure, but this results in the NHS being in an endemic crisis from lack of funds. Whilst the search for efficiency may continue, the problem of escalating costs has not been solved.

Health insurers are therefore experimenting with tackling moral hazard using market incentives on the demand side. The conventional response is to impose charges on patients at the point of consumption (for part or all the costs of care). Although this conventional response is used extensively in the US, its efficacy is hotly contested (Evans, 1987). This is for two reasons. First, given patients' imperfect information, and variations in ability to pay, it is not a simple task to set co-payments at a level that discourages unnecessary use of services without also reducing appropriate and important service use. Second, due to asymmetry of information in health care, it is

doctors, not patients, who make decisions on demand, they act as agents for patients. Thus, decision makers will not respond to patient charges, because the decision makers are not the patients.

Insurers are now exploring the effects of tackling the moral hazard of *doctors'* decision making on demand for health care. US insurers have developed various ways of doing this through "managed care", of which the family of Health Maintenance Organisations (HMOs) is particularly important. HMOs are funded by capitation; and in one classic form of HMO, insurers have integrated into the demand side by employing primary care physicians to act as gatekeepers to specialists.

Health authorities in the NHS market between 1991 and 1999 were insurers who practised "unmanaged care". Neither hospital doctors nor GPs (both making decisions about care) were employees of the authority. Thus many authorities faced the recurrent problem towards the end of each financial year of hospital trusts "overperforming" - delivering volumes agreed in contracts before the end of the year. GPs and hospital doctors continued to identify cases needing care, but authorities had no extra money to pay for these cases. Integrating the insurer role with that of the clinician could help address this issue, decreasing moral hazard and containing costs.

Transactions costs and contracting

Transactions costs are the costs associated with buying or selling products, and affect firms' decisions on whether to contract to buy products from another firm, or make the products in-house (Bannock et al. 1998). Decisions on whether or not to produce a good in-house depend on the in-house transactions costs relative to those incurred through contracting on the market. A contract can be seen as being negotiated between a buyer (principal) and supplier (agent). Where there is information asymmetry and the players are facing different incentives, there are trade-offs between the costs of reducing the information asymmetry and ensuring a complete contract, and the risks of agreeing an incomplete contract. The costs of agreeing complete contracts are dependent on the interaction of a number of conditions (Williamson, 1975; Williamson, 1985), referred to as the organisational failures framework, summarised below.

Bounded rationality – This condition assumes that human behaviour is rational, but only to the point where the limits of human understanding and language are reached. Beyond this limit, fully informed agreements cannot be reached.

Uncertainty and complexity – for uncertain and/or complex contracts, a full range of contingent actions cannot by specified at the outset – the range of possible responses is too lengthy and complex.

Where decisions are made under uncertain conditions or complex environments, the costs of determining the full range of potential options is very high, and so the bounded rationality constraint is reached more quickly.

Opportunism - Conventionally, rational economic agents are guided by self-interest. Opportunistic behaviour extends this convention to 'self-interest seeking with guile', that is, making false or empty threats and

promises, in an attempt to gain an advantage. With information asymmetry, the consumer or payer is not able to distinguish true from false promises.

Small numbers - Where there are a large number of players in competition for contracts, it pays to be honest about intentions – dishonesty will result in other players taking over the contract at renewal. However, where a small number of players exist, and there is little threat of losing a future contract, it is in the interest of each party to seek terms most favourable to themselves. In addition, although large numbers of firms may bid for an initial contract, if the winner gains cost advantages for the contract renewal, small numbers may be exhibited in subsequent rounds.

Small numbers make opportunism appear to pay, resulting in opportunistic representations and subsequent bargaining. This strategy may be advantageous to individual players, but increases the total transactions costs in the system as a whole.

Information impactedness - Information impactedness exists when there is information asymmetry between players, and when it is not costless to reduce this asymmetry. One player may have more information than another, or each may have different information. Whenever information is distributed asymmetrically, any exchange will be subject to a degree of risk.

Atmosphere - Atmosphere relates to the *process* of the exchange, that is, the value gained from a satisfactory exchange process becomes an integral part of the economic problem.

Asset specificity - Asset specificity exists where a supplier invests heavily in delivering a contract, and there are limited alternative outlets for the supplier's assets. The supplier is therefore dependent upon future contracts with the buyer.

In summary, limits to understanding and ability to form a complete contract are reached quickly when the future is uncertain or complex. The opportunity to make false promises in order to gain a contract is increased where the number of players is small. Information asymmetry and suppliers' dependence on a single purchaser place further uncertainty on the process and can reduce the number of players. Exchanges take place within an environment that can either enhance or worsen the trading experience (Williamson, 1975).

The existence of some or all of the above market failures makes the decision to contract in-house more attractive. Williamson (Williamson, 1975) argues that moving from a market to a hierarchical (in-house) exchange system could bring the following advantages:

1. adaptive, sequential decision making (economising on bounded rationality)
2. reduction of opportunism resulting from (*ex ante* or *ex post*) small numbers exchange
3. promotion of convergent expectations, thus reducing uncertainty
4. reduction of conditions that create information impactedness, and reduction of strategic behaviour as a result of information impactedness

5. a more satisfying trading atmosphere.

Transactions costs, contracts and health care

The 1991 purchaser provider split introduced costs due to contracting. A relatively large number of purchasers (HAs and GP fundholders) began to contract for services with providers. The NHS moved from a system of internal hierarchical organisation to one of independent trusts and purchasers trading in a (quasi-) market. The pre-1991 NHS, with the absence of a market, seemed to satisfy all Williamson's arguments for a hierarchical system of exchange. The reforms, however, introduced the associated problems of organisational failure as detailed above.

Throughout the 1990s there were small numbers of suppliers (in the form of local trust monopolies) that had incentives to act opportunistically, as many purchasers had no viable alternative supplier. Similarly, there was information asymmetry, both before and after contracts had been agreed; the purchaser was dependent on the provider for information about the costs of care. This may have been particularly important where purchasers were attempting to reduce the value of contracts and invest elsewhere – providers were better informed about the marginal and fixed cost components of care. Where such reconfiguration of services was considered (and will be in the future), re-contracting was almost certainly necessary, with the associated transactions costs of re-negotiation. Assets are certainly specific to the provision of health care services, but it has been argued that this is not an important issue, as assets are owned by the state and are specific to a local population, not to a purchaser or provider (Dawson and Goddard, 1999).

Directly measurable transactions costs associated with contracting would have been expected to rise due to the annual re-negotiation of cost per case or cost and volume contracts, with GP fundholders in particular. For HAs, evidence from Dawson and Goddard (Dawson and Goddard, 1999) suggests that many contracts were three year or three year rolling. Despite this, these contracts were still re-negotiated annually, and so did not help to reduce costs. In addition to direct costs, complications associated with information asymmetry, complexity and small numbers may also have increased.

In the new NHS, contracts are now termed 'long term agreements' (LTAs). These are in effect rolling contracts of between 3 and 5 years. However, information asymmetry between purchaser and provider still exists. It remains to be seen whether the new LTAs can reduce yearly direct transaction costs or solve some of the problems of organisational failures.

SECTION 3: ADDRESSING THE ISSUES – THE FUTURE FOR PRIMARY CARE GROUPS AND EVIDENCE FROM TOTAL PURCHASING PILOTS

This section takes the moral hazard and transactions costs theories in turn, to consider how the set up of PCG/Ts may be related to these issues, and considers some experiences from total purchasing pilots.

Moral hazard and cost containment

GP insurers

Recent experiences in the USA and UK show that, increasingly, the role of health care insurer is being integrated with that of health care provider/decision maker in an attempt to contain escalating costs that result in part from moral hazard. In the USA, a classic model of health maintenance organisation was based on employing primary care gatekeepers; in the UK, the last decade saw various experiments with GP budget holding. Insurer and clinical roles are now more fully integrated than ever before in the UK, with PCG/Ts managing their HCHS, prescribing and cash limited GMS budgets, shared between practices.

The total purchasing pilot scheme (TPP) experimented with this integration in the late 1990's. Volunteer practices worked alone or in groups to manage at least some of their hospital and community health services (HCHS) budgets that were outside the fundholding scheme. The extent to which this new insurer role was fully integrated with the traditional clinical decision making role varied. In many TPPs, a single or small group of GPs took responsibility for budgetary management whilst the remainder continued with their usual approach to clinical decision making (Baxter et al. 2000).

Risk pools and incentives

The number of GPs in a single PCG/T is large, compared to the group practice sizes within which GPs have been used to working. If the average size of a PCG/T is around 100 000 patients, the number of whole time equivalent GPs is about 50. In fact, average sizes may increase as a number of PCGs plan to merge (Wilkin et al. 2000). As discussed in the theoretical background section, the size of the risk pool can affect the incentives faced by those insured, the potential for moral hazard increasing with the size of the risk pool.

The TPP evaluation showed differences in ability to manage budgets between single and multi-practice TPPs, which arose as a result of the numbers of GPs and practices within the pilots. Single practice and smaller TPPs were able to manage budgets better (and were initially more successful in implementing change) than their larger counterparts (Baxter et al. 2000; Mays et al. 1998). One reason for this could be that smaller groups of GPs were able to communicate and work together as a team better than larger groups. In insurance terms, the single practice TPPs were smaller insurers, and had fewer GPs in the risk pool. The impact of each individual GP's demand for

care (in terms of referral costs) would have had a more noticeable impact on total costs, and so moral hazard would be less of a problem.

Integration of general practitioners' insurance and gatekeeping roles could be encouraged within primary care groups by either developing internal budgeting at practice level, or for small groups of general practitioners to volunteer to work together. Such small groups could remove moral hazard altogether. NHS circulars suggest both devolved budgets at practice level (DoH, 1998a), and managing budgets at the level at which spending decisions are made (DoH, 1998b). Developing indicative practice-based budgets is in keeping with the findings from TPPs, but the logistics of information management may not allow it (see 'Adequacy and timing of information').

Where PCGs choose to allow the HA to manage the budget for a proportion of services, for those services, there may be less incentive to manage demand than for services for which a budget is held. In many PCGs there will be practices with little or no direct purchasing experience. Level 1 and 2 PCGs will be sharing their insurance role with their HAs. The situation may be similar to that pre-1991, with the HA as third party insurer.

Free riding

Our TPP research discovered general practitioners who were "free riders", happy to accept benefits of holding a budget for their patients (for example, more practice based services) provided that it did not interfere with their clinical autonomy. In multi-practice total purchasing pilots, the impact of each general practitioner on the whole organisation is relatively small, thus it is easier for individual GPs to "free ride". A key difference between PCG/Ts and TPPs is that whereas participation in TPPs was voluntary and selective, participation in PCG/Ts is compulsory and universal. The incentives for individual general practitioners to "free ride" within these groups could outweigh the incentives for integrating their clinical role with that of insurer (given that not all GPs chose the additional role of insurer). On the other hand, as TPPs were voluntary, it would have been relatively easy for practices to withdraw from the scheme, or even to be asked to leave if they were not 'pulling their weight'. With PCG/Ts, individual practices have no choice, at best, they can leave one PCG to join another. Perhaps free riders will be tolerated less than they were in TPPs, as their lack of commitment could affect future performance for all. It is not clear what incentives there will be to encourage full participation, or sanctions to discourage free riding.

Adequacy and timing of information

Where a PCG/T is holding a budget, all GPs will be affected by a greater awareness, and the more direct consequences, of their own referral actions on resources available to colleagues, within their practice as well as within their PCG/T. TPP overspends were often covered by host HAs, that is, from funds allocated to non-TPP practices. In the 'New NHS', HAs will hold funds for level one PCGs and for services blocked back by level two PCGs; HAs will not be able to use these funds to cover overspends by other PCG/Ts. Thus, if PCG/Ts overspend, this will be a first call on the next year's budget.

To avoid this plight, PCG/Ts must monitor demand (i.e. expenditure) as it is committed. In TPP, less than a quarter did this, the remainder accounting for expenditure when invoiced by the hospital. If expenditure is not known, then the PCG/T cannot manage demand effectively, and this negates the effect on cost containment of being a combined gatekeeper and insurer. The result will either be an overspend, or overheating of contracts, with patients having to wait until the next financial year before they can be treated. This would be no different to the situation when HAs were third party insurers. Indicative practice level budgets could help control spending, but if contracts are set at PCG/T level, management (and activity) information may only be available at that level. It is not clear how the management of devolved budgets will relate to the management of whole group contracts. PCG/T (or multi-PCG/T) level contracts may achieve managerial economies of scale (Posnett *et al*, 1998), but impair incentives for GPs to accept responsibility for the costs of their decisions.

Capitation targets and variations

Historic funding levels have placed little pressure on expenditure or patterns of care, and create no incentives to limit moral hazard. Under capitation, PCG/Ts that have a level of expenditure greater than their capitation target will face increased budgetary pressure, and as a result should attempt to abate moral hazard. Capitation funding raises questions around variation in resource use. Despite extensive research, high variation in referral rates and subsequent resource use by individual GPs and practices remains unexplained (Wilkin, 1987; Wilkin, 1992). Capitation funding, which assumes average referral rates and costs, places full responsibility for referral and admission variations with PCG/Ts.

However, losses or gains at PCG/T level may not be passed on to individual practices. A survey of resource allocation and financial incentives in health maintenance organisations (HMOs) in the United States found the degree to which capitation payments (and therefore risk) were devolved from HMO, through middle tier management, to individual primary care physicians or risk pool level was variable (Hillman, 1992). Therefore, although the PCG/T may be under financial pressure, individual practices may not feel the effects of (and thus incentives to stop) overspending.

Transactions costs and contracting

Long term agreements

A long term agreement is a contract that commits resources for more than one year, in many cases, for three to five years. The main features are stated in EL(97)39 as a continuing relationship between commissioner and provider, shared risks, fixed or mechanistic funding regimes, and an aim to increase quality and efficiency (DoH, 1997). The hoped for advantages over traditional short term cost-driven contracts are greater involvement of clinicians, users and carers, better focus of time on quality and effectiveness, and better planning for investment and change (DoH, 1997). It is not clear how big a change this is compared to the contracting that has been

undertaken by HAs in the past (Dawson and Goddard, 1999), nor how marginal changes in activity will be translated into price changes.

Economies of scale

NHSE circulars (DoH, 1998a; DoH, 1998b) suggest the importance both of integrating insurance and clinical management roles, and of devolving budgets to small group or practice level. This could increase the number of transactions and hence measurable transactions costs within PCG/Ts. It could also increase transactions costs due to organisational failure. There is therefore a trade-off between the size of the risk pool and the level of (measurable and non-measurable) transactions costs. Efforts to minimise moral hazard and to reduce transactions costs are pulling in opposite directions. The smaller the risk pool, the greater the chances of tackling moral hazard, but the greater the costs of negotiating and agreeing contracts. To reduce transaction costs, one would expect large purchasing groups to be created, especially for purchases of the same services. However, the larger the group, the smaller the cost consequences to individual GPs of not controlling demand. Thus, moral hazard is reintroduced. With many PCGs planning to merge (Wilkin et al. 2000), the size of risk pool will increase. Evidence from TPPs shows that smaller and single practice pilots were more successful in managing budgets and avoiding overspends (Baxter et al. 2000).

Information impactedness and reliance on providers

Information impactedness is the situation in which it is not costless to reduce information asymmetry. In the current NHS, PCG/Ts and hospital trusts both possess information about their own needs, but have limited information about each others needs or costs. In particular, PCG/Ts lack full information on the true costs of hospital provision, and costs of changes in the amount of care provided. Many contracts in the past have been based on, for example, changes in the purchasers budget, cost pressures of providers, and purchaser priorities for new developments (Dawson and Goddard, 1999). Such decisions are not based on costs of activity. A similar situation was seen with TPP where managers suggested that the amount of payment offered in contract negotiations was dependent not on the level of activity required, but on historic spend and the amount of funds available. In many cases, changes in levels of activity for TPPs, for example, shifting services to primary care, resulted in less than proportionate changes in payments (Mays, forthcoming). Although this is due to the costs of activity changes at the margin, it is not possible for the purchasers to elicit the true costs changes. PCG/Ts are instead dependent upon their providers to give true information. TPPs found information technology and gathering relevant information a problem. Many TPPs were critical of provider and HA information, some felt that the information available was not useful, others felt that providers and HAs were uncooperative or did not have the data (Mahon, 1998). With limited information, GPs face incentives to reduce services where marginal savings are greatest, not where efficiency is maximised.

Complexity and incomplete contracts

Trusts cannot state a single price in advance for the majority of treatments. The costs of many inpatient stays depend on the duration of stay and nature of care in addition to the type of treatment. Trusts have large fixed costs and relatively small variable costs, making the average cost (and price) per case high compared to the marginal cost. Prices paid for care therefore relate more to the overall commitment of funding to a hospital than to individual packages of care.

EL(97)39 states that 'No agreement can cover every eventuality – nor should it try to.' (DoH, 1997). This statement underlines the complexity of delivering and costing health care, and of agreeing complete contracts. The advice given is that pre-agreed events should trigger pricing rules or re-negotiations, and that agreements should be approached 'as part of a long term, ongoing relationship rather than as a chance for short term opportunism.' (DoH, 1997). Despite this official guidance on LTAs, it is not clear exactly how these contracts will work, what will happen if contracts are overheating and additional funds are not available, or how changes in activity will be translated into changes in prices.

SECTION 4: PROPOSED EMPIRICAL WORK

This section gives a brief overview of our proposed research. The context for the research is the management of PCG/T budgets within a framework of unclear relationships between activity and expenditure, variations in referral, and pressures to contain costs. Limited budgets necessitate incentives or regulations to control spending. These incentives and regulations place GPs under pressure to alter referral activity to stay within budget. A change in referral activity, however, may not result in a proportionate change in expenditure. As a result, budgets may still be under pressure, and incentives and regulations could be strengthened further. Too much or too little stress results in poorer decisions than a moderate level of stress (Janis 1977; Dowie 1983). Pressures on GPs may increase, possibly resulting in poorer decision making, higher costs (as a result of delayed or inappropriate referrals), and a reduction in work satisfaction. We are interested to know if this is the situation, and whether the theoretical frameworks outlined are an appropriate starting point for analysis.

We aim to explore these issues by studying financial management arrangements in PCG/Ts, the nature of long term agreements, and the impact of budgetary pressures on GPs and their referrals.

Specific objectives are to:

- determine the nature of PCG/T budgeting systems, specifically, the regulations and incentives in place to control referrals and spend, and the way in which financial responsibility is passed on from PCG to practice level.

- describe the working arrangements of long term agreements between PCG/Ts and hospital trusts. Particular attention will be given to the relationship between changes in hospital activity and subsequent changes in expenditure by PCG/Ts.
- explore how additional pressures from regulations and incentives affect GP referrals and decision making.

The study will use a mix of both qualitative and quantitative techniques: a postal survey of PCG/Ts; semi-structured interviews with contracting managers; observations of contracting meetings; and in-depth interviews with GPs from selected PCG/Ts.

Ideally we will select and study different models of PCG/T (for example, capitation gainers with weak control mechanisms compared to capitation losers with a strong degree of budgetary control), to determine how budgetary pressures affect GPs and referral activity. This is dependent upon being able to categorise PCG/Ts in this way, and upon PCG/Ts managing their budgets and having control systems in place, in the way we anticipate. It may be more appropriate to compare those with devolved budgets to those without. If it is not practical or possible to categorise PCG/Ts in these ways, we will develop referral scenarios for discussion with high and low referring GPs/practices, and try to elicit which of recent referrals would not have been undertaken in a tighter budgetary regime.

The research will cover three stages:

(a) Survey of PCG/T budgetary control systems (to explore how GPs are taking on the role of insurer).

We are about to begin a survey of PCG/Ts in the south west region. This will give us information about PCG/T characteristics, whether the PCG is a capitation gainer or loser, budgeting arrangements such as the mechanisms and incentives in place to encourage integration of clinical and financial roles, and to stay within budget. It will also explore the extent to which budgets and incentives are relevant at practice and individual GP level. Results should enable us to categorise PCGs according to strength of control mechanisms, difference between capitation target and current expenditure, and level of budgetary devolvement.

(b) Nature of relationships between activity and expenditure (LTAs, complexity and information impactedness).

PCG/T and hospital contract/finance managers will be approached for interview, and a number of contracting meetings observed. Particular issues to be addressed will be the problems of setting and changing contracts and prices when a large proportion of costs are fixed, and the impact of activity changes on expenditure – both within and across years. We will also attempt to find out what information purchasers and providers have available when agreeing contracts, the extent of information asymmetry, and their willingness to share data with each other.

(c) GPs responses to budgetary pressures, and effects on referrals (the success of the system in reducing moral hazard and coping with organisational failures in contracting).

GPs from PCG/Ts with contrasting levels of budgetary control and finance will be interviewed. We are not sure of the exact content and style of these interviews at the moment. These interviews may be based around recent

referral decisions, and be undertaken with GPs with high, low and average rates of referral in contrasting PCG/Ts. They will aim to determine how referral patterns would change under a greater or lesser degree of budgetary pressure, and, importantly, how GPs can be helped to make choices (and be satisfied with their decision making) within systems of budgetary control. Issues for consideration could include mechanisms for coping with uncertainty and stress, satisfaction with freedom to refer, and perceptions of ability to control own working environment and patterns of decision making.

The outcomes will be predominantly descriptive and include items such as mechanisms of budgetary control, the pattern capitation gainers and losers, the divisions (and degree of devolvement) of financial responsibility, types of long term agreements, the relationship between activity and expenditure, flexibility within LTAs with respect to time and fixed costs, and the impact (or perceived impact) of budgetary pressure on GPs, referrals and activity. Results will be discussed in relation to the theories outlined above.

SECTION 5: SOME POINTS FOR DISCUSSION

Our main purpose in putting this paper forward for discussion is to ask the question – how can we use economic theories to evaluate the financial arrangements in the NHS? Our aim is not to explore the options for achieving efficiency in the NHS (for example, competition between purchasers or provider regulation), but to take the current situation and assess its effectiveness within a conventional economics framework. This paper has suggested a possible framework for an analysis of the structure and impact of budgeting arrangements. We are now interested in whether these theories provide the right framework for analysis, and whether or not they are the only relevant theories. Are there others that are equally appropriate?

The NHS is not a free market; it is subject to many of the causes of market failure, in particular, a small number of suppliers and no competition between purchasers. It is highly regulated to compensate for these failures. In a fully competitive market, inefficient firms are forced into dissolution or bankruptcy. Even if the NHS did function within a fully competitive market, firms (hospitals and PCG/Ts) would not be allowed to fail. Where there are inefficiencies that in a free market would result in bankruptcy, the government steps in with a ‘solution’. Both in its attempt to control moral hazard and costs, and to contract efficiently, the NHS is sheltered from the rigours of the free market, and as such from the market as judge of efficiency. Many economic theories are founded, in one way or another, upon the principals of a free market, which ensures economic efficiency without regulation. Most market failures are addressed in an attempt to eliminate or regulate against them in order to return to a competitive (or quasi-competitive) market. Given that the NHS will never be permitted to ‘fail’, can these market-based theories be used to address efficiency in the NHS?

We are interested in views and comments on the following areas:

- Using these theories to explore clinical and financial integration in the new NHS.

- Other theories that could be used to analyse budgeting systems in the NHS.
- Empirical work on transactions costs that has taken place in regulated markets where firms are not permitted to fail.
- Conflicting incentives for GPs as insurers – the desire to provide the best treatment for individual patients and the practice population, versus the need to control costs within a capitation allowance.
- Judging policies that pull in opposite directions (simultaneous attempts to reduce moral hazard and transactions costs).
- Interdependency - who is dependent upon whom in the purchaser-provider relationship?

REFERENCES

Arrow, K.J. (1963) Uncertainty and the welfare economics of medical care. *American Economic Review* 53, 941-973.

Bannock, G., Baxter, R.E. and Davis, E. (1998) *The Penguin Dictionary of Economics*, 6 edn. London: Penguin Books.

Baxter, K., Bachmann, M.O. and Bevan, G. (2000) Primary care groups: trade-offs in managing budgets and risk. *Public Money and Management* 20, 53-62.

Dawson, D. and Goddard, M. (1999) Long term contracts in the NHS: A solution in search of a problem? *Health Economics* 8, 709-720.

Department of Health. (1997) *NHS priorities and planning guidance 1998/99*. EL(97)39. Leeds: NHS Executive

Department of Health. (1998a) *The New NHS Modern and Dependable. Establishing Primary Care Groups*. Leeds: NHSE. (HSC 1998/65).

Department of Health. (1998b) *The New NHS Modern and Dependable. Developing Primary Care Groups*. Leeds: NHSE. (HSC 1998/139)

Dowie, R. (1983) *General practitioners and consultants: a study of outpatient referrals.* London: King Edward's Hospital Fund

Evans, R.G. (1987) Public health insurance: the collective purchase of individual care.*Health Policy* 7:115-134.

Fertig, A., Roland, M., King and Moore, T. (1993) Understanding variation in rates of referral among general practitioners: are inappropriate referrals important and would guidelines help to reduce rates? *British Medical Journal* 307, 1467-1470.

Hillman, A.L. (1991) Managing the physician: Rules versus incentives. *Health Affairs* :138-146

Hutchinson, A. (1993) Explaining referral variation. *British Medical Journal* 307, 1439-1439.

Janis, I.L., Mann, L. (1977) *Decision making: a psychological analysis of conflict, choice and commitment.* New York : Free Press

Mahon, A., Stoddart, H., Leese, B., Baxter, K. (1998) *How do total purchasing projects inform themselves for purchasing?* London: King's Fund.

Mays, N., Goodwin, N., Bevan, G. and Wyke, S. (1997) *Total purchasing. A profile of national pilot projects*. London: King's Fund Publishing.

Mays, N., Goodwin, N., Malbon, G., Leese, B., Mahon, A. and Wyke, S. (1998) *What were the achievements of total purchasing pilots in their first year and how can they be explained?* National Evaluation of Total Purchasing Pilot Projects Working Paper. London: King's Fund Publishing.

Mays, N., Wyke, S., Malbon, G., Goodwin, N. (eds) (forthcoming) *Can general practitioners purchase health care? The total purchasing experiment in Britain*. London: King's Fund Publishing.

NHS Executive. (1994) *Developing NHS purchasing and general practitioner fundholding.* (EL(94)79) Leeds: NHS Executive

Posnett, J., Goodwin, N., Griffiths, J., Killoran, A., Malbon, G., Mays, N., Place, M., Street, A. (1998) *The transaction costs of total purchasing*. London: King's Fund Publishing.

Secretaries of State for Health, Wales, Northern Ireland and Scotland. (1989) *Working for patients*. London: HMSO.

Secretary of State for Health. (1997) *The New NHS. Modern and Dependable*. London: HMSO.

Wilkin, D. and Smith, A. (1987) Explaining variation in general practitioner referrals to hospital. *Family Practice* 299:304-308

Wilkin, D. Patterns of referral: explaining variation. (1992) In Roland M and Coulter A. *Hospital Referrals*. Oxford: Oxford University Press

Wilkin, D., Gillam, S., Leese, B. (eds) (2000) *National Tracker Survey of primary care Groups and Trusts. Executive Briefing. Progress and Challenges 1999/2000.* Manchester: University of Manchester, NPCRDC

Williamson, O.E. (1975) *Markets and hierarchies: analysis and antitrust implications*. New York: Free Press.

Williamson, O.E. (1985) *The economic institutions of capitalism.* New York: Free Press.

Wright, J. and Wilkinson, J. (1996) General practitioners' attitudes to variations in referral rates and how these could be managed. *Family Practice* 13, 259-263.