Optimal Price-Setting in Pay for Performance Schemes in Health Care

5 October 2015

Abstract

The increased availability of process measures implies that the quality of health care is in some areas *de facto* verifiable. Optimal price-setting for verifiable dimensions of quality is well-described in the theoretical literature on incentive design. We seek to narrow the large gap that remains between actual price-setting behaviour in Pay for Performance schemes and the incentive design literature. We present a stylised model for hospital price setting for process measures of quality and show that optimal hospital prices should reflect the marginal benefit of the expected health gains, the weight given to patients' benefit relative to profits, and the opportunity cost of public funds. Based on published estimates, we derive the optimal prices for three measures of quality that have been incentivised in the English National Health Service since April 2010 (treatment in an acute stroke unit, rapid brain imaging, and thrombolysis with alteplase). We then compare the optimal prices with the actual prices offered to hospitals in England under the Best Practice Tariffs scheme for emergency stroke care.

Keywords: Pay For Performance; provider behaviour; optimal price-setting.

JEL: I11, I18

1 Introduction

Pay for performance (P4P) schemes link provider payments to performance indicators of quality. They receive much attention from both policy makers and scholars. The empirical evidence on the effectiveness of P4P is mixed. However, there is an emerging consensus that the key to effective P4P schemes is in their design elements (Epstein, 2012; Maynard, 2012; Roland, 2012). These design elements include who to pay, what to pay for, the criteria for bonuses or penalties and how much to pay for each unit of increase in quality (Ryan, 2009).

The size of the performance payments (i.e. the price, or the 'power' of the incentive scheme) is obviously critical, but has received surprisingly little attention in the applied literature. It has been treated mainly as an empirical question in ex-post evaluations of implemented schemes rather than as a key parameter that could be set optimally on the basis of economic theory. In an early review of the effects of P4P, Petersen et al. (2006, p. 269) stated that the "[s]ize of the bonus is *probably also* important [our emphasis]" and suggested that "the lack of effect or small effect in some studies *may* include the small size of the bonus [our emphasis]" (See also Cashin et al., 2014).

Normative statements about the size of incentive payments in the literature on design choices have been extremely vague. For example, Conrad and Perry (2009 p. 361) suggested that the optimal incentive size should "follow the Goldilocks principle: not too little, but not too much", while Eijkenaar (2013 p. 124) stated that "[a]ll else equal, the higher the revenue potential for providers, the larger their response and the impact on performance, up to a certain point".

Empirically, the size of incentive payments is often measured as a percentage of provider income. For example, the largest hospital P4P scheme in the US (the Premier Hospital Quality Incentives Demonstration (HQID) program) set bonuses and penalties as percentages (1-2%) of Medicare revenue (Das and Anderson, 2007). Similarly, the English adaptation, Advancing Quality, set bonuses of 2-4% of revenue for the associated activities (Sutton et al., 2012), and the Commissioning for Quality and Innovation framework determined that 0.5% in the first year rising to 2.5% of provider income be tied to performance on locally selected performance indicators (Kristensen et al., 2013). In their review of the literature, Conrad and Perry (2009) found that incentive sizes in the US varied between 2-9% of provider income.

The theoretical literature on incentive design suggests that setting incentive payments relative to revenue is not appropriate. Rather, as we emphasize in this paper, a regulator should focus on the expected health gains of improved performance and the costs of these performance improvements when setting payments for performance. An extensive theoretical regulation

literature has investigated how to set optimal prices when health care quality is verifiable (Chalkley and Malcomson, 1998a, 1998b; Ellis and McGuire, 1986; Holmstrom and Milgrom, 1991; Kaarboe and Siciliani, 2011; Laffont and Tirole, 1993). The key insight is that price should be set equal to the marginal benefit of health care (discounted downwards for the opportunity cost of public funds and for altruistic motives of the provider; Ellis and McGuire, 1986; Chalkley and Malcomson, 1998a and 1998b). Given the large increase in availability of indicators of quality, the assumption that many dimensions of quality are verifiable is not unreasonable in many areas of care (Eggleston, 2005; Goddard et al., 2000; Kaarboe and Siciliani, 2011). If quality is verifiable, it is still the case that the optimal price should be basically set equal to the (adjusted) marginal benefit of the verifiable quality (Kaarboe and Siciliani, 2011).

However, the literature on optimal price-setting is purely theoretical, and no attempt has been made to compare the derived optimal price solutions with incentive schemes implemented in practice. This may explain why the optimal price-setting literature appears to have been neglected by the practical P4P literature.

The aim of this paper is to make a first serious attempt at bridging the gap between the theory and the applied literature. We provide a theory model of hospital price setting for P4P schemes, and compare it with the actual implementation of such a scheme. Our example of actual price-setting behaviour is the Best Practice Tariffs (BPTs) hospital scheme for emergency stroke—a national P4P scheme introduced in the English NHS from 2010/11. BPTs are now the main vehicle for supplementing activity-based tariffs with performance related payments in the English NHS. We therefore build a theoretical model whose key assumptions match this scheme closely. The main feature of our model is that hospitals' optimal prices should reflect the marginal benefit of the health gain associated with the incentivised dimensions of care.

For our implementation, we searched the published literature for estimates of the health gains associated with the incentivised dimensions of care (treatment in an acute stroke unit, rapid brain imaging, and thrombolysis with alteplase). Using a monetary social value of a Quality-Adjusted Life Year of £50,000 (previously used by the English Department of Health), we show how the optimal prices depend on the assumed weight given to patients' benefit relative to profits, and the opportunity cost of public funds.

Our application relates to patients affected by stroke. Stroke is the second most common cause of death in the world, causing 10-12% of deaths in the western world (Donnan et al., 2008). The estimated total societal costs of stroke in the UK is £9 billion per year, including approximately £4 billion direct treatment costs, meaning that stroke treatment costs make up 5% of total UK

NHS costs (Saka et al., 2009). Timely and appropriate treatment of stroke is thus important both from an individual and a societal perspective.

The framework presented here can be used to improve scholars' and policymakers' thinking about price-setting for quality. Our analysis highlights the importance of setting prices based on expected benefits, not only costs, and the weight which hospitals assign to patients' benefits relative to profits. A key policy implication is that current incentive schemes appear either low-powered or imply a relatively high hospitals' weight given to patients' benefits relative to profits. The interpretation that current schemes are low-powered is consistent with a recent review of existing P4P schemes which suggests that current hospital schemes (that pay up to 5% of the revenues) have achieved very limited or no improvement in incentivising process measures of quality (Cashin et al., 2014, p. 86).

The paper is organised as follows. In Section 2 we describe the BPT incentive scheme for emergency stroke care in English hospitals. In Section 3 we provide a theory model for optimal hospital tariff setting in a context similar to BPTs, i.e. aimed at incentivising processes of health care for emergency stroke treatment. In Section 4 we simulate the theoretical model numerically and compare the result with the actual price set in the BPT incentive scheme. We end the paper with a discussion of our key results and venues for future research.

2 Background

In this section we review the information needed to setup a model that matches the key assumptions of the English BPT scheme including the financial incentives for quality before and after the scheme (section 2.1), the verifiability of emergency stroke care quality (Section 2.2), and provider performance on the incentivised dimensions of care before the BPT scheme (Section 2.3).

2.1 Financial incentives for quality in the English NHS

Hospital care in the English National Health Service (NHS) is provided through an internal market in which 211 Clinical Commissioning Groups (CCGs) have the responsibility of purchasing health care for their populations from 160 Acute NHS Trusts (henceforth termed "providers").

The Payment by Results (PbR) framework links hospital reimbursement to activity through a fixed tariff per admission. Hospital activity is classified into a manageable number of homogenous, clinically-meaningful healthcare resource groups (HRGs) – the English equivalent of diagnosis related groups (DRGs). The tariff or price paid per HRG is usually set equal to the

national average cost of treating patients in a given HRG. The Best Practice Tariffs analysed in this paper represent a deviation from this rule (Department of Health, 2012a).

Reimbursing hospitals on the basis of average costs has been shown theoretically to provide incentives for efficiency through cost reductions (Shleifer, 1985). If patient demand does not reflect quality it has been argued that these cost containment incentives may adversely affect the level of quality provided (Chalkley and Malcomson, 1998a). If quality is verifiable, however, a regulator can achieve the desired level of quality through contracting.

The Best Practice Tariffs (BPT) analysed in this paper can be seen as an attempt to include verifiable dimensions of quality into the agreements between third-party purchasers and providers. BPTs were first introduced in the English NHS from April 2010 for four procedures. This was extended in the following years, and from 2013/14 BPTs cover more than 50 procedures (Department of Health 2013). BPTs are tariffs that have been "structured and priced to adequately reimburse and incentivise care that is high quality and cost effective" (Department of Health 2013, s.61). This aim is pursued using a number of different pricing regimes, of which we will focus on the payment regime known as *Paying for Best Practice.* This regime is similar to the most common type of P4P today, in which health care providers are paid on the basis of their performance on process measures of quality that are assumed to be linked to better outcomes.

In the financial year (FY) 2013-14 the *Paying for Best Practice* model was used for four different conditions: emergency stroke care; diabetic ketoacidosis and hypoglycaemia; fragility hip fracture; and transient ischaemic attack (TIA/mini-stroke[1]). This pricing model consists of a base payment for all admissions, plus one or more additional payments conditional on performance. In this paper we focus on the performance indicators for emergency stroke care.

2.2 The verifiability of emergency stroke care quality

Townsend et al. (2012) estimated that there were a total of 125,945 stroke incidences in England in 2009, while NHS England (2013), using a narrower definition of stroke, estimated that the median provider admitted about 400 stroke patients in 2012-13.

Stroke has been described as a “brain attack” and is caused by a disturbance in the blood supply to the brain. The most common type of stroke is ischaemic stroke (representing approximately 80% of emergency strokes). Ischaemic strokes are caused by a blood clot narrowing or blocking the blood supply to the brain leading to the death of brain cells due to lack of oxygen. The less

[1] A TIA is a temporary disruption of the blood flow to the brain and cause similar symptoms as emergency stroke, but the symptoms resolve within 24 hours.

common haemorrhagic stroke is caused by a bursting of blood vessels leading to damaging bleeding into the brain (Department of Health, 2007a).

Untreated stroke typically leads to a loss of 1.9 million neurons (brain cells) per minute, so stroke treatment should be initiated as early as possible (Department of Health, 2007a). The appropriate treatment of stroke depends on whether the stroke is ischaemic or haemorrhagic, which can be determined by an experienced health care professional on the basis of either a computed tomography (CT) scan or magnetic resonance imaging (MRI). If the stroke is ischaemic, within 4.5 hours from the stroke, an attempt can be made to dissolve the blood clot medically in a procedure known as thrombolysis with alteplase. It is of key importance that alteplase is *not* administered to patients with an haemorrhagic stroke, in which case the treatment could be fatal.

There is good clinical consensus on what constitutes high quality care for emergency stroke patients. In England, National Clinical Guidelines for Stroke were first published in 2000 and have recently been published in their fourth edition (Intercollegiate Stroke Working Party 2012). The Department of Health published a National Stroke Strategy in 2007 (Department of Health, 2007a). The National Institute for Health and Care Excellence published a guideline for interventions in the acute stage of stroke and transient ischaemic attack (TIA) in 2008 based on clinical and economic evidence and expert consensus (National Collaborating Centre for Chronic Conditions 2008). This was backed up by publication of a quality standard in 2010 (NICE 2010) and a NICE pathway—a visual representation of the NICE guidelines and quality statements in the form of online interactive topic-based diagrams.

In addition, the verifiability of stroke care quality is high and increasing. Until recently, the quality of stroke care was monitored on a range of key indicators for samples of patients in biennial and quarterly reports (The National Sentinel Stroke Audit (NSSA) published from 1998 to 2010; The Stroke Improvement National Audit Programme (SINAP) from 2010 to 2012). The Sentinel Stroke National Audit Programme (SSNAP), for which data collection started in 2012, will provide a minimum dataset with process and outcome data for all stroke patients in England, Wales and Northern Ireland which include the indicators on the NICE quality standard and the NHS Outcomes Framework.

The detailed coverage of all patients means that emergency stroke care can now reasonably be assumed to be fully verifiable in some key quality dimensions which, in principle, allows for very detailed contracts to be written. Notably, the National Clinical Guideline for stroke contains recommendations for Clinical Commissioning Groups (CCGs) on how to purchase stroke care.

2.3 Best practice tariffs for emergency stroke

The BPT for stroke uses the high verifiability of stroke care quality to include quality in the contract arrangements between purchasers and providers. The BPTs for stroke are designed with the intention of supporting the key components of clinical best practice in the acute phase of the stroke following the recommendations of the NICE clinical guidelines and the National Stroke Strategy. Specifically the tariffs incentivise the treatment of patients in an acute stroke unit, rapid performance of brain imaging and administration of thrombolysis, if appropriate.

The tariffs are designed as a base tariff paid for all stroke patients irrespective of performance, and extra payments for a) rapid brain imaging, b) treating the patient in an acute stroke unit, and 3) alteplase (see table A.1 for a full description of the indicators). Alteplase was already paid for separately on top of the stroke tariff from 2008/09 and was not formally considered a part of the BPT scheme in the first two years of the programme. From 2012/13 the level of the separate alteplase payment was kept the same but considered a part of the BPT for emergency stroke.

2.4 Provider performance on the incentivised dimensions of stroke care quality

There is no set of indicators available that exactly match hospitals' performance on the indicators incentivised in the BPT scheme before and after its implementation. In the following, we describe performance on some indicators related to the incentivised dimensions of care. We do not attempt to evaluate whether improvements in performance after the introduction of BPT are attributable to the introduction of the P4P scheme. An earlier evaluation did not find this to be the case (McDonald et al., 2012).

2.4.1 Stroke care delivered in an acute stroke unit

According to the NSSA, in 2008, 59% of English stroke patients spend at least 90% of their time on a stroke unit (Royal College of Physicians 2011; Royal College of Physicians 2009). This proportion increased to 62% in 2010. Column 2 of Table 1 shows that this increase continued steadily over time. There is however, sign of a stabilisation around 85% towards the end of the period, which possibly reflects capacity constrains with the current level of acute stroke units. Note that the definition of stroke unit used for these data is broader than what is required to satisfy the requirement for the BPT (the BPT incentivises admission to *acute* stroke unit as defined in Table A.1).

2.4.2 Rapid Brain Imaging

According to the NSSA (Royal College of Physicians 2011; Royal College of Physicians 2009; Royal College of Physicians 2007), the percentage of patients who had a brain scan carried out

within 24 hours from arrival at hospital increased from 42% in 2006 to 59% in 2008 and to 70% in 2010. The percentage of patients who had a brain scan carried out within 3 hours increased from 9% in 2007 to 21% in 2008 and 25% in 2010. Columns 3—5 of Table 1 show some continued improvement in scanning times, but again, especially for a scan within 24 hours of arrival, performance seems to stagnate at around 90% of patients scanned within this time. Note however, that these numbers do not provide information as to whether all patients eligible according to the brain imaging guide incentivised by the BPT were scanned as quickly as possible.

2.4.3 Thrombolysis by alteplase

Column 6 of Table 1 shows an increase in the proportion of patients eligible for thrombolysis who are thrombolysed, although the increase was most rapid in the first three quarters. Due to the non-mandatory participation in SINAP, the estimates could be biased by selection into the SINAP audit on which these numbers are based.

Table 1: Hospital performance on indicators close to the quality dimensions incentivised by Best Practice tariffs

Financial year and quarter	% of patients that spend at least 90% of time on a stroke unit[a]	% of patients who receive a brain scan within		Median time (minutes) between arrival and first brain scan	% of patients eligible for thrombolysis who are thrombolysed
		1 hour	24 hours		
2011-12 Q1	80	25	82	218	44
2011-12 Q2	84	28	86	237	49
2011-12 Q3	85	31	90	208	60
2011-12 Q4	83	31	91	196	58
2012-13 Q1	86	34	91	157	64
2012-13 Q2	87	36	92	135	63
2012-13 Q3	86	37	92	129	65

Source: The Intercollegiate Stroke Working Party (2013) and NHS England (2013).
Note: a) These data relate to the broad definition of stroke units, not the narrower definition of *acute* stroke units used for the BPT (See table A.2 for definitions).

3 A model for optimal price-setting for verifiable quality

In this section we present a model of optimal price-setting in a context similar to the BPT for emergency stroke. We focus on price-setting for verifiable process indicators of quality for stroke emergency care. We first describe in Section 3.1 a model of *hospital* decision making and derive the providers' optimal incentivised qualities for a given incentive scheme. In Section 3.2 we derive the optimal prices set by the regulator.

Throughout the analysis we assume that it is the hospital as a whole that takes decisions over quality choices, and these are not chosen unilaterally by the doctors. Hospitals are complex organisations and the decisions made are the outcome of the involvement of several parties, including hospitals' managers concerned about costs and potential deficits or surpluses (see fuller discussion below).

3.1 The hospital

We assume that each hospital receives a basic tariff p^0 for every patient admitted with a stroke. The number of patients admitted with a stroke is assumed to be exogenous since stroke requires emergency treatment.

Hospitals provide four different dimensions of services to stroke patients where the type of care is denoted with $i=0,1,2,3$. We assume that three out of the four dimensions are incentivised by the regulator (when $i=1,2,3$). All patients receive the basic care (service 0). The hospital receives three additional payments p^1, p^2 and p^3 for three incentivised dimensions of care: rapid brain imaging (service 1), thrombolysis with alteplase (service 2) and delivery in an acute stroke unit (service 3). The number of services provided in each dimension of quality is $N^i \leq N^0$. The number of services provided may in general differ from the number of patients since a patient may receive more than one service (as highlighted in more detail below).

Patient's benefit depends on the services received. There are potentially 8 possible combinations of types of care that a patient could receive: basic services only (type 0); basic services and rapid brain imaging (type 1); basic services and thrombolysis with alteplase (type 2); basic services and delivery in an acute stroke unit (type 3); basic services, rapid brain imaging and thrombolysis with alteplase (denoted with type 12); basic services, rapid brain imaging and delivery in stroke unit (denoted with type 13); basic services, thrombolysis with alteplase and delivery in an acute stroke unit (denoted with type 23); and basic services, rapid brain imaging, thrombolysis with alteplase and delivery in an acute stroke unit (denoted with type 123).

Alteplase administered to patients with an haemorrhagic stroke can be fatal. Hence, the drug should not be given without conducting rapid brain imaging first. We therefore ignore types 2 and 23.[2] We assume that patients' benefit from basic care is b^0 and that the additional benefit (on top of the benefit from basic care) from each of the five residual combinations of services is b^i with *i = 1, 3, 12, 13, 123*. Moreover, we assume that the additional benefit from delivery in an acute stroke unit is separable so that $b^{123} = b^{12} + b^3$ and $b^{13} = b^1 + b^3$.

Define n^i as the number of patients for each of the six possible combinations of services (with *i = 0, 1, 3, 12, 13, 123).* n^0 patients do not receive any of the three incentivised services. n^1 includes patients who receive brain imaging only. n^3 includes patients who are admitted in a stroke unit. n^{12} refers to the number of patients who are *considered* for alteplase treatment and therefore also require brain imaging. Similarly, n^{123} refers to the number of patients *considered* for alteplase in a stroke unit and also require brain imaging. Since alteplase can be administered only after brain imagining and is safe only to a subset of patients (who do not have an haemorrhagic stroke), we assume that only a fraction γ out of (n^{12} +n^{123}) are administered alteplase (and therefore have an additional benefit b^{12}) and for a fraction $(1-\gamma)$ alteplase is not administered since brain imaging revealed that patients had an haemorrhagic stroke. These patients have an additional benefit of b^1 (from brain imaging only).

The total benefit of patients treated, denoted with *B*, is:

$$B = n^0 b^0 + n^1(b^0 + b^1) + n^3(b^0 + b^3) + \gamma n^{12}(b^0 + b^{12}) + (1-\gamma)n^{12}(b^0 + b^1)$$

$$+n^{13}(b^0 + b^1 + b^3) + \gamma n^{123}(b^0 + b^{12} + b^3) + (1-\gamma)n^{123}(b^0 + b^1 + b^3),$$

which can be re-arranged as:

$$B = b^0(n^0 + n^1 + n^3 + n^{12} + n^{13} + n^{123}) + b^1(n^1 + n^{13})$$

$$+b^3(n^3 + n^{13} + n^{123}) + \left(b^{12}\gamma + b^1(1-\gamma)\right)(n^{12} + n^{123}).$$

We can further rewrite the benefit function as a function of the number of services provided N^i, defined above, instead of the number of patients treated (since recall each patient can receive more than one service). The expression simplifies to:

[2] This is a simplifying assumption. We could alternatively endogenise the possibility that alteplase is given to an haemorrhagic stroke. However, this would make the exposition of the model considerably more complex without adding any new insights. In equilibrium we would have two groups of patients with a corner solution where types 2 and 23 are never provided. Since the health penalties are incredibly high when alteplase is provided, any small level of patients' concern in our model would be consistent with this outcome. Moreover, since this is a basic mistake on the side of the doctor and easily verifiable, doctors could face severe personal penalties (such as a suspension) if this would happen.

$$B = b^0N^0 + b^1N^1 + b^3N^3 + \left(b^{12} + \frac{1-\gamma}{\gamma}b^1\right)N^2, \quad (1)$$

where $N^0 = n^0 + n^1 + n^3 + n^{12} + n^{13} + n^{123}$, $N^1 = n^1 + n^{13}$, $N^2 = \gamma(n^{12} + n^{123})$, $N^3 = n^3 + n^{13} + n^{123}$.[3] In words, every patient receives the basic care. N^1 patients receive and benefit from brain imaging without being considered for alteplase. N^3 patients benefit from delivery in a stroke unit. N^2 patients receive both brain imaging and alteplase.[4] There are also $\frac{1-\gamma}{\gamma}N^2$ patients who received brain imaging since potential users of alteplase but then do not receive it because brain imaging suggests that it is not appropriate for them. These patients still receive some benefits from brain imaging (which is equal to b^1).

The revenues of the hospitals, denoted by *R*, are a function of the services provided:

$$R = n^0p^0 + n^1(p^0 + p^1) + n^3(p^0 + p^3) + \gamma n^{12}(p^0 + p^1 + p^2) + (1-\gamma)n^{12}(p^0 + p^1)$$

$$+n^{13}(p^0 + p^1 + p^3) + \gamma n^{123}(p^0 + p^1 + p^2 + p^3) + (1-\gamma)n^{123}(p^0 + p^1 + p^3),$$

which can be re-arranged as:

$$R = p^0(n^0 + n^1 + n^3 + n^{12} + n^{13} + n^{123}) + p^1(n^1 + n^{12} + n^{13} + n^{123})$$

$$+p^2\gamma(n^{12} + n^{123}) + p^3(n^3 + n^{13} + n^{123}).$$

Analogously to the benefit function, we can further rewrite the revenues as a function of the number of services provided N^i (instead of patients treated):

$$R = p^0N^0 + p^1(N^1 + \frac{N^2}{\gamma}) + p^2N^2 + p^3N^3. \quad (2)$$

The cost function of each service i is $c^iN^i + K^i(N^i) + F^i$ with c^i>0, $K^{i\prime}(N^i) > 0$, $K^{i\prime\prime}(N^i) > 0$. F^i is a fixed cost (for example the fixed cost of setting up a stroke unit or the fixed cost associated to an MRI machine). c^i is a constant marginal-cost component (for example the unit cost of administering alteplase or unit cost of a CT scan). $K^i(N^i)$ includes monetary and non-monetary costs of providing the service. The increasing marginal cost assumption is justified by capacity constraints on beds and the fixed number of personnel of the hospital.[5]

[3] To obtain the last term in the benefit function we have used $(n^{12} + n^{123}) = N^2/\gamma$.

[4] Notice that although both N^1 patients and N^2 patients receive brain imaging, N^1 and N^2 are separate choice variables for the hospital. The first acts on the pool of patients who do not need alteplase but for whom a brain imaging is useful to the patient (for example to improve diagnosis).

[5] We could use a more parsimonious notation of the cost function of the type $C^i(N^i)$ which accounts for both fixed and variable costs. Splitting in its three components makes it is easier to relate the model to the application below.

We assume that the total cost, denoted by C, is additively separable in the three services, and therefore the total cost function of the hospital is:

$$C = c^1\left(N^1 + \frac{1}{\gamma}N^2\right) + K^1\left(N^1 + \frac{1}{\gamma}N^2\right) + F^1 + c^2N^2 + K^2(N^2) + F^2 + c^3N^3 + K^3(N^3) + F^3. \quad (3)$$

The financial surplus, denoted with π, is equal to $\pi = R - C$.

3.1.1 Hospital payoff function

We assume that hospitals are *motivated* to provide quality and that they care about patients' benefit. Such motivation is captured by the parameter α and gives the relative weight given to patients' benefit versus profits. Analytically, we assume that the hospital's utility is separable and additive in profits and in the motivation component:

$$U = \alpha B + \pi . \quad (4)$$

This specification is standard and in line with the seminal papers by Ellis and McGuire (1986) and Chalkley and Malcomson (1998a) and current health and public economics literature[6]. Equation (4) can be interpreted as the reduced form of the payoff function of a complex hospital organisation involving different key agents working within the hospital, such as the physicians and the hospital managers. Different agents will have a different payoff functions *within* the hospital, but they will ultimately have to agree on the level of quality to be provided as a result of a negotiation process.

Suppose for example that physicians are salaried and care only about patients' benefits (due to altruistic concerns) and managers maximise profits. We can think of the payoff function, $\alpha B + \pi$, as the sum of the payoffs of the different agents within the hospital where α can be interpreted as a parameter which is proportional to the relative bargaining power of doctors versus the managers. In this specific example, a value of α less than one implies that managers have a stronger influence in the decision process compared to doctors. In practice, the decisions may be less clear-cut with managers also caring about patients' benefit, and physicians also giving importance to costs, despite being salaried (if for example, they are concerned that the hospital may be closed or subject to strict monitoring if a deficit arises). This will reduce differences

[6] Amongst health economics papers see Eggleston (2005), Heyes (2005), Jack (2005), Choné and Ma (2011), Kaarbøe and Siciliani (2011). Amongst public economics papers see Besley and Ghatak (2006, 2005), Dixit (2005), Murdock (2002), Lakdawalla and Philipson (2006), Delfgaauw and Dur (2008, 2007), Glazer (2004), Makris (2009), Brekke et al. (2011) and Siciliani et al. (2013)

between the quality desired by the doctor and the managers but would still be consistent with the suggested payoff function.[7]

The scenario described matches with the institutional setting in the English National Health Service. In England hospital physicians are salaried employees of the hospital, and so there is no link between the P4P bonuses received by the hospital and the personal income of physicians. The hospital may choose to redistribute received bonuses to the departments generating the income, but there is no requirement for them to do so. Still, it is reasonable to expect hospital managers to monitor that policies are implemented to ensure alignment between the external reimbursement incentives and the effort of individual physicians. This is already a part of the normal duties of hospital management that face substantial personal incentives to keep budgets (or hospital managers may lose their jobs) and there is no reason to expect that financial incentives in the form of P4P are any different in that respect.

Note that high degree of altruism on the side of the physicians is not incompatible with low levels of α as along as hospital' managers can exert pressure on physicians to contain costs.[8] As an illustration, a value of $\alpha = 0.75$ could be the result of a setting with highly altruistic physicians and moderately motivated managers to contain costs, or equally a situation where physicians only maximise patients' benefit but managers have a strong influence on containing costs.

Although the parameter α has often been interpreted as related to physicians' altruism, this parameter could encompass other sources of motivation, such as reputational concerns, peer pressure, fear of malpractice suits and other pre-existing internal auditing mechanisms within the hospital. Higher levels of α imply that providers have already some existing incentives to provide quality on top of the pay-for-performance scheme. In summary, we interpret α as the relative weight given to patients' benefits versus hospital's profits.

[7] Although modelling the bargaining process is outside of the scope of this analysis and would be a contribution to the literature on its own, we could think as the current approach as the result of a cooperative Nash-bargaining outcome (Nash, 1953). In practice, there may be several ways to model such negotiations (eg cooperative versus non-cooperative bargaining) and much more structure on the details of the negotiations would have to be put in place (especially, in a non-cooperative set-up; (Osborne and Rubinstein, 1990)).

[8] Using experimental data from medical students acting as physicians, Godager and Wiesen (2013) found significant evidence of altruism. More precisely, 26% had a level of altruism less than one, 29% had altruism not statistically different from one, and the remaining 44% above one. There is unfortunately very little evidence on estimates on physician's level of altruism. Although the study by Godager and Wiesen is very interesting, the sample is small and it is not clear how the results from students could be extrapolated to physicians in a working environment (e.g. subject to pressures exerted by the managers) and in different countries.

One reason for a hospital to give higher weight to patients' benefit relates to their ownership status, i.e. whether the hospital has for-profit or non-profit status. Brekke, Siciliani and Straume (2012) model profit constraints adopting a payoff function of the type $U = \tilde{\alpha}B + (1-\delta)\pi$, where $\tilde{\alpha}$ is the weight given to patients' benefits. The parameter δ is zero for for-profit hospitals and is positive for non-profit hospitals. This formulation is analytically equivalent to ours by simply re-defining $\alpha = \tilde{\alpha}/(1-\delta)$. The presence of profit constraints is therefore equivalent to a higher weight given to patients' benefits.

As it will be shown below what matters for the design of the incentive scheme set by the regulator is the relative weight given to patients' benefits versus profits. The exact mechanism which leads to a given α is less critical. Therefore, for our purposes it is not necessary to model the various agency problems that may be arise within a hospital in further detail (physicians' incentives versus managers' incentives).

We do not restrict the parameter α to have an upper bound, though we show in Section 3.3 that a value of α above one implies that the regulator would set optimal negative prices, therefore eliminating the scope for pay-for-performance schemes. In other words, there is a value of α sufficiently high that quality is above the one desired by the regulator.

3.1.2 Optimal number of services chosen by the hospital

The hospital chooses the amount of services for each type of care N^i to maximise utility U. The hospital does not choose N^0 because the number of emergency patients admitted as stroke is exogenous but can choose the number of incentivised services (rapid brain imaging, alteplase and admission to an acute stroke unit).

Substituting (1), (2) and (3) into (4), the hospital's problem is to maximise:

$$U = \alpha\left(b^0N^0 + b^1N^1 + b^3N^3 + \left(b^{12} + \frac{1-\gamma}{\gamma}b^1\right)N^2\right)$$

$$+ p^0N^0 + p^1(N^1 + \frac{N^2}{\gamma}) + p^2N^2 + p^3N^3 \qquad (5)$$

$$-\left(c^1\left(N^1 + \frac{1}{\gamma}N^2\right) + K^1\left(N^1 + \frac{1}{\gamma}N^2\right) + F^1 + c^2N^2 + K^2(N^2) + F^2 + c^3N^3 + K^3(N^3) + F^3\right).$$

The optimality (first-order) conditions for N^1, N^2 and N^3 are:

$$\alpha b^1 + p^1 = c^1 + K^{1\prime}\left(N^{1*} + \frac{1}{\gamma}N^{2*}\right), \qquad (6)$$

$$\alpha\left(b^{12} + \frac{1-\gamma}{\gamma}b^1\right) + p^2 + \frac{1}{\gamma}p^1 = c^2 + K^{2\prime}(N^{2*}) + \frac{1}{\gamma}\left[c^1 + K^{1\prime}\left(N^{1*} + \frac{1}{\gamma}N^{2*}\right)\right] \qquad (7)$$

$$\alpha b^3 + p^3 = c^3 + K^{3\prime}(N^{3*}). \quad (8)$$

The marginal benefit from the non-monetary component and price is equated to the marginal cost. The optimality condition on the number of patients receiving alteplase also includes the marginal (monetary and non-monetary) benefits and costs of those patients who were considered for alteplase but ultimately do not receive it. The second order conditions are satisfied under weak regularity conditions on the convexity of the cost function.

Equation (7) can be re-written as:

$$\alpha(b^{12} - b^1) + p^2 + \frac{1}{\gamma}\left(\alpha b^1 + p^1 - c^1 - K^{1\prime}\left(N^{1*} + \frac{1}{\gamma}N^{2*}\right)\right) = c^2 + K^{2\prime}(N^{2*}), \quad (9)$$

where the third term is equal to zero using equation (6). Intuitively, although an increase in the number of patients receiving alteplase generates benefits and costs from imaging, the marginal (monetary and non-monetary) benefits from imagining are equal to the marginal costs and can therefore be ignored in the optimality condition for alteplase, which therefore simplifies to:

$$\alpha(b^{12} - b^1) + p^2 = c^2 + K^{2\prime}(N^{2*}).$$

The condition on alteplase is intuitive and analogous to (6) and (8). The main difference is that the marginal non-monetary benefit includes the benefit from alteplase *net* of the benefit of imaging, since patients considered for alteplase would obtain the benefits from imagining even if imaging revealed that patients had an haemorrhagic stroke (and therefore would not be eligible for alteplase).

3.2 The regulator

We assume that the regulator is utilitarian. It maximises the sum of patients' benefit net of transfers to the hospital and the utility of the hospital, $B-(1+\lambda)R+U$, where $\lambda>0$ accounts for the opportunity cost of public funds.

Substituting for U, we obtain $B(1+\alpha) - C - \lambda R$. It has been argued (e.g. Chalkley and Malcomson (1998a); Hammond (1987)) that this specification leads to double-counting of the benefits, which is due to altruistic motives. Following this suggestion, we eliminate double-counting and assume that welfare is given by $W = B - C - \lambda R$. This expression is intuitive. It gives the difference between patients' benefits and hospital costs minus the cost associated with raising public funds.

We assume that the purchaser designs the optimal contract subject to (i) a participation constraint, $U \geq 0$; and (ii) a profit constraint, $\pi \geq 0$. Since the provider is motivated, the first constraint is always satisfied when the profit constraint is satisfied. Since leaving a profit to the hospital is costly for welfare (due to the assumption of positive opportunity costs of public funds), it is optimal to set profits to zero, $\pi=0$ and $R=C$. The welfare function reduces to $W = B - (1+\lambda) C$.

The maximisation problem for the regulator is to maximise:

$$W = b^0 N^0 + b^1 N^1 + b^3 N^3 + \left(b^{12} - b^1 + \frac{1}{\gamma} b^1\right) N^2$$

$$-(1+\lambda)\begin{pmatrix} c^1\left(N^1 + \frac{1}{\gamma} N^2\right) + K^1\left(N^1 + \frac{1}{\gamma} N^2\right) + F^1 + c^2 N^2 + K^2(N^2) + F^2 \\ + c^3 N^3 + K^3(N^3) + F^3 \end{pmatrix}.$$

The optimality (first-order) condition for N^i is from the purchaser's perspective such that:

$$b^1 = (1+\lambda)\left(c^1 + K^{1\prime}\left(N^{1f} + \frac{1}{\gamma} N^{2f}\right)\right), \tag{10}$$

$$(b^{12} - b^1) + \frac{1}{\gamma}\left[b^1 - (1+\lambda)\left(c^1 + K^{1\prime}\left(N^{1f} + \frac{1}{\gamma} N^{2f}\right)\right)\right] = (1+\lambda)\left(c^2 + K^{2\prime\prime}\left(N^{2f}\right)\right), \tag{11}$$

$$b^3 = (1+\lambda)\left(c^3 + K^{3\prime}\left(N^{3f}\right)\right). \tag{12}$$

where f denotes first best, and the second term in the square bracket in equation (11) can be set to zero using equation (10). This suggests that the optimal number of services is such that patients' benefits equate to marginal costs.

3.3 Implementation

In the following we assume that the regulator sets prices only on the three incentivised services. We therefore take the number of incentivised dimensions of quality as exogenous. In the Appendix, we show that the optimal prices are:

$$p^{1f} = (1-\alpha) b^1 - \lambda\left(c^1 + K^{1\prime}\left(N^{1f} + \frac{1}{\gamma} N^{2f}\right)\right), \tag{13}$$

$$p^{2f} = (1-\alpha)(b^{12} - b^1) - \lambda\left(c^2 + K^{2\prime}\left(N^{2f}\right)\right), \tag{14}$$

$$p^{3f} = (1-\alpha) b^3 - \lambda\left(c^3 + K^{3\prime}\left(N^{3f}\right)\right). \tag{15}$$

The optimal price is equal to patients' benefits discounted by the weight given to patients' benefit. When such weight is higher, the regulator needs to incentivise the hospital less. A higher opportunity cost of public funds implies a lower optimal price. While prices 1 and 3 are proportional to the marginal benefit from the service, price 2 is proportional to the benefit from alteplase net of the benefit from imaging since, as already outlined in the previous section, patients considered for alteplase would obtain the benefits from imagining even if imaging revealed that patients had an haemorrhagic stroke (and therefore would not be eligible for alteplase).

Note how values of α larger than one always imply negative optimal prices in the presence of a positive opportunity costs of public funds. The intuition is straightforward. Providers are motivated and take patients' benefits fully into account. They also take costs fully into account through the profit motive. The fixed tariff is a lump-sum payment and therefore does not alter incentives to provide quality. The hospital equates patients' benefits with its costs. With zero opportunity cost of public funds, the optimal price is zero since the optimal quality provision coincides with the first best. With a positive opportunity cost of public funds, costs are inflated from the purchaser's perspective and multiplied by $(1+\lambda)$ and the provider chooses a level of quality in excess of the one desired by the regulator.

Moreover, from the profit constraint which has to be satisfied with strict equality, we can compute the basic tariff

$$p^{0,f} = \left\{ \sum_{i=0}^{3} [c^i N^{i,f} + K^i\left(N^{i,f}\right) + F^i] - \sum_{i=1}^{3} p^i N^{i,f} \right\} / N^0. \qquad (16)$$

This condition simply states that the basic tariff is equal to the average cost net of other transfers to the hospital.

4 Comparing actual and optimal price-setting in BPT for emergency stroke

In this section we describe the implementation of our proposed (optimal) price-setting scheme, and the actual one adopted by the Department of Health. We then compare the actual tariff with the optimal tariff.

4.1 Optimal price-setting for Best Practice Tariffs for emergency stroke

4.1.1 Information requirements

Equations (13-15) show that setting the optimal hospital tariff requires knowledge of the marginal benefit and cost of the incentivised dimension of care and the opportunity cost of public funds. We illustrate our results for levels of the hospital's weight given to patients'

benefits relative to profits, α, which vary between 0 and 1 (as mentioned in section 3.1.1 a value of α above one would imply negative optimal hospital prices). We present our estimates at plausible ranges of benefits and costs to illustrate the robustness of optimal prices to uncertainty about the benefit and cost estimates. We fix the opportunity cost of public funds at the plausible value equal to $\lambda = 0.2$ since this is consistent with previous literature and policy practice.[9]

In the following we focus on obtaining plausible estimates for the marginal benefits of the three processes incentivised in the three BPTs for emergency stroke. For assessing the marginal benefit of the interventions incentivised by the BPT we use Quality Adjusted Life Years (QALYs). As the BPT performance indicators for stroke are all based on clinical evidence, we searched the medical literature behind the national clinical guidelines and NICE guidance to find estimates of the per-patient QALY gains associated with the incentivised interventions.

We focus on studies that present estimates of per-patient QALY gains as close as possible to the counterfactual treatment, i.e. the type of care an NHS patient would have received without the incentivized process of care. Where possible, we sought studies with a lifetime perspective on the benefits associated with the incentivised dimensions of care. We adopt the monetary social value of a QALY most often used by the Department of Health in its policy Impact Assessments of £50,000[10] (Shah et al., 2012).

4.1.2 Marginal benefit of treatment delivered in an acute stroke unit

Saka et al. (2009b) used data from the South London Stroke Register and Markov modelling to assess the 10-year cost-effectiveness of emergency stroke care in a general medical ward

[9] The seminal work on estimating the opportunity costs of public funds was carried out by Browning (1976) who estimated a value of λ between 0.09-0.16 for the United States which in subsequent work (Browning, 1987) was re-estimated to between 0.18-0.47. Although numerous definitions of λ exist (Massiani and Picco, 2013) there is consensus that estimates of λ depend on the specific tax regime and assumptions about how economic agents respond to new taxes. While there are no published estimates for the UK, Ruggeri (1999) estimated a value of λ for a small open economy of 0.13-0.18 using Canada as a an example. Some countries have guidance values of λ for use in national cost-benefit appraisals. While the UK Treasury Green Book (HM Treasury 2003) does not suggest a specific value of λ, a survey of appraisal methods for infrastructure projects in the Nordic countries (Lyk-Jensen, 2007) showed that the official value of λ for project appraisal in Denmark and Norway is 0.2 and $\lambda = 0.3$ in Sweden. Considering the similarity of the UK to Canada and the Scandinavian countries we assume a value of λ of 0.2 in this paper.

[10] The Impact Assessment of End of Life Care (Department of Health, 2008a) explained that the £50,000 figure was based on a projection of a 2004 estimate of willingness to pay for an additional life year of £29,000 by the Department for Environment, Food and Rural Affairs. The increase to £50,000 was justified by a wish to reflect price changes from 2004 to 2006, the older age and poorer life quality of the likely target group of DH interventions, and an upward rounding due to concern that the figure might be an under-estimate of the social value of a QALY.

compared to stroke unit care[11]. The study found an incremental QALY gain per patient of 0.472 QALYs associated with care in a stroke unit. As the time horizon of the study was restricted to 10 years and the average age of patients was 64 years, the QALY gain of treating patients in a stroke unit rather than in a general ward is potentially larger. When assessing the robustness of our results to the uncertainty around this estimate we follow the approach taken by Saka et al. (2009b) in their univariate sensitivity analysis and use QALY values of above and below 20% the central value.

4.1.3 Marginal benefit of rapid brain Imaging

Wardlaw et al. (2004b) used a decision tree and a deterministic model to compare the cost effectiveness of 12 different CT-scanning strategies.[12] Usual care at the time of the study was to scan all patients within 48 hours. The strategy included in the study that was most similar to the strategy incentivized by the BPT was to "scan patients on anticoagulants, in life-threatening condition, or candidates for thrombolysis immediately, and scan all remaining patients within 24 hours". This strategy was associated with a gain of just 0.1 QALYs per 1,000 patients over a five-year period compared to usual care. Wardlaw et al. (2004) justified the seemingly low QALY gain by the high proportion of haemorrhagic stroke patients (85%) in the study population. For these patients the main treatment strategy is aspirin, which only needs to be given within 48 hours. The study assumed that just 4% of patients would reach the hospital in time to be considered for thrombolysis, and suggested that the cost effectiveness of scanning all patients immediately would be higher if the proportion of potentially-eligible patients was higher. Wardlaw et al. (2004a) further explored the sensitivity of their estimates to the assumptions of their model. While these sensitivity analyses suggested some uncertainty around the absolute level of QALYs gained from different scanning strategies, few of these suggested any uncertainty about the incremental QALY gain of 0.1 per 1,000 patients. Allowing the proportion of actual strokes in a population attending hospital with stroke-like symptoms to be 77% rather than the base level of 81%, suggests a potential incremental QALY gain of 0.2 per

[11] The Stroke unit studied in the paper admitted only stroke patients and was a mixed unit with 4 acute beds and 23 rehabilitation beds. The paper defined stroke units as units fulfilling at least 4 of the criteria set out for stroke units by the Royal College of Physicians (see the left hand side of Table A.2), but was not explicit about the configuration of the stroke unit from which the data was collected. It is unlikely that the unit fulfilled the stricter criteria of an acute stroke unit (see the right hand side of Table A.2) and the similar definition given by the DH (see Table A.1). Although this might suggest that the estimated QALY gain is in the lower end of what is being incentivised, it should be noted that the Cochrane review carried out by the Stroke Units Trialist's Collaboration did not find a statistically significant difference in the odds of death, or death or requiring institutional care or death or dependency when comparing acute monitoring with acute non-intensive units.

[12] The analysis was carried out for a cohort of 1000 patients (age 70-74) and repeated for 1000 60-64 year and 80-84 year old patients in teaching urban and rural general hospitals using data from a range of sources.

1,000 patients, and we use this estimate as an upper bound when exploring the robustness of our results to the benefit estimate.

4.1.4 Marginal benefit of thrombolysis with alteplase if clinically indicated

A deterministic cost-effectiveness analysis based on a Markov modelling simulation study of patients with emergency ischaemic stroke receiving alteplase within 4.5 hours of onset of symptoms (NICE 2012) found an incremental per-patient gain of 0.333 QALYs from alteplase treatment.[13] The study considered lifetime effects of treatment, assuming no change in health status after 12 months (other than death). The probabilistic cost-effectiveness analysis from the same study suggested upper and lower bounds of 0.235 and 0.411 QALYs per patient.

4.1.5 Marginal costs of incentivised processes

Ideally, we would use estimates of marginal costs, which according to our model depend on the provider's level of performance. As estimates of marginal costs of the incentivised processes are not available from the literature, we instead compute the difference between the average cost of standard care and average cost of incentivised care from the literature instead. In the presence of fixed costs, the marginal cost is likely to be lower than the average cost. For each care process we discuss to which extent this approximation is likely to be valid and further explore the robustness of our results to the choice of cost estimate in the sensitivity analysis. We also note that with an opportunity cost of public funds of 0.2, uncertainty in costs (including from estimates of the marginal costs) have a relatively small impact on price. This is described in more detail below.

For admission to a stroke unit we use the per-patient per-diem costs reported by Saka et al. (2009). Multiplying the difference in costs between the stroke unit (£164.80 per day) and the general ward (£114.80 per day) by the average length of stay (34.4 days) yields a per-patient cost of treatment in a stroke unit of £1,720. Note that this price may be an underestimate, since the BPT for emergency stroke requires patients to be admitted to an *acute* stroke unit, which requires more potentially costly characteristics to be fulfilled (summarised in Table A.1 and A.2). In assessing the robustness of our results to this estimate, we follow again the approach by Saka et al (2009b) and assess how an increase or decrease in costs of 20% affects our results. The Department of Health (2007b) suggest that the only additional fixed costs in a stroke unit

[13] The estimate was submitted by the manufacturer of alteplase (Boehringer Ingelheim), but reviewed by Davis et al (2012). The review noted that manufacturer failed to model the correlation between the risks of death and death or dependency but that this was unlikely to affect the incremental cost effectiveness ratio (ICER) of administering alteplase. In addition, Davis et al. (2012) commented that the utility values for patients in dependent and independent health states stays fixed over the lifetime not allowing for deterioration of health related quality of life over time. Davis et al. (2012) suggest that this may potentially favour alteplase over standard care, but also note that the model is not very sensitive to the utility values applied and so consider the effect likely to be small.

compared to a general ward is low-cost monitoring equipment. The largest component by far is the variable staffing input of auxiliary nurses, physiotherapist and assistant, occupational therapist and assistant, speech and language therapist, psychologist to ensure delivery of the stroke standards (24 hour access consultant stroke specialist input, continuous monitoring by specialist nursing staff, rapid access to specialist advice on neurointensive care and respiratory, swallowing, dietary and communication issues; see Tables A1 and A2). The fixed cost component is thus likely to be very small, and we conjecture that differences in the average cost are a good approximation of those in the marginal cost.

We assume that the direct costs to providers of achieving the stroke BPT to scan patients within 24 hours are the additional costs relative to scanning all patients within 48 hours. Wardlaw et al (2004) reported that the mean costs of a CT scan at 2000 prices were £43 in normal working hours and £79 after hours. Their model suggested that the cost of CT scanning patients was £47 if required within 48 hours for all patients and £71 if all patients were required to be scanned immediately. We take the mid-point of these estimates as the per-patient cost of achieving a scan within 24 hours for all patients (£59), and therefore the additional costs to the provider per patient of achieving 24 hours compared to 48 hours is £12 (£59 minus £47). We follow the same approach for calculating upper and lower bounds of the marginal cost of more rapid CT scan based on cost estimates from Wardlaw et al. (2004) yielding a lower cost estimate of £8 and an upper bound of £34. Cost analyses in Wardlaw et al. (2004) suggest that variable costs make up 49%-88% of the average costs per CT scan depending on the time of day. This suggests that the marginal cost is lower than the average cost, which we reflect in the sensitivity analysis investigating the effect of assuming lower costs. Given that labour costs make up approximately 39%-85% of the average costs per scan (higher share for out of office hours) and the fact that the required increase in the rapidity of CT scans is likely to take place out of hours, we conjecture that £8 is a plausible lower bound of the marginal cost.

For thrombolysis with alteplase we rely on the cost estimate from Saka et al. (2009) which reflects the additional cost of administering alteplase including the cost of the drug. Sandercock et al. (2004) suggested that the plausible range of cost for alteplase treatment is between £480 and £1000. As the main cost components for treatment with alteplase is the cost of the drug and the nursing time taken to administer the drug, the average cost of alteplase treatment is thus unlikely to vary significantly from the marginal cost.

We adjust all cost estimates to 2013/14 prices using the hospital and community health services price index (Curtis, 2014).

4.2 Optimal price-setting in Best Practice tariffs for emergency stroke

Figures 1-3 present the implementation of the optimal best practice tariffs for emergency stroke as described by Equations (13-15) for plausible levels of benefits of the incentivised dimensions of care and for a given central estimate of costs (see Tables A3-A5 in the Appendix for exact monetary values).

The optimal price is highest when hospitals are purely profit-maximising ($\alpha = 0$). In this case, the price should be set equal to the marginal benefit associated with the incentivised process (net of the marginal cost weighted by the opportunity cost of public funds) and, as a result, the benefit estimate has a substantial impact on the optimal price. For example, at the central estimate of a per-person QALY gain from admission to a stroke unit of 0.472, assuming a social value of a QALY of £50,000, the optimal price of admission to a stroke unit is £23,226.

The figures show that the optimal prices decrease with the weight given to patients relative to profits α. The result is intuitive. If hospitals give a higher weight to patients' benefit then they need less of a financial incentive to incentivise performance. For example, at $\alpha = 0.8$ the optimal price of admission to a stroke unit reduces dramatically to £4,346.

The analysis therefore highlights the critical role of the relative weight given to patients' benefits as opposed to profits, α, and of patients' benefits, b, in determining the optimal price. Assuming a social value of a QALY of £50,000, and considering $\alpha = [0.1,0.9]$ and a given $\lambda = 0.2$, the optimal price at the central cost estimate for treatment in an acute stroke unit lies in the interval [£1,514, £ 27,946]; the optimal price for rapid brain imaging lies in the interval [£-3, £7]; and the optimal price for alteplase lies in the interval [£1,011, £20,382]. For treatment in an acute stroke unit and alteplase, α matters relatively more, because the marginal benefit is substantially higher than the cost of carrying out the intervention.

4.2.1 Uncertainty in benefits

Uncertainty on the size of the benefits, as described in section 4.1.2-4.1.3, can have a substantial impact on prices especially at low levels of α. For example, at $\alpha = 0$, the maximum optimal price for care in an emergency stroke unit is £27,946, assuming a social value of a QALY of £50,000 and a per-patient gain of 0.566 QALYs and £18,506 if the marginal benefit of treatment in a stroke unit is assumed to be 0.378 QALYs per patient. The difference in the estimated QALY gain is thus carried directly over to the difference in the optimal price. However, when hospitals give a higher weight to patients' benefit relative to profits, e.g. at $\alpha = 0.8$, the difference in QALY gain is only partially reflected in the optimal price which would vary between £3,402 and £5,290. Similarly, for treatment with alteplase, if no weight is given to patients' benefit ($\alpha = 0$) the optimal price varies between £11,582 for a QALY gain per person of 0.235 and £20,382 at the

upper bound QALY gain estimate of 0.411. Again, as the weight attached to patients' benefit increases (assume again $\alpha = 0.8$) the variation in optimal prices is less sensitive to the QALY estimate, and would vary between £2,186 and £3,946.

Figure 1: Optimal performance payment (£) for treatment in an Acute Stroke Unit at different levels of treatment benefits

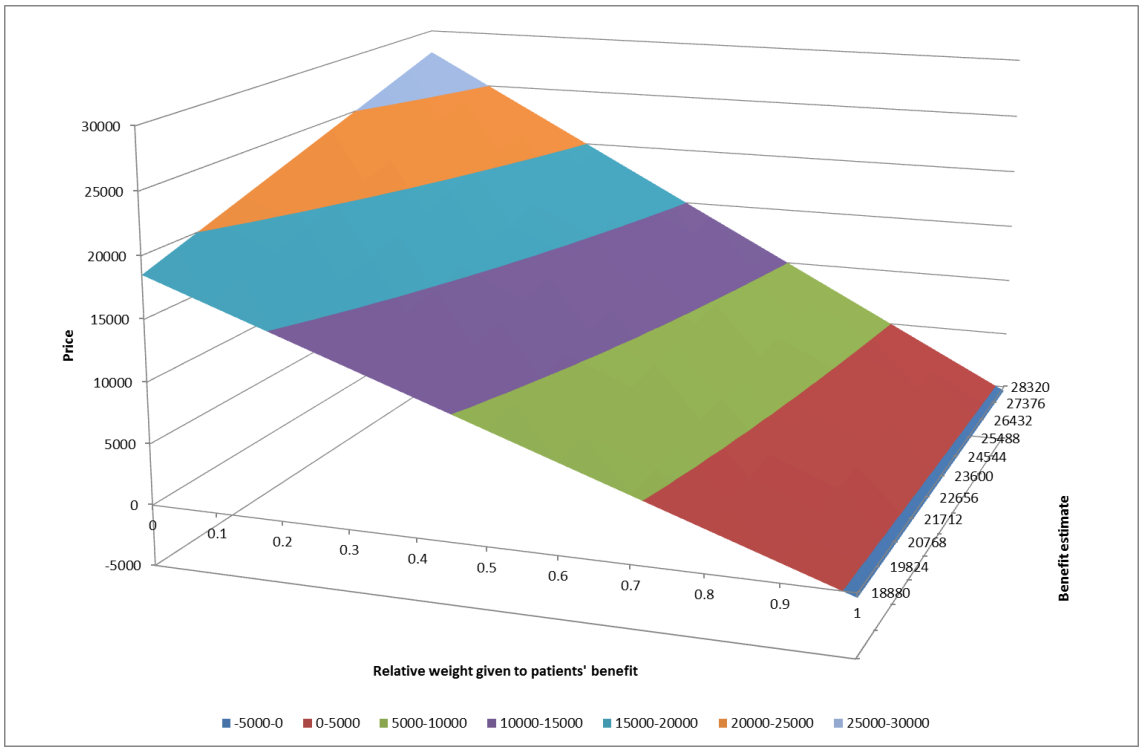


Figure 2: Optimal performance payment (£) for rapid brain imaging at different levels of treatment benefits

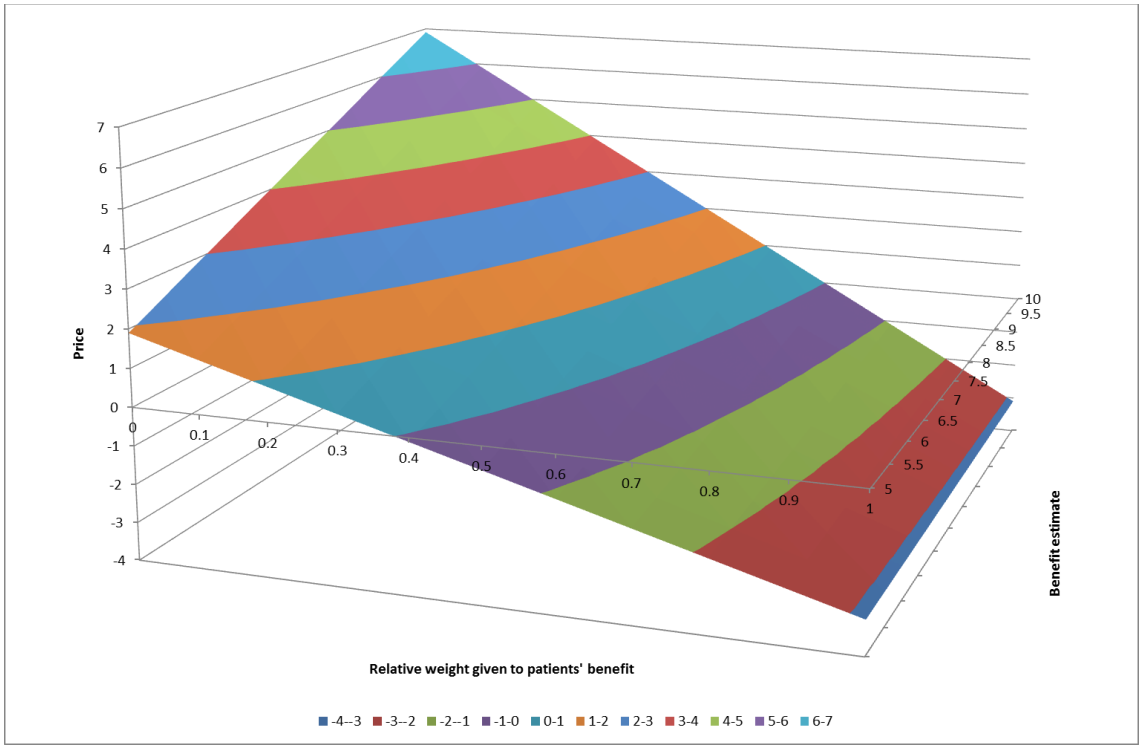
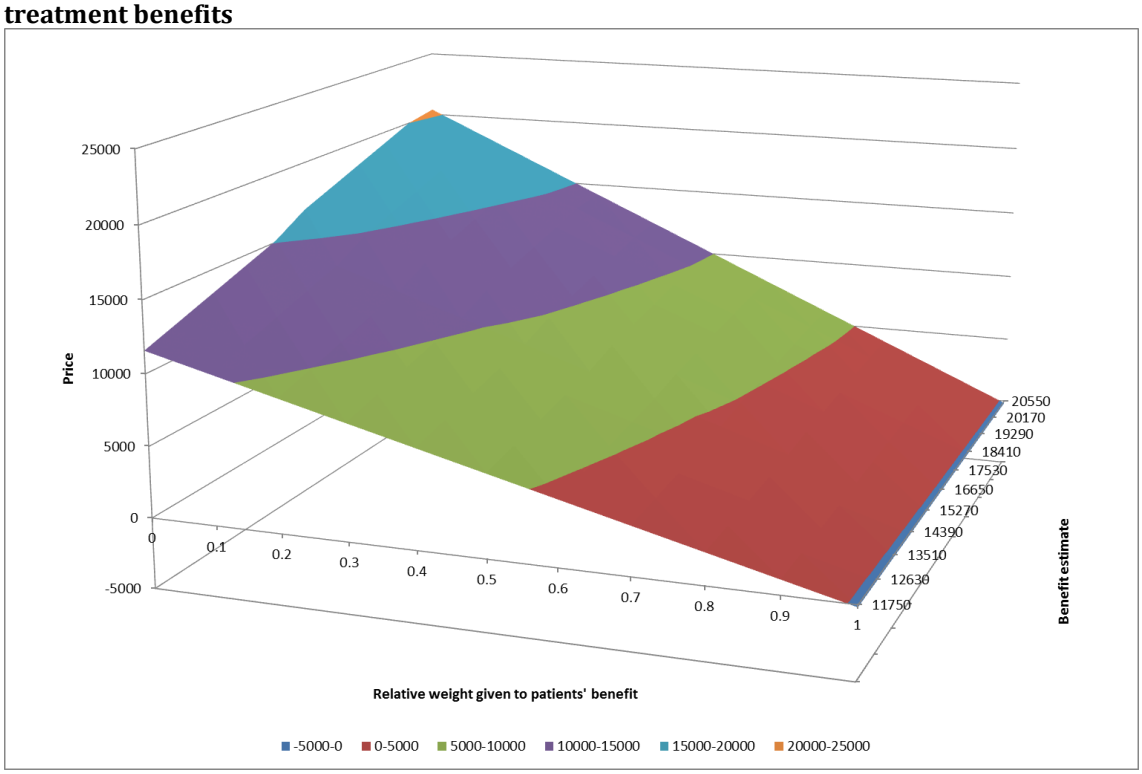


Figure 3: Optimal performance payment (£) for thrombolysis with alteplase at different levels of treatment benefits



4.2.2 Uncertainty in costs

Figures 4-6 and Tables A6-A8 illustrate how the optimal prices vary with the cost estimate at a given level of benefits. The optimal prices are relatively insensitive to changes in estimates of the costs for a given level of α. This is in contrast to Figures 1-3, which shows great price sensitivity in relation to changes in estimates in benefits. The result is due to relatively lower levels of costs (compared to benefits), smaller variations in uncertainty around costs, and the fact that costs are weighted by the opportunity cost of public funds. It is still the case that the level of α has a considerable impact on the optimal price.

Figure 4: Optimal performance payment (£) for treatment in an Acute Stroke Unit at different levels treatment costs

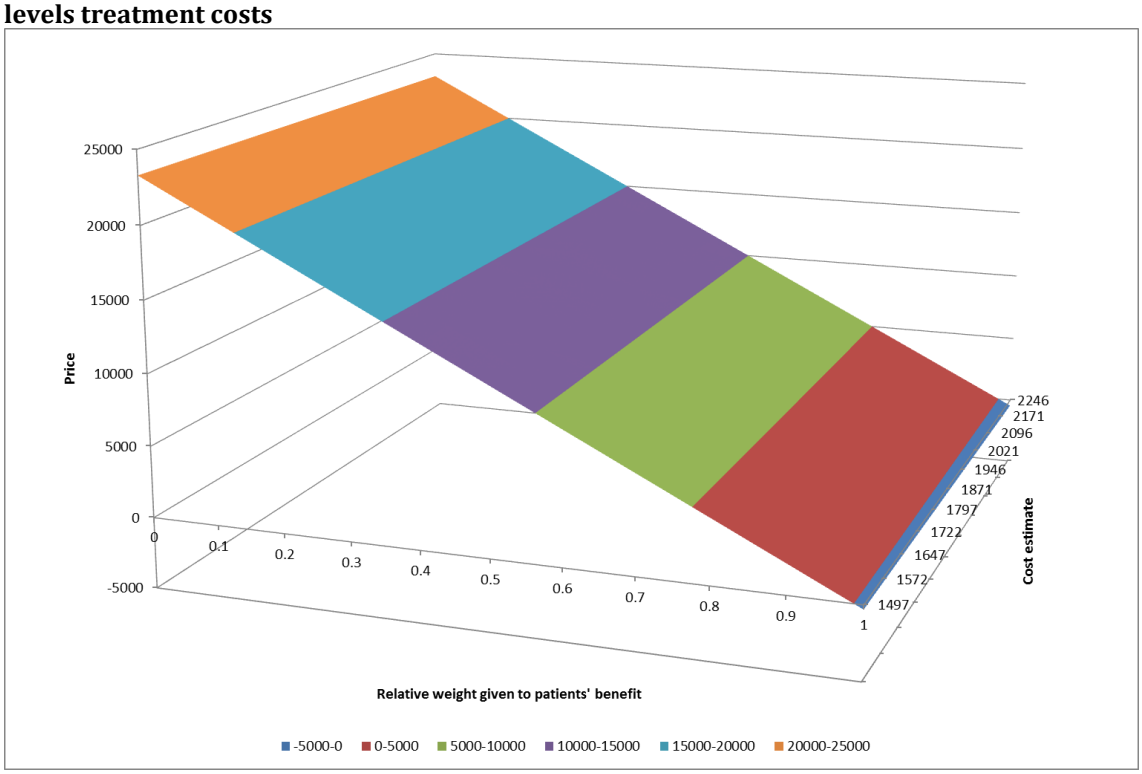


Figure 5: Optimal performance payment (£) for rapid brain imaging at different levels of treatment costs

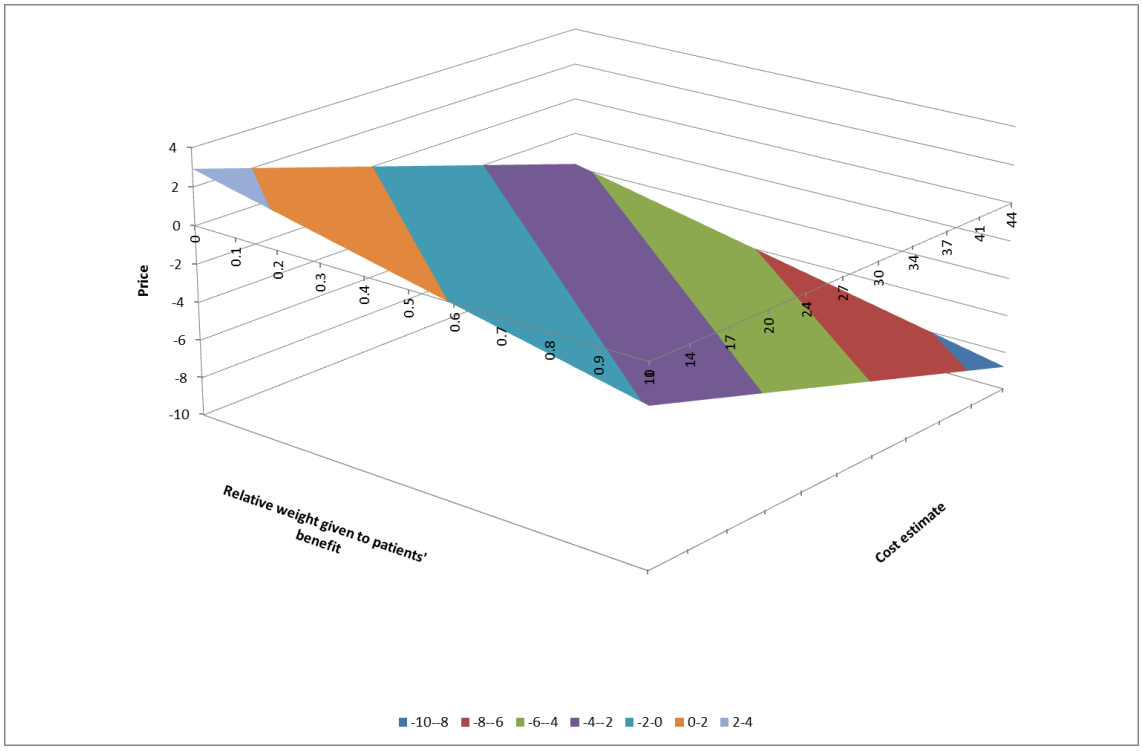
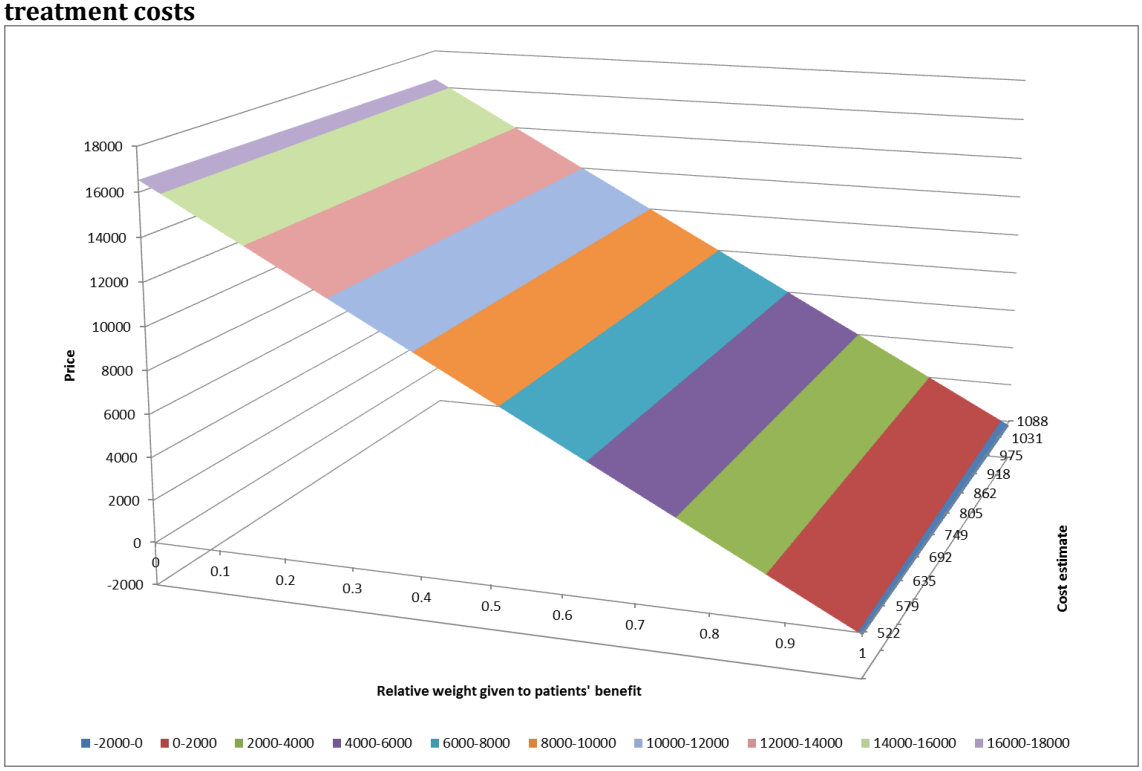


Figure 6: Optimal performance payment (£) for thrombolysis with alteplase at different levels of treatment costs



4.3 Actual price-setting in Best Practice Tariffs for emergency stroke

When the English Department of Health introduced the BPT in the financial year 2010/11, it explained that it wished to set prices "not just at the national average but instead to better reflect the costs of delivering best practice," with a built-in financial incentive "to encourage uptake of best practice in the early stages." The financial incentive was expected to be removed in the future and "align[ed] … with the actual cost of best practice" (Department of Health, 2010a, p. 51).

The BPT tariff was *top-sliced* from the basic HRG tariff implying that a hospital providing none of the incentivised services to any patient would make a financial loss. This can be seen in Table 2. When BPT was introduced in 2010/11 the basic tariff reduced. This was further reinforced in the following years where the BPT for rapid brain imaging and delivery in the stroke unit was further raised in conjunction with a further reduction in the basic tariff. Using the figures provided in Table 1 for 2011/12 we can infer however that, compared to 2009/10, hospitals received on average larger payments, after the inclusion of the BPT. Therefore, taking into account the hospital behaviour in responding to the scheme, hospitals ended up on average with a financial surplus following the introduction of the BPT.

The tariffs and their development over time are described in more detail in Table 2. Between 2009/10 and 2012/13 there were two base tariffs— *AA22Z Non-Transient Stroke or Cerebrovascular Accident*, *nervous system infections or encephalopathy* and *AA23ZZ Haemorrhagic Cerebrovascular Disorders.* From 2013/14 new groups were introduced for patients with and without complications and co-morbidities. Co-morbidities are defined as "additional conditions that the patient might come into hospital with that increase the complexity of the primary intervention" and complications as "events during treatment that […] increase complexity" (Department of Health, 2012a, p. 22).

While the initial BPT prices were calculated close to the costs of providing the service, the subsequent adjustments were justified by a desire to increase the incentive for delivering best practice (Department of Health, 2012b). Both before and after the reform, alteplase has been paid for separately in addition to the base-DRG tariff for stroke patients. When the P4P was introduced, the DRG payment was lowered. In the first year, the base tariff for patients with and without co-morbidities and complications (CC) was lowered by £253, while the available bonus payments (not considering alteplase which was paid for separately) was £475. In the second year, the baseline tariff was reduced by a further £383 for patients without CC and by £579 for patient with CC. Simultaneously, the incentive payment was doubled to a total of £950. This development continued in subsequent years with further reductions to the base tariffs and increases in the bonus payment. In the final year, the bonus payment for alteplase was £828, for rapid brain imaging £399 and for treatment delivered in an acute stroke unit £1,026.

Table 2: Actual prices (£) for emergency stroke care and incentivised dimensions of quality 2009/10-2013/14

Component	2009/10	2010/11	2011/12	2012/13	2013/14
Base tariff					
AA22Z	4,348	4,095	3,712	3,005	N/A
AA22A	N/A	N/A	N/A	N/A	2,764
AA22B	N/A	N/A	N/A	N/A	2,851
AA23Z	4,411	4,158	3,579	2,987	N/A
AA23A	N/A	N/A	N/A	N/A	1,764
AA23B	N/A	N/A	N/A	N/A	1,377
Additional BPTs					
Rapid brain imaging	0	133	266	399	399
Treatment delivered in an acute stroke unit	0	342	684	1,026	1,026
Alteplase	828	828*	828*	828	828

Notes: AA22Z: Non-Transient Stroke or Cerebrovascular Accident, Nervous System Infections or Encephalopathy. AA23Z: Haemorrhagic Cerebrovascular Disorders. *A: with Co-morbidities and Complications (CC), *B: without CC.
Source: Payment by Results guidance for 2009/10-2013/14.

4.4 Comparison of actual and optimal price-setting

The Department of Health description of the scheme suggests that BPT tariffs are mainly set as a function of costs in combination with an added incentive that appears to be arbitrarily set.[14] Our analysis highlights the importance of also considering the patients' benefits of the incentivised processes and hospitals' weight given to patients relative to profits.

The current performance payment for treatment at an acute stroke unit is £1,026. At $\lambda = 0.2$ this is consistent with a weight given to patients' benefit relative to profits α between 0.92 and 0.95 for the plausible range of marginal benefit from treatment in a stroke unit. Remarkably, for alteplase, at a QALY value of £50,000, the tariff of £828 is also consistent with α between 0.92 and 0.95.

The performance payment for rapid brain imaging set by the Department of Health has tripled from £133 when BPT was first introduced to £399 in 2013/14. This price is not consistent with any positive levels of α in our framework which suggests low or even negative optimal prices. This is because the major benefit of rapid brain imaging only arises for patients with an ischaemic stroke who can subsequently be treated with alteplase, which has a high benefit for these patients. Due to the high prevalence of ischaemic strokes (about 80%) and the high expected benefit for ischaemic patients, the incentive payment for brain imaging need not be very high.

Aside from brain-imaging, there are two possible ways to interpret our results. First, the current tariff is optimal and hospitals currently assign a relatively high weight to patients' benefit (ie above 0.9); therefore the need to incentivise performance through financial incentives is limited. Second, that the current price-setting in the BPTs for emergency stroke appear relatively low compared to the optimal price, and higher prices could be welfare-improving. Which interpretation is correct depends on the correct assessment of the relative weight given

[14] Specifically for stroke, the PbR guidance explains that the cost of the initial CT scan was originally paid for by the conventional tariff, but that the costs relating to the scan has now been removed "so that providers will only be reimbursed for scans that are in line with best practice." The Step-by-step guide to calculating the national tariff (Department of Health, 2010b) further specified that reduction in the conventional tariff relating to delivering a CT scan was £133 and that the tariff was reduced by an additional £120 to reflect the current compliance to the delivery of care on an acute stroke unit estimated using data from the National Sentinel Stroke Audit 2008 and Vital Signs 14. The tariff for treating patients in an acute stroke unit was derived on the basis of the cost estimate of implementing the National Stroke Strategy from the Stroke Strategy Impact Assessment (Department of Health, 2007b) using the fraction of strategy implementation costs relating to delivering care in an acute stroke unit. The impact assessment based its cost estimates on the assumption that an additional 37% acute beds were required to ensure that all stroke units would have sufficient bed capacity to be run at an 85% bed occupancy rate which was assumed to be enough to allow 95% of all stroke patients to be admitted to an Acute stroke unit.

to patients' benefits relative to profits. The empirical evidence, as far as the authors are aware, does not provide an estimate of such relative weight.[15] This would seem an area for future development to advance the design of pay-for-performance schemes.

5 Discussion and concluding remarks

This paper aims at bridging the gap between the theory and the practice of pay-for-performance incentive schemes. Price-setting has been treated informally in practice and the emphasis has been placed on the proportion of total revenue accounted for by incentive payments. We have presented a model of hospital optimal price-setting of process measures of performance for stroke patients. We have compared the derived optimal price with the actual price set in the English NHS as part of the Best Practice Tariffs scheme from 2010/11.

The main features of our model are that optimal prices should reflect the marginal benefit of the health gain associated with the incentivised dimensions of care, the level of hospital weight given to patients' benefit relative to profits, and the opportunity cost of public funds. In our implementation we have searched the medical literature for estimates of the health gains expected from delivering the incentivised dimensions of care. Using a monetary social value of a QALY of £50,000 (previously used by the Department of Health), we have described the optimal prices for treatment in an acute stroke unit, rapid brain imaging, and thrombolysis with alteplase in intervals depending on uncertainty of key parameters.

Overall, the model provides a framework in which scholars and policymakers can consider how incentive theory should inform price-setting for quality. There is currently no good overview of how prices are actually set in pay for performance schemes, but the focus appears to be on the costs of delivering the process measures of quality. Our analysis highlights the importance of setting prices based on expected benefits, not only costs, and the weight which hospitals assign to patients' benefits relative to profits. Given the monetary estimates of the benefits are large, we conjecture that prices for stroke care in the English NHS are currently set below the optimal level. Alternatively, hospitals are assigning a weight to patients' benefit which is above 0.9 and therefore close to the weight associated to a societal perspective which maximises the difference between benefits and costs (the latter weighted by the opportunity cost of public funds). This highlights the importance of obtaining accurate empirical estimates on hospitals'

[15] As already discussed above, Godager and Wiesen (2013) provide estimates of physicians level of altruism using experimental data from medical students acting as physicians. However, these do not necessarily reflect the relative weight give to patients' benefit within a more complex organisation such as hospitals. See Section 3.1 for a more detailed discussion.

weight given to benefits relative to profits in future research since this plays a critical role in the design of optimal incentive schemes.

We conclude by briefly discussing some limitations and other avenues for further research. We have not considered the potential multitasking problem with respect to unmonitored or unverifiable aspects of quality of care. Although this cannot be excluded, the BPT incentive scheme covered all the key quality dimensions, leaving little as unmonitored. Therefore, we do not think that allowing for the minor dimensions of unmonitored quality would alter our key results. If minor unmonitored aspects of quality were complements, then we would expect the optimal prices to be marginally higher. Moreover, setting prices that take account of multitasking would require considerably more information (in addition to the heavy information requirements already included in this study), including the responsiveness of unmonitored quality to changes in prices for monitored quality and the benefits and costs of changes in unmonitored quality.

The study does not deal explicitly with uncertainty and heterogeneity of patients. It may be argued that patients may vary in severity and the cost and benefit that arise from treatment. If benefits and costs were noisy and the provider had risk aversion or limited-liability constraints, the presence of uncertainty may imply prices being lower than under certainty. We do not think these issues are critical in our current application. Hospitals are large organisations and may be thought as risk neutral. Moreover, temporary deficits within some treatments can be compensated by profits in other treatments, and hospitals provide treatments across hundreds of distinct DRGs. In addition, strokes are common: over a financial year volumes will be high and highly predictable. The high volumes also imply that the average costs of heterogeneous treatments are likely to be stable with more costly patients being compensated by less costly patients. This is also consistent with DRG pricing which is based on average costs, and prices do not differ across hospitals. Therefore, although the tariff may not exactly correspond to the price of a given patient, it will on average cover treatment costs. Finally, the processes incentivised by the P4P scheme are mostly routine processes, and thus while patient severity may vary, it is unlikely that this would affect the costs of administering alteplase (which is a drug), or taking a CT/MRI scan of the patient. The supply of services may only vary to some extent with patient severity in a specialised stroke unit. In contrast, uncertainty may be more relevant for incentives schemes tailored at small organisations, for examples family doctors working on their own or in small practices (e.g. the QOF scheme for general practitioners in England). Family doctors working on their own may be risk averse. If the volume of patients for the incentivised process is small and erratic, uncertainty may play a more significant role. We leave this for future research.

Uncertainty may also play a more prominent role for incentive schemes that are based on health outcomes rather than process measures of quality. The current analysis focuses on process measures of quality and this has been the key focus in the development of pay-for-performance schemes. It may be argued however that providers' incentive schemes should be based on health outcomes, rather than process, since ultimately health systems aim at maximising health gains. In practice, schemes based on outcomes are still at an early developmental stage. Available measures of health outcomes on a routine basis are mainly based on mortality or readmission rates, which are very crude, imperfect and noisy. These do not capture the distribution of health gains but only the lower tail of the health distribution, when the outcome is as extreme as death. They are also irrelevant for most elective care where mortality rates are negligible or infrequent and readmissions are very low. Moreover, health outcome measures are more sensitive to the severity of the patients, where risk adjustment becomes paramount. An interesting recent development within the English NHS is the development of Patient Reported Outcome Measures (PROMs) which aim at measuring health gains before and after the surgery. However, these only apply to four elective conditions (hip and knee replacement, hernia and varicose veins). Although incentivising health outcomes is attractive, its' development is still subject to a range of measurement issues and for this reason policymakers are likely to privilege process measures of quality in the future.

If reliable and robust health outcome measures become available in the future, policymakers will have to decide whether to incentivise outcomes, process or both. Incentivising outcomes is likely to generate more uncertainty in the payoffs: outcomes are noisier and they also depend on patients' compliance, therefore making physicians more exposed to factors which they cannot control. They entail a greater transfer of financial risk from the payer to the provider. The cost of an improvement in outcome is also unknown to the payer and cannot be used when setting the optimal price. This is in contrast to process measures of quality that are under full control of the physicians. This seems to generate a trade-off between incentivising outcomes versus processes. The choice between the two is likely to depend on the details of the treatment and measurement constraints. We leave the detailed investigation of these trade-offs for future research though we conjecture that each may be optimal for some conditions or treatments, and are likely to coexist in the future.

References

Besley, T., Ghatak, M., 2006. Sorting with motivated agents: implications for school competition and teacher incentives. Journal of the European Economics Association (Papers and Proceedings) 4, 404–414.

Besley, T., Ghatak, M., 2005. Competition and incentives with motivated agents. American Economic Review 95, 616–636.

Brekke, K.R., Siciliani, L., Straume, O.R., 2012. Quality competition with profit constraints. Journal of Economic Behavior & Organization 84, 642–659.

Brekke, K.R., Siciliani, L., Straume, O.R., 2011. Hospital competition and quality with regulated prices. Scandinavian Journal of Economics 113, 444–469.

Browning, E.K., 1987. On the marginal welfare cost of taxation. The American Economic Review 11–23.

Browning, E.K., 1976. The Marginal Cost of Public Funds. Journal of Political Economy 84, 283–298. doi:10.2307/1831901

Cashin, C., Chi, Y.-L., Smith, P., Borowitz, M., Thomson, S., 2014. Paying for performance in healthcare: implications for health system performance and accountability. Open University Press, Buckingham, UK.

Chalkley, M., Malcomson, J.M., 1998a. Contracting for health services when patient demand does not reflect quality. Journal of Health Economics 17, 1–19. doi:10.1016/S0167-6296(97)00019-2

Chalkley, M., Malcomson, J.M., 1998b. Contracting for Health Services with Unmonitored Quality. The Economic Journal 108, 1093–1110.

Choné, P., Ma, C.A., 2011. Optimal health care contracts under physician agency. Annals of Economics and Statistics 101/102, 229–256.

Conrad, D.A., Perry, L., 2009. Quality-Based Financial Incentives in Health Care: Can We Improve Quality by Paying for It? Annu. Rev. Public. Health. 30, 357–371.

Curtis, L.A., 2014. Unit costs of health and social care 2014. Personal Social Services Research Unit.

Das, K., Anderson, G., 2007. Premier Hospital Quality Incentive Demonstration. Health Policy Monitor.

Davis, S., Holmes, M., Simpson, E., Sutton, A., 2012. Alteplase for the treatment of acute ischaemic stroke (review of technology appraisal 122). School of Health and Related Research (ScHARR), The University of Sheffield, Sheffield.

Delfgaauw, J., Dur, R., 2008. Incentives and Workers' Motivation in the Public Sector. The Economic Journal 118, 171–191.

Delfgaauw, J., Dur, R., 2007. Signaling and Screening of Workers' Motivation. Journal of Economic Behavior and Organization 62, 605–624.

Department of Health, 2013. Payment by Results Guidance for 2013-14.

Department of Health, 2012a. A simple guide to Payment by Results. Department of Health Payment by Results team, London.

Department of Health, 2012b. Payment by Results Guidance for 2012-13. Leeds.

Department of Health, 2010a. Payment by Results Guidance for 2010-11. Leeds.

Department of Health, 2010b. Step-by-Step Guide: Calculating the 2010/11 National Tariff.

Department of Health, 2008a. Impact Assessment: End of Life Care. London.

Department of Health, 2008b. Implementing the National Stroke Strategy – an imaging guide. London.

Department of Health, 2007a. National Stroke Strategy.

Department of Health, 2007b. Impact Assessment: National Stroke Strategy. London.

Dixit, A., 2005. Incentive contracts for faith-based organisations to deliver social services, in: Lahiri, S., Maiti, P. (Eds.), Economic Theory in a Changing World: Policy Modelling for Growth. Oxford University Press, New Delhi.

Donnan, G.A., Fisher, M., Macleod, M., Davis, S.M., 2008. Stroke. Lancet 371, 1612–1623.

Eggleston, K., 2005. Multitasking and mixed systems for provider payment. Journal of Health Economics 24, 211–223.
Eijkenaar, F., 2013. Key issues in the design of pay for performance programs. Eur J Health Econ 14, 117–131.
Ellis, R.P., McGuire, T.G., 1986. Provider behavior under prospective reimbursement : Cost sharing and supply. Journal of Health Economics 5, 129–151.
Epstein, A.M., 2012. Will Pay for Performance Improve Quality of Care? The Answer Is in the Details. New England Journal of Medicine 367, 1852–1853.
Glazer, A., 2004. Motivating devoted workers. International Journal of Industrial Organization 22, 427–440.
Godager, G., Wiesen, D., 2013. Profit or patients' health benefit? Exploring the heterogeneity in physician altruism. Journal of Health Economics 32, 1105–1116.
Goddard, M., Mannion, R., Smith, P., 2000. Enhancing performance in health care: a theoretical perspective on agency and the role of information. Health Economics 9, 95–107.
Hammond, P., 1987. Altruism. The New Palgrave: A Dictionary of Economics. Macmillan, London 85–87.
Heyes, A.G., 2005. The economics of vocation or why is a badly paid nurse a good nurse'? Journal of Health Economics 24, 561–569.
HM Treasury, 2003. The Green Book: appraisal and evaluation in central government: Treasury guidance. Stationery Office.
Holmstrom, B., Milgrom, P., 1991. Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. Journal of Law, Economics, and Organization 7, 24.
Intercollegiate Stroke Working Party, 2012. National clinical guideline for stroke. Royal College of Physicians of London.
Jack, W., 2005. Purchasing health care services from providers with unknown altruism. Journal of Health Economics 24, 73–93.
Kaarboe, O., Siciliani, L., 2011. Multitasking, quality and pay for performance. Health Economics, Health Economics 20, 225–238.
Kristensen, S.R., Mcdonald, R., Sutton, M., 2013. Should pay-for-performance schemes be locally designed? Evidence from the commissioning for quality and innovation (CQUIN) framework. Journal of Health Services Research & Policy 18, 38–49.
Laffont, J.J., Tirole, J., 1993. A theory of incentives in procurement and regulation. the MIT Press.
Lakdawalla, D., Philipson, T., 2006. The nonprofit sector and industry performance. Journal of Public Economics 90, 1681–1698.
Lyk-Jensen, S.V., 2007. Appraisal methods in the Nordic Countries. Danish Transport Research Institute, Lyngby.
Makris, M., 2009. Incentives for motivated agents under an administrative constraint. Journal of Economic Behavior & Organization 71, 428–440.
Massiani, J., Picco, G., 2013. The Opportunity Cost of Public Funds: Concepts and Issues. Public Budgeting & Finance 33, 96–114.
Maynard, A., 2012. The powers and pitfalls of payment for performance. Health Economics 21, 3–12. doi:10.1002/hec.1810
McDonald, R., Zaidi, S., Todd, S., Konteh, F., Hussain, K., Roe, J., Allen, T., Fichera, E., Sutton, M., 2012. A Qualitative and Quantitative Evaluation of the Introduction of Best Practice Tariffs.
Murdock, K., 2002. Intrinsic motivation and optimal incentive contracts. RAND Journal of Economics 33, 650–671.
Nash, J., 1953. Two-Person Cooperative Games. Econometrica 21, 128–140. doi:10.2307/1906951
National Collaborating Centre for Chronic Conditions, 2008. Stroke: national clinical guideline for diagnosis and initial management of acute stroke and transient ischaemic attack (TIA). Royal College of Physicians, London.
NHS England, 2013. Integrated Performance Measures Monitoring.

NICE, 2012. Alteplase for treating acute ischaemic stroke (review of technology appraisal guidance 122) (NICE technology appraisal guidance No. 264). Manchester.
NICE, 2010. Stroke quality standard (QS2).
Osborne, M.J., Rubinstein, A., 1990. Bargaining and markets. Academic press San Diego.
Petersen, L.A., Woodard, L.D., Urech, T., Daw, C., Sookanan, S., 2006. Does Pay-for-Performance Improve the Quality of Health Care? Annals of Internal Medicine 145, 265–W71. doi:Article
Roland, M., 2012. Pay-for-Performance: Not a Magic Bullet. Ann Intern Med 157, 912–913. doi:10.7326/0003-4819-157-12-201212180-00014
Royal College of Physicians, 2011. National Sentinel Stroke Clinical Audit 2010 Round 7.
Royal College of Physicians, 2009. National Sentinel Stroke Audit Phase II (clinical audit) 2008: Prepared on behalf of the Intercollegiate Working Party.
Royal College of Physicians, 2007. National Sentinel Stroke Audit Phase I (organisational audit) 2006 Phase II (clinical audit) 2006: Prepared on behalf of the Intercollegiate Working Party.
Ruggeri, G., 1999. The marginal cost of public funds in closed and small open economies. Fiscal Studies 20, 41–60.
Ryan, A., 2009. Hospital-based pay-for-performance in the United States. Health Economics 18, 1109–1113.
Saka, Ö., McGuire, A., Wolfe, C., 2009. Cost of stroke in the United Kingdom. Age Ageing 38, 27–32. doi:10.1093/ageing/afn281
Shah, K., Praet, C., Devlin, N., Sussex, J., Appleby, J., Parkin, D., 2012. Is the aim of the English health care system to maximize QALYs? J Health Serv Res Policy 17, 157–163.
Shleifer, A., 1985. A theory of yardstick competition. RAND Journal of Economics 16, 319–327.
Siciliani, L., Strume, O.R., Cellini, R., 2013. Quality competition with motivated providers and sluggish demand. Journal of Economic Dynamics and Control 37, 2041–2061.
Sutton, M., Nikolova, S., Boaden, R., Lester, H., McDonald, R., Roland, M., 2012. Reduced Mortality with Hospital Pay for Performance in England. N Engl J Med 367, 1821–1828.
The Intercollegiate Stroke Working Party, 2013. SINAP - Combined Quarterly Public Report (Quarters 1-7).
Wardlaw, J.M., Keir, S., Seymour, J., Lewis, S., Sandercock, P., Dennis, M., Cairns, J., 2004a. What is the best imaging strategy for acute stroke? Health Technology Assessment 8.
Wardlaw, J.M., Seymour, J., Cairns, J., Keir, S., Lewis, S., Sandercock, P., 2004b. Immediate Computed Tomography Scanning of Acute Stroke Is Cost-Effective and Improves Quality of Life. Stroke 35, 2477–2483.

Appendix

Table A.1: Best practice tariff for emergency stroke

BPT component	Description
Stroke care delivered within an acute stroke unit	Patients are admitted directly (intending to be within 4 hours of arrival in hospital) to an acute stroke unit (Or similar facility where the patient can expect to receive the service set out in quality marker 9 of the National Stroke Strategy[16]) either by the ambulance service, from A&E or via brain imaging. Patients should not be directly admitted to a Medical Assessment Unit. Patients should then also spend the majority (Defined as greater than or equal to 90% of the patient's stay within the spell that groups to either AA22Z or AA23Z) of their stay in the acute stroke unit.
Urgent brain imaging for all suitable patients	Initial brain imaging is delivered in accordance with best practice guidelines as set out in Implementing the National Stroke Strategy – An Imaging Guide (Department of Health, 2008b) . The scan should not only be done in these timescales but immediately interpreted and acted upon by a suitably experienced physician or radiologist. A CT scan should be undertaken urgently if (a) indication for thrombolysis/anticoagulation (b) On anticoagulants and/or known bleeding tendency (c) Depressed level of consciousness (GCS<13) (d) Unexplained fluctuating or progressive symptoms (e) Severe headache at onset (e) Papilloedema, neck stiffness, fever. Otherwise a CT scan should be performed within 24 hours. An MRI scan should be performed if (a) Diagnostic uncertainty after CT (e.g. suspected non stroke pathology but unsure) (b) Atypical clinical presentation including: (b.1) "Young" stroke (<50 years) (b.1) Strong clinical suspicion of vessel dissection (c) Delayed clinical presentation (>7 days after symptom onset)
Alteplase	Patients are assessed for thrombolysis, receiving it if clinically indicated in accordance with the NICE technology appraisal guidance on alteplase. Alteplase is a drug that can dissolve the blood clot. It must be administered within 4.5 hours from the onset of symptoms and should not be given to patients if brain imaging has indicated that the patient has a bleeding in the brain.

Source: Department of Health (2012b) and Department of Health (2008b).

[16] (a) all stroke patients have prompt access to an acute stroke unit and spend the majority of their time at hospital in a stroke unit with high-quality stroke specialist care (b) hyper-acute stroke services provide, as a minimum, 24-hour access to brain imaging, expert interpretation and the opinion of a consultant stroke specialist, and thrombolysis is given to those who can benefit (c) specialist neuro-intensive care including interventional neuroradiology or neurosurgery expertise is rapidly available (d) specialist nursing is available for monitoring of patients (e) appropriately qualified clinicians are available to address respiratory, swallowing, dietary and communication issues (Department of Health, 2007a).

Table A.2: Key characteristics of stroke units and acute stroke units

Key characteristics of stroke units	Key characteristics of acute stroke units
• Consultant physician with responsibility for stroke • Formal links with patient and carer organisations • Multidisciplinary meetings at least weekly to plan patient care • Provision of information to patients about stroke • Continuing education programmes for staff	• Continuous physiological monitoring (ECG, oximetry, blood pressure) • Access to scanning within 3 hours of admission • if not 3 hours, access to 24 hour brain imaging • Policy for direct admission from A&E • Specialist ward rounds at least 5 times a week • Acute stroke protocols/guidelines

Source: Royal College of Physicians (2007).

Table A.3: Optimal performance payment (£) for treatment in an Acute Stroke Unit at different levels of treatment benefit

		Benefit estimate										
		18880	*19824*	*20768*	*21712*	*22656*	*23600*	*24544*	*25488*	*26432*	*27376*	*28320*
Relative weight given to patients' benefit	*0*	18506	19450	20394	21338	22282	23226	24170	25114	26058	27002	27946
	0.1	16618	17467	18317	19167	20016	20866	21715	22565	23415	24264	25114
	0.2	14730	15485	16240	16995	17751	18506	19261	20016	20771	21527	22282
	0.3	12842	13503	14163	14824	15485	16146	16807	17467	18128	18789	19450
	0.4	10954	11520	12087	12653	13219	13786	14352	14919	15485	16051	16618
	0.5	9066	9538	10010	10482	10954	11426	11898	12370	12842	13314	13786
	0.6	7178	7555	7933	8311	8688	9066	9443	9821	10199	10576	10954
	0.7	5290	5573	5856	6139	6423	6706	6989	7272	7555	7839	8122
	0.8	3402	3591	3779	3968	4157	4346	4535	4723	4912	5101	5290
	0.9	1514	1608	1703	1797	1891	1986	2080	2175	2269	2363	2458
	1	-374	-374	-374	-374	-374	-374	-374	-374	-374	-374	-374

Note: Assuming per person MC of treatment in Acute Stroke Unit=£1871 QALYs, Social value of a QALY =50,000, Lambda= 0.2

Table A.4: Optimal performance payment (£) for rapid brain imaging at different levels of treatment benefit

		Benefit estimate										
		5	*5.5*	*6*	*6.5*	*7*	*7.5*	*8*	*8.5*	*9*	*9.5*	*10*
Relative weight given to patients' benefit	*0*	2	2	3	3	4	4	5	5	6	6	7
	0.1	1	2	2	3	3	4	4	5	5	5	6
	0.2	1	1	2	2	2	3	3	4	4	4	5
	0.3	0	1	1	1	2	2	2	3	3	4	4
	0.4	0	0	0	1	1	1	2	2	2	3	3
	0.5	-1	0	0	0	0	1	1	1	1	2	2
	0.6	-1	-1	-1	-1	0	0	0	0	0	1	1
	0.7	-2	-1	-1	-1	-1	-1	-1	-1	0	0	0
	0.8	-2	-2	-2	-2	-2	-2	-2	-1	-1	-1	-1
	0.9	-3	-3	-3	-2	-2	-2	-2	-2	-2	-2	-2
	1	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3

Note: Assuming per person MC of rapid brain imaging=£16, Social value of a QALY =50,000, Lambda = 0.2

Table A.5: Optimal performance payment for thrombolysis with alteplase at different levels of treatment benefit

		Benefit estimate										
		11750	*12630*	*13510*	*14390*	*15270*	*16150*	*17030*	*17910*	*18790*	*19670*	*20550*
Relative weight given to patients' benefit	*0*	11582	12462	13342	14222	15102	16482	17362	18242	19122	20002	20382
	0.1	10407	11199	11991	12783	13575	14817	15609	16401	17193	17985	18327
	0.2	9233	9937	10641	11345	12049	13153	13857	14561	15265	15969	16273
	0.3	8058	8674	9290	9906	10522	11488	12104	12720	13336	13952	14218
	0.4	6884	7412	7940	8468	8996	9824	10352	10880	11408	11936	12164
	0.5	5709	6149	6589	7029	7469	8159	8599	9039	9479	9919	10109
	0.6	4535	4887	5239	5591	5943	6495	6847	7199	7551	7903	8055
	0.7	3360	3624	3888	4152	4416	4830	5094	5358	5622	5886	6000
	0.8	2186	2362	2538	2714	2890	3166	3342	3518	3694	3870	3946
	0.9	1011	1099	1187	1275	1363	1501	1589	1677	1765	1853	1891
	1	-163	-163	-163	-163	-163	-163	-163	-163	-163	-163	-163

Note: Assuming per person MC of treatment with alteplase=816, Social value of a QALY =50,000, Lambda = 0.2

Table A.6: Optimal performance payment (£) for treatment in an Acute Stroke Unit at different levels of treatment costs

		Cost estimate										
		1497	*1572*	*1647*	*1722*	*1797*	*1871*	*1946*	*2021*	*2096*	*2171*	*2246*
Relative weight given to patients' benefit	*0*	23301	23286	23271	23256	23241	23226	23211	23196	23181	23166	23151
	0.1	20941	20926	20911	20896	20881	20866	20851	20836	20821	20806	20791
	0.2	18581	18566	18551	18536	18521	18506	18491	18476	18461	18446	18431
	0.3	16221	16206	16191	16176	16161	16146	16131	16116	16101	16086	16071
	0.4	13861	13846	13831	13816	13801	13786	13771	13756	13741	13726	13711
	0.5	11501	11486	11471	11456	11441	11426	11411	11396	11381	11366	11351
	0.6	9141	9126	9111	9096	9081	9066	9051	9036	9021	9006	8991
	0.7	6781	6766	6751	6736	6721	6706	6691	6676	6661	6646	6631
	0.8	4421	4406	4391	4376	4361	4346	4331	4316	4301	4286	4271
	0.9	2061	2046	2031	2016	2001	1986	1971	1956	1941	1926	1911
	1	-299	-314	-329	-344	-359	-374	-389	-404	-419	-434	-449

Note: Assuming per person MB of treatment in Acute Stroke Unit=0.479 QALYs, Social value of a QALY =50,000, Lambda= 0.2

Table A.7: Optimal performance payment (£) for rapid brain imaging at different levels treatment costs

		Cost estimate										
		10	*14*	*17*	*20*	*24*	*27*	*30*	*34*	*37*	*41*	*44*
Relative weight given to patients' benefit	*0*	3	2	2	1	0	0	-1	-2	-2	-3	-4
	0.1	2	2	1	0	0	-1	-2	-2	-3	-4	-4
	0.2	2	1	1	0	-1	-1	-2	-3	-3	-4	-5
	0.3	1	1	0	-1	-1	-2	-3	-3	-4	-5	-5
	0.4	1	0	0	-1	-2	-2	-3	-4	-4	-5	-6
	0.5	0	0	-1	-2	-2	-3	-4	-4	-5	-6	-6
	0.6	0	-1	-1	-2	-3	-3	-4	-5	-5	-6	-7
	0.7	-1	-1	-2	-3	-3	-4	-5	-5	-6	-7	-7
	0.8	-1	-2	-2	-3	-4	-4	-5	-6	-6	-7	-8
	0.9	-2	-2	-3	-4	-4	-5	-6	-6	-7	-8	-8
	1	-2	-3	-3	-4	-5	-5	-6	-7	-7	-8	-9

Note: Assuming per person MB of rapid brain imaging=0.0001 QALYs, Social value of a QALY =50,000, Lambda = 0.2

Table A.8: Optimal performance payment for thrombolysis with alteplase at different levels of treatment costs

		Cost estimate										
		522	*579*	*635*	*692*	*749*	*805*	*862*	*918*	*975*	*1031*	*1088*
Relative weight given to patients' benefit	*0*	16541	16529	16518	16507	16495	16484	16473	16461	16450	16439	16427
	0.1	14876	14865	14853	14842	14831	14819	14808	14797	14786	14774	14763
	0.2	13212	13200	13189	13178	13166	13155	13144	13132	13121	13110	13098
	0.3	11547	11536	11524	11513	11502	11490	11479	11468	11457	11445	11434
	0.4	9883	9871	9860	9849	9837	9826	9815	9803	9792	9781	9769
	0.5	8218	8207	8195	8184	8173	8161	8150	8139	8128	8116	8105
	0.6	6554	6542	6531	6520	6508	6497	6486	6474	6463	6452	6440
	0.7	4889	4878	4866	4855	4844	4832	4821	4810	4799	4787	4776
	0.8	3225	3213	3202	3191	3179	3168	3157	3145	3134	3123	3111
	0.9	1560	1549	1537	1526	1515	1503	1492	1481	1470	1458	1447
	1	-104	-116	-127	-138	-150	-161	-172	-184	-195	-206	-218

Note: Assuming per person MB of treatment in Acute Stroke Unit=0.479 QALYs, Social value of a QALY =50,000, Lambda = 0.2

Appendix

The three optimality conditions for the hospital and the regulator are respectively

$$\alpha b^1 + p^1 = c^1 + K^{1\prime}\left(N^{1*} + \frac{1}{\gamma}N^{2*}\right), \tag{A1}$$

$$\alpha(b^{12} - b^1) + p^2 = c^2 + K^{2\prime}(N^{2*}) \tag{A2}$$

$$\alpha b^3 + p^3 = c^3 + K^{3\prime}(N^{3*}). \tag{A3}$$

$$b^1 - \lambda\left(c^1 + K^{1\prime}\left(N^{1f} + \frac{1}{\gamma}N^{2f}\right)\right) = c^1 + K^{1\prime}\left(N^{1f} + \frac{1}{\gamma}N^{2f}\right), \tag{A4}$$

$$b^{12} - b^1 - \lambda\left(c^2 + K^{2\prime}\left(N^{2f}\right)\right) = c^2 + K^{2\prime}\left(N^{2f}\right) \tag{A5}$$

$$b^3 - \lambda\left(c^3 + K^{3\prime}\left(N^{3f}\right)\right) = c^3 + K^{3\prime}\left(N^{3f}\right). \tag{A6}$$

By equating the LHS of (A1)-(A3), with the LHS of (A4)-(A6), we obtain

$$\alpha b^1 + p^1 = b^1 - \lambda\left(c^1 + K^{1\prime}\left(N^{1f} + \frac{1}{\gamma}N^{2f}\right)\right),$$

$$\alpha(b^{12} - b^1) + p^2 = (b^{12} - b^1) - \lambda\left(c^2 + K^{2\prime}\left(N^{2f}\right)\right),$$

$$\alpha b^3 + p^3 = b^3 - \lambda\left(c^3 + K^{3\prime}\left(N^{3f}\right)\right).$$

The optimal prices are obtained:

$$p^{1f} = b^1(1-\alpha) - \lambda\left(c^1 + K^{1\prime}\left(N^{1f} + \frac{1}{\gamma}N^{2f}\right)\right),$$

$$p^{2f} + \frac{1}{\gamma}p^{1f} = \left(b^{12} + \frac{1-\gamma}{\gamma}b^1\right)(1-\alpha) - \lambda\left(c^2 + K^{2\prime}\left(N^{2f}\right)\right),$$

$$p^{3f} = b^3(1-\alpha) - \lambda\left(c^3 + K^{3\prime}\left(N^{3f}\right)\right).$$