Pricing in the English NHS quasi market: a national study of the allocation of financial risk through contracts

Pauline Allen* and Christina Petsoulas
Department of Health Services Research and Policy,
London School of Hygiene and Tropical Medicine,
15/17 Tavistock Place
London WC1H 9SH

*Corresponding author
Tel: 0207 927 2460
Email: Pauline.allen@lshtm.ac.uk

Pauline Allen is Reader in Health Services Organisation and Head of the Department
Christina Petsoulas is a Research Fellow

Key words: pricing; financial risk; contracts; NHS; markets; healthcare

This research is funded by the Policy Research Programme of the Department of Health. The views expressed are those of the researchers and not necessarily those of the Department of Health.

Summary

We investigate how the formal national provisions for pricing in the NHS (which are a form of prospective payment, known as Payment by Results) are operationalised at local level. Transactions costs theory and existing evidence predict that actual practice often does not comply with contractual rules. Our national study of pricing between 2011 and 2015 confirms this and indicates that such payment systems may not be appropriate to address the current financial and organisational challenges facing the NHS.
Pricing in the English NHS quasi market: a national study of the allocation of financial risk through contracts

Introduction

As in all markets, the negotiation and implementation of contracts for healthcare allow buyers and sellers to exchange information and provide a framework for the allocation of financial risk, primarily through pricing. Pricing of healthcare in markets is problematic because of the complexity of health services and the concomitant asymmetry of information between providers of care and its purchasers (Arrow, 1963). Firstly, it is difficult for purchasers to ascertain the costs of providing care. Secondly, it is even more difficult for purchasers to measure all aspects of the quality of care. If providers are forced to compete in respect of prices in order to improve efficiency, decreases in prices may well have the effect of reducing unobserved aspects of quality (Zwanziger et al, 2000; Propper et al, 2008). Across the developed world, various countries have made attempts to address the problem of how to arrive at prices for healthcare which would encourage efficiency while protecting the quality of care (O’Reilly et al, 2012). This paper concerns attempts to change the pricing of healthcare in the English National Health Service (NHS) quasi market\(^1\) over the past decade, concentrating on how the formal national provisions for pricing are operationalised at local level. This focus is important because it is misleading simply to analyse the formal provisions alone, as actual practice often does not comply with contractual rules (Macneil, 1981).

The intention of introducing a quasi market into the NHS in 1990 was, inter alia, to improve efficiency by the use of competition between providers of care (DH, 1989; NHS ME 1990). It was envisaged that commissioners and providers of care would initially use local prices set by calculating the short run average total cost of every set of procedures, adjusted to allow for a six percent return on net assets (NHS ME, 1993; Dawson, 1994; Propper and Bartlett, 1997).

\(^{1}\) The NHS quasi market consists of tax funded purchasing of care by state actors on behalf of patients; and competition between providers of care which may be state owned or independent (Bartlett and Le Grand, 1993). It has had three major incarnations: 1) 1990 to about 2004, spanning the Conservative and early period New Labour regimes under which negotiated prices were expected (although fundholding was abolished in 1997); 2) 2004 to 2013 spanning the later New Labour regime after pricing reform using PbR until the coming into force of the HSCA 2012 under the Coalition government; and 3) 2013 to date when pricing is designed by a new national economic regulator, Monitor.
Price competition between providers of care was destined to form part of the operation of the quasi market (Propper et al, 2008).

In practice, for the first few years of the internal market, scant information was available about the costs of care, and most contracts did not contain prices in respect of individual episodes of care. Instead, most contracts took the form of either block contracts or cost and volume contracts (Bartlett and Le Grand, 1993; Rafftery et al, 1996). Block contracts amount to a fixed budget allocating the financial risks of over performance to the provider. Cost and volume contracts involve the setting of the volume of cases in advance, and may lead to additional payments where the target volume of cases is exceeded, thus mitigating the risk to the provider. At the same time contracts for limited volumes of some elective care were on a cost per case basis, which allocated the financial risk to the purchaser, as they could not cap their expenditure. These were made by individual GP fundholders (another form of purchasing that existed alongside health authorities, which were responsible for commissioning the majority of care).

Greater sophistication was introduced into pricing of acute services in 2004 by ‘Payment by Results’ (PbR) (Sussex and Street, 2004), and this can be regarded as an important development in the quasi market. This prospective payment system entails that health care is categorised into a series of predefined activities (called Healthcare Resource Groups, HRGs). The idea of PbR is to sharpen incentives, as each episode of care reimbursed (or lost to another provider) is charged at national tariff rates, which are average costs across the whole country. This is meant to improve provider efficiency by driving down the costs of those providers whose costs are above average costs. PbR allocates the risk of over performance to the purchaser, while the provider is at risk of losing income if patients are not treated in sufficient numbers. In principle, PbR should obviate the need to negotiate prices. However, economic and socio-legal theories of contracting generally, together with empirical evidence on contracting and pricing in healthcare, indicate that the allocation of financial risk is often handled differently from the stipulations of formal contractual provisions (Williamson, 1985; Petsoulas et al, 2011).

This paper reports a unique study of the changes in the NHS national pricing rules and the actual allocation of financial risk on the ground during the four financial years 2011/12 to 2014/15. This period is particularly interesting due to the regulatory changes. The regulatory framework for the NHS was altered by the coming into force of the Health and Social Care Act 2012 (HSCA) in April 2013. Under the HSCA, PbR was retained, and renamed the National
Tariff, and the responsibility for setting the level of the National Tariff was given to two bodies: Monitor (the economic regulator of the NHS market); and NHS England (NHS E), the body responsible for commissioning health care (either through other bodies, such as new local clinical commissioning groups of GPs, CCGs, replacing Primary Care Trusts, PCTs, or directly itself). The HSCA can be seen as introducing a second set of major changes to the quasi market. The national standard contract, which was introduced in 2007, was retained. The form of this contract, including rules concerning pricing, was subject to changes introduced by NHS E each year. Although the HSCA in many ways continued the direction of travel set by previous reforms under the New Labour government concerning the use of competition to drive up efficiency and quality of care (Allen et al, 2011), the HSCA significantly altered the regulation of competition in the NHS through the creation of an economic regulator (i.e. Monitor) and the clear extension of competition law (and thus the jurisdiction of the national competition authorities) to apply to the planning and provision of NHS services (Sanderson et al. under review)

It has been argued that the introduction of the new regime under the HSCA would have the effect of juridifying decision making, and thus removing the internal flexibility previously enjoyed by the NHS (Davies, 2013). The rationale for this was that, in order to produce a fair playing field between all types of provider (including independent ones), it was necessary for pricing rules to be transparent and applied equally to all providers. Prior to the introduction of HSCA, prices had been subject to flexibility in practice (Monitor, 2013). This paper will show that there continues to be a marked difference between the rules promulgated at national level and the allocation of financial risk by local actors. These findings are to be expected, given the well-known difficulties in contracting for healthcare, and the continuing influence of hierarchical factors (in particular the cash limited budget) on the NHS quasi market.

Pricing policy in the NHS quasi market

Before discussing pricing policy in the NHS quasi market, it is necessary to describe some aspects of that market. The salient point is that the total national budget for the NHS is cash limited, and the budgets allocated to those bodies commissioning care on behalf of patients are capped on an annual basis. Traditionally, commissioners kept within budget by making patients wait for non-urgent care (Bevan and Hood, 2006). Until the introduction of the NHS Foundation Trust status in 2004, all NHS Trusts were required to break even on their income and expenditure account, taking one year with another. NHS Foundation Trusts are permitted
to carry over certain levels of deficits and surpluses, agreed with the regulator (Health and Social Care Act, 2003).

As explained above, prior to the introduction of PbR, prices were meant to be set at cost. In fact there was little relationship between cost and price (Bartlett and Le Grand, 1994; Propper and Wilson, 1996; Propper and Bartlett, 1997). In practice, local pricing continued to be unsophisticated for years. It consisted mainly of agreeing that the same overall sum would be paid to the provider as in the previous year, adjusted to take account of any efficiency savings demanded at national level (and any large changes in volumes of activity).

One aim of PbR was to incentivise providers to undertake larger volumes of care to increase throughput and thus shorten waiting times. The payment of average cost for every HRG provided was thought to encourage hospitals to do more, as they would be paid at above marginal cost for additional work (assuming their costs were being brought down to near national average costs). However, Dawson (1994) argued that the local market structures were of small numbers of providers and high fixed costs, in which setting national prices for each procedure was inappropriate. It would work for markets characterised by a large number of sellers who are price takers, but not where a few large providers exerted significant market power and were able to restrict the supply of quantity and quality with little fear of new entry by competitors. Although Dawson was discussing the pricing policy under the ‘Costing for Contracting’ rules (NHSME, 1993), the point is also relevant to PbR.

Despite a large increase in its use over the past decade, PbR does not apply to all activity. Some activity in acute hospitals is not subject to a National Tariff price, and is still subject to local agreement. It has proved difficult to create HRGs, and thus National Tariff prices, for care which is less episodic than acute hospital stays. There are no National Tariff prices for community health services (CHS) or, as yet, mental health services (MH) (Monitor and NHS E, 2014).

Research has been undertaken, both in respect of the English NHS and other health care systems (e.g. Street and AbdulHussain, 2004; Kobel et al, 2011) concerning the best way to set prices in prospective payment systems resembling the PbR system in the NHS, and also into the effects on provider behaviour when these systems are introduced (e.g. Farrar et al, 2011). However, there has been very little research about how these pricing rules are operationalised at local level (Petsoulas et al, 2011 being the most recent example) and none since the introduction of the HSCA. This paper aims to fill that gap.
Theoretical framework

In order to understand how pricing and allocation of financial risk in NHS contracts is likely to be undertaken in practice, it is necessary to understand how contracts for healthcare operate. The relevant theoretical frameworks are new institutional economics (Williamson, 1985) and socio-legal (Macneil, 1981). A key concept is whether contracts are capable of being complete. Completeness means that the contractual document covers all eventualities. Contracts for health care are unlikely to be complete because health care has a number of features which mean that the transaction costs (TCs) are likely to be high. TCs result from imperfect information, either about the other party involved in the exchange (asymmetric information) or about the future (uncertainty). Imperfect information means that it is costly to enter into contracts, since the parties will have to incur the costs of negotiating and writing contracts. It also makes it costly to monitor, enforce and renegotiate contracts. Health care is characterised by high levels of uncertainty and asymmetric information (Arrow, 1963). Parties to long-term contracts often do not plan and specify their contractual relationships completely (Macaulay, 1963) and socio-legal and economic theories suggest that this might be an efficient strategy to reduce TCs. In these circumstances, relational contracts might evolve and permit efficient trade (Macneil, 1981). In relational contracts, adjustments are made to the initially agreed terms during the course of the contractual relationship to deal with unforeseen contingencies (Vincent-Jones, 2006). Risk can be managed by the parties as events arise (when co-operative strategies can be developed) (Sabel, 1991).

This theoretical framework leads us to hypothesise that, despite the fact that the NHS national standard contract set out clear rules for the allocation of financial risk through pricing of healthcare, these contractual rules may not have been followed in all cases. These circumstances are likely to apply in relation to some of contracts for healthcare in the NHS than others. Although PbR appears to be more ‘complete’ in contractual terms, as prices are fixed nationally, volume of activity is not. Thus, in the context of fixed local commissioning budgets, it is possible that financial risk will need to be managed in the local health economy in ways not foreseen by the PbR regime. Where pricing is on the basis of block contracts, the contracts can be seen as ‘complete’ in relation to total expenditure, which may be the most important parameter in the NHS.

Our study
During 2012 to 2014 we undertook a study of the pricing rules in the NHS national standard contract in the financial years 2011/12 to 2014/15. We examined both the changing contractual provisions and the behaviour of contracting parties at local level.

**Design and methods**

The project consisted of three aspects:

1. Detailed analysis of each year’s standard NHS national contract from 2011/12 to 2014/15.
2. Two national telephone surveys of commissioners in 2012 and 2014 to find out what pricing mechanisms were being used in formal written contracts, and how they were implemented (or not). We were able to interview 23 PCT commissioners in 2012, which amounted to 15% of PCTs and a further 25 CCG commissioners in 2014, which amounted to 13% of CCGs. They were spread out across England in rural, suburban and inner city areas.
3. A series of three in depth case studies of three local health economies, looking at the contractual relationships between commissioning organisations and their providers of acute, mental health and community healthcare. Interviews of 27 contracting personnel in commissioners and providers (4 with CSU personnel, 6 with Acute Trust personnel, 15 with CCG personnel and 2 with a combined Acute and Community Care Trust personnel); observation of 21 contracting meetings (13 in Acute Care Trusts, 3 in combined Acute and Community Care Trust, 4 in combined Community and Mental Health Trust, and 1 PCT internal contract strategy meeting) and analysis of local documents were used. This allowed us to gather in depth and contextual information not available from the surveys.

This triangulated approach enabled us to put together a broader and more reliable picture of our findings: the surveys helped us put the case studies within the overall national context, whereas the case studies enabled us to pursue our research questions in greater depth.

Data analysis was conducted with the help of the qualitative research software NVivo. The authors agreed the main themes derived from the research questions, the literature on contracting, and additional themes suggested by the data.
**Findings**

**Provisions of the national standard contracts**

The relevant provisions of the national standard contracts remained relatively stable during the research period until 2014/5 when greater local financial flexibility was permitted.

For acute services the contract provided for the use of both national tariff (formerly PbR) prices and the negotiation of local prices in respect of care which was not covered by it. Although the principle behind the use of PbR was that providers should be paid for every episode of care delivered, the 2011/2 national contract included limits on activity which would be reimbursed. Commissioners could refuse to pay for more activity than had been forecasted. Since 2012, contractual provision was deleted because the then economic regulator, the Cooperation and Competition Panel, ruled that commissioners could not place a cap on activity, as it restricted patient choice. Nevertheless, each year’s contract provided that emergency admissions exceeding a local baseline figure from 2008/09 would only be reimbursed at 30% of tariff.

The 2014/15 contract, for the first time, contained provisions specifically designed to allow the parties greater flexibility in pricing. ‘Local variations’ were designed to allow adjustments to prices or currencies to facilitate significant service redesign or reconfiguration. ‘Local modifications’ were allowed in the case of unavoidable higher local costs. Moreover, for 2014/15 the contract allowed the parties to vary the base line figure over which emergency activity would be reimbursed at 30%.

The standard contract did not contain pricing rules for mental health and community services, and these were negotiated locally in the form of block contracts.

**Negotiation, monitoring and enforcement of contracts**

The information collected in the surveys and the case studies were similar, which allows us to report them together, and to draw conclusions that we can be confident are likely to apply across the English NHS. We did not observe any salient differences between commissioning areas (or case study sites) which could be attributed to their geographical locations. We did not collect sufficient data in the surveys to allow us to correlate personal relationships with differences in handling pricing.
In respect of contracts between CCGs and NHS acute trusts, the allocation of financial risk outside the framework of the various formal pricing mechanisms was striking. Most of the contractual relationships between NHS acute providers and commissioners were characterised by the use of general annual financial settlements outside the terms of the contract. Whatever detailed financial provisions had been agreed and implemented during the course of the year, a final overall agreement was made at year-end which did not adhere strictly to the contractual provisions. It was not always possible for commissioners to pay the full contractually designated amount for activity undertaken, as their budgets were insufficient. This appeared to be increasing over time, with more commissioners reporting not being able to afford to pay the full amount for the level of activity provided by the time of the second survey in mid 2014.

Year-end deals were seen as pragmatic and inevitable in the context of the NHS.

Well, this is, you see, this is where PbR shows its - how can I say? – its limitations. At the end of the day, health economies need to be in balance, ... it’s in nobody’s interests to bankrupt any of the parties associated with the relevant health economy, and that’s almost a diktat in terms of public policy, okay? (Case Study B, Commissioning Consultant, CSU)

But there was variation between the commissioners who responded. Several confirmed that they simply followed the contract and paid for all activity undertaken, even if it was more than they expected. These were areas where there was sufficient money available to commissioners for them to afford to do so.

The new provisions in the 2014/15 contract allowing for variation in allocation of financial risk were used by some CCGs, although by no means the majority. This was because other commissioners continued to make informal arrangements to allocate risk at the end of the financial year, and did not see any point in formalising these by notifying national regulators.

Three commissioners had used the new provisions allowing for flexibility in national tariff prices (‘local variations’). One had submitted variations in relation to six tariff areas to Monitor. The provider is a tertiary hospital and the new approach was needed due to the fact that NHS E had taken over commissioning of some of the services. Variations to national tariffs were needed to keep within the CCG budget. Another commissioner had agreed a lower local tariff with its provider if a patient was admitted for under two hours, which was sent to NHSE and Monitor for agreement.
A few parties ignored the national instructions on pricing and agreed a block contract for all acute services at the beginning of the financial year in 2012 and 2014. This was because they could foresee that the local health economy could not bear the cost of PbR, with unlimited financial exposure for the commissioners. Not all of these agreements were reported to Monitor. By 2014, some commissioners had reported this to NHS E as ‘local modifications’. One commissioner explained that they might be forming an integrated care organisation, and this, in addition to the financial risks to the acute provider posed by transferring money to the local Better Care Fund (i.e. out of the acute sector), required different payment provisions

As a local community with financial pressures that each organisation is under, we agreed that it is financially less risky to have a block contract. We submitted like an 'excusing note' to NHSE explaining why we did that. [...] We [may] have in our area an Integrated Care Organisation. That was part of the reasoning for looking at a different way of contracting this year to assure our providers, [...] was to ensure that our acute Trust had that guaranteed income, so we were able to do the long term financial modelling to allow an integrated organisation to proceed.

There was a conflict between different policy objectives: financial solvency for each NHS provider trust on the one hand; and reconfiguring services so that more care is provided outside hospital on the other. One commissioner director of finance pointed out:

We did look, this year, at trying to move the Acute Trust contract away from PbR to more caps and collars [i.e. block] contract base, but, again, the Trust Development Agency [i.e. national regulatory agency] were not supportive with that approach. Because the Acute Trust is financially challenged, they didn’t want them to do that. But the only way we’re going to be able to redesign services and reduce the secondary care footprint is to try and take some of the discussion away from finance and prices to an agreement where we have a set level of income for a set level of activity and we collectively work together to step it down because the PbR system and the contracting is slightly perverse in that why would the Acute Trust want to step down activity when it gets paid for what it delivers? (CS A, Director of Finance, CCG)

Despite the increased flexibility in the 2014/15 contract, most of the commissioners were still using the 2008/09 baseline for setting the point at which the 30% marginal rate for emergency admissions applied. Several had moved to later figures. One commissioner had done so because
there had been major changes in the configuration of acute services since 2008/09. Other commissioners had moved to later base lines, such as 2011/12, in one case after having been forced to do so after arbitration.

In some places, commissioners felt obliged to pay so called ‘non-recurring’ additional amounts to their NHS acute providers in order to help the hospitals balance their books at the end of the year. This was related to the need in some areas to facilitate the reconfiguration of local services, which might require transitional payments to support changes in service delivery in the short term. This could only occur in areas where commissioners had sufficient additional funds available.

We investigated how prices for non-tariff activity were agreed. In 2012 in most areas, prices paid the previous year were reduced by the current NHS-wide efficiency target. By 2014, a wider range of techniques was in use. In a few areas, there were attempts to benchmark local prices with those in other areas. And very rarely there were attempts to undertake more accurate costing exercises in respect of some of these services. These were undertaken in areas where there was particular concern that current prices were inaccurate. In addition, by 2014/15, increasing numbers of commissioners were insisting on agreeing a fixed sum in respect of these non PbR services – effectively another block contract. This was related to the poor financial situation in those areas, where money had to be saved in order to try to stay within commissioning budgets.

In contrast to the behaviour with NHS providers, all commissioners reported being able to pay independent providers of acute services in accordance with the PbR rules. This may have been due to the fact that there were insignificant volumes of such activity, so the financial viability of the local health economy was not threatened.

Block contracts were used in respect of CHS and MH services, in accordance with national contracting rules. Volumes of activity were monitored throughout the year and any over or under performance informed the setting of the next year’s block amount. Although the block contract limited financial risk for commissioners, it also impeded moving the activity from one type of provider (e.g. acute Trust) to another (e.g. community health care setting), which was
a national policy priority. In order to make such moves easier, the participants suggested that national tariffs were needed in respect of CHS so that money would follow the patient.2

There's something about as we move forward, we recognise that we've got to transform services. So essentially, less direct acute provision and more alternative provision, either a primary care, community care or social care setting. Now, in order for that to happen, it's simpler that the money follows the patient. So from that perspective, I think, going forward, a tariff type contract works better, because the money's more easily moved. (CS C, Director of Finance, CCG)

Context for contracting

The context in which the contractual relationships we studied took place was very important. First, the case studies demonstrated that personal relationships between staff were a vital element in facilitating effective contractual relationships. The degree of flexibility required could only be achieved where these worked well.

Secondly, the increasing financial stringency affecting the whole NHS during the course of the study had an effect on the way in which the contracts could be used at local level. As less money was available, it became increasingly difficult for commissioners to adhere to the national tariff rules.

Thirdly, in the later years of the study national and local policies entailed major service reconfigurations at local level, mainly aimed at shifting resources from acute care to social care and CHS. The contractual pricing rules impeded this.

Discussion

The limitations of the study should be noted. The response rates for the two national telephone surveys were not very high, mainly due to the national organisational changes in respect of commissioning being carried out. Secondly, it was not possible to interview the same people in the two consecutive surveys, as staff had moved on. But the use of the case studies allowed us to observe some trends.

2 PbR pricing does not take account of trusts’ costs structures, so that the amount of money lost to a provider when the money ‘follows the patient’ is greater than the savings made by the provider due to not treating that patient.
The extent to which the evidence about contracting for healthcare has been consistent over the past two and a half decades is striking. It remains the case (as found earlier by e.g. Bartlett and Harrison, 1993; Petsoulas et al, 2011) that allocation of financial risk is often dealt with outside the formal structures of the contractual document.

The informal flexibilities used by local staff have been incorporated into the formal system by the introduction of flexibilities into the 2014/15 standard contract. This is a new development for NHS contracting, and appears to be due to the new role of Monitor in setting transparent pricing rules. The changes can be understood as evidence that fixed national tariff prices are not appropriate in the NHS quasi market, as Dawson (1994) predicted. And it appears that there has been some increase in the divergence from the formal pricing rules over the past three years. The evidence from the study indicates that this is probably due to two factors: firstly the fact there has been increasing financial stringency in the NHS as a whole, while levels of activity have not diminished; and secondly, there has been an increasing national policy focus on reconfiguring local services. This latter factor requires greater flexibility in the allocation of resources in local health economies.

As time goes on, it has become increasingly difficult for many local commissioners fully to use the terms of the national pricing rules to regulate their relationships with local providers. Although this was recognised to some extent in the 2014/15 contract, Monitor and NHS E will be reconsidering pricing in the NHS. This is particularly important in the light of The Five Year Forward View (NHS E, 2014), which indicates that different configurations of providers should be explored. The current pricing rules are insufficiently flexible to facilitate these changes. Monitor (2014b) has stated that the NHS should be moving towards a blended payment system including ‘activity-based, outcomes-based and capitated payment approaches’ (p. 7). It is not clear how easy it will be to reconcile this wider range of pricing mechanisms with both the current system of contracting and the organisational changes which the Five Year Forward View envisages. The latter appear to be more aligned with notions of hierarchy and planning than with market structures that use contracts with nationally set tariffs as a way to allocate financial risk between autonomous parties.

Thus, we can conclude that not only is pricing in the English NHS more complex in practice than the official rules would indicate, but also that the current nationally determined prices set by the National Tariff are not appropriate in all circumstances for which they were designed. As the NHS struggles radically to reconfigure services, it is necessary to
reconsider the appropriateness of a wider range of pricing mechanisms to facilitate moving care out of hospitals. One promising approach is capitation (Monitor, 2014b).

References


Department of Health (1989) Working for Patients; the Health Service in the 1990s Cmnd 555 (HMSO, London)


Health and Social Care Act 2003

Health and Social Care Act 2012


Monitor (2013) Local price setting and contracting practices for NHS services without a nationally mandated price: A research paper

Monitor and NHS England (2014) National Tariff Payment System engagement workshops (mental health) proceedings

Monitor (2014a) 2014/5 National Tariff Payment System

Monitor (2014b) Reforming the Payment System for NHS services: supporting the Five Year Forward View

National Health Service Management Executive (1990) Contracts for Health Services: Operating Contracts EL(90) MB/24

National Health Service Management Executive (1993) Costing for contracting FDL (93) 51 Annex A


Sanderson, M., Allen, P. and Osipovic, D. ‘The regulation of competition in the NHS – what difference has the Health and Social Care Act 2012 made?’ (under review)


