An economic model of the NHS under Payment by Results and Patient Choice

Gavin Roberts, Rob Unsworth

Department of Health, Quarry House, Leeds

Paper presented at the 69th Health Economics Study Group, University of York, 26-28 July 2006

Draft. Not for quotation or circulation

Abstract

Aims
NHS funding in England has changed from fixed hospital budgets to payment per unit of activity ("Payment by Results"). Also patients are now able to choose among competing providers ("Patient Choice"). This study seeks to understand the effect of these policies on the quantity and quality of care produced in the NHS by modelling changes in supplier behaviour under the new incentives implied by reforms.

Methods
We develop a formal model of the demand for treatment, and supplier behaviour. Equilibrium activity and quality levels are compared when funding changes from block contracts to national tariff payments under Payment by Results, and when Patient Choice introduces competition among suppliers.

Results
Preliminary findings show that Payment by Results alone is unlikely to change the quality of care, and activity cannot increase without exceeding commissioner budgets. However introduction of competition under Patient Choice drives quality toward the maximum possible, given the tariff available. In perfect competition, the entire tariff is diverted to maximising quality, which is the optimum outcome for society under this system. However many classical market failures prevent this outcome, including information problems, principal-agent problems and factor market failures.

Conclusions
Our model provides a general framework for understanding the impacts of current reforms on health production and consumption in the NHS. It defines society's optimal outcome under the new system and identifies the market failures which prevent this result from being realised. The role of government in managing the new system must be in mitigating these failures as far as possible.

Introduction

The NHS in England has introduced a system of paying hospitals and other providers on the basis of the work they do. Hospitals receive a fixed tariff each time they provide a treatment. Different treatments attract different fixed tariffs. Termed "Payment by Results" (PbR), the policy rewards providers for volumes of work adjusted for differences in casemix. In addition to changing the funding mechanism, health provision is being opened to competition between providers – both NHS hospitals and private suppliers – by the policy of "Patient Choice" whereby patients are able to select the provider they wish to treat them.

Prior to Payment by Results, transactions between purchasers (Primary Care Trusts, PCTs) and providers stipulated a total bulk contract value, usually specified at specialty level. The PCT would decide how much of its budget to devote to its contract with each of its providers and negotiate with them how much activity would be made available. The price per unit of activity was arrived at as the by-product of negotiations about total contract value and the volume of activity. These arrangements allowed for tight control of expenditure but crucially provided disincentives for actually treating patients, because providers' marginal revenue for additional treatments was zero. Providers effectively operated with fixed budgets and made their own decisions on the quantity and quality of treatment to offer – which may explain the development of significant waiting lists.

Under Payment by Results budgets are no longer given to providers in fixed contracts, but instead are managed directly by the PCTs, who act on patients' behalf to purchase treatments from providers at a marginal unit price – the tariff – set nationally based on average costs across providers. Providers' revenues are therefore directly dependent on the number of treatments they carry out, giving them strong incentives to undertake more activity.

Payment by Results is not just an end in itself, but it also provides the structure necessary for enabling competition among providers. The introduction of Patient Choice means that, in principle, hospitals operate in a market where they must compete to attract commissions and earn revenues. Because prices are fixed, competition must occur on other bases – these may include location, waiting times and service quality.

This paper seeks to understand the effect of Payment by Results and Patient Choice policies on the quantity and quality of care produced in the NHS by modelling changes in supplier behaviour under the new incentives implied by reforms. We first examine the situation before Payment by Results, where providers received fixed

budgets, and effectively determined the quality and quantity of treatments supplied. The transition to Payment by Results is then modelled for the hospitals retaining an effective monopoly, and then for the introduction of Patient Choice and consequent competition among suppliers.

Accounting for treatment quality and efficiency

In attempting to describe the output of hospitals and the NHS as a whole, it is important to consider not only the quantity of treatments supplied and consumed, but also the quality of those treatments. Quality is multi-dimensional, and may be considered to include aspects of clinical effectiveness and patient experience. Here we consider quality to be associated only with the treatment provided and not with the waiting time for that treatment, or the distance the patient must travel to receive it. In this section we develop a general description of treatments in which a single measure of quality is determined by the amount of spending at maximum efficiency.

The cost, c, of any actual treatment can be conceptually divided into effective spending, e, and waste, w.

$$c = e + w$$

Here the effective spending is the minimum expenditure necessary to provide the treatment, by using the most efficient possible production technique. The residue, w, can be considered to be expenditure that is wasted due to inefficiency.

We consider that treatment quality, q, can be represented by the effective spending manifested in the treatment[1], e. That is

$$q = e$$

and

$$c = q + w$$

Therefore if a treatment of a certain quality was to cost £1200, but could be provided at the same quality for £800 using the most efficient possible technique, then the effective spending is considered to be £800 and £400 has been wasted. The treatment would be of higher quality if it manifested more than £800 of effective spending, so the measure of quality embodied in the treatment can be represented as £800.

[1] If quality is a utility function reflecting the preferences of fully informed consumers, and the effective spending of a treatment, e, is the minimum expenditure required to generate that quality of treatment, then q is a monotonic transformation of e. As q is an ordinal function, it is therefore possible to say that the quality level of a treatment is represented exactly by the effective spending manifested in that treatment, or $q = e$. So if one treatment embodies higher effective spending than another, it must be preferred by patients, and must be of greater quality. However as q is ordinal, it is not possible to say that increasing effective spending twofold will result in a treatment of double the quality.

Demand for treatment

The number of treatments demanded by patients, *D*, is an increasing function of quality, or effective spending. If the effective spending embodied in a treatment is zero, and quality is zero, patients will derive no utility from it, and will demand no treatments. As effective spending and quality increase, so patients will demand more treatments. This is shown in figure 1.

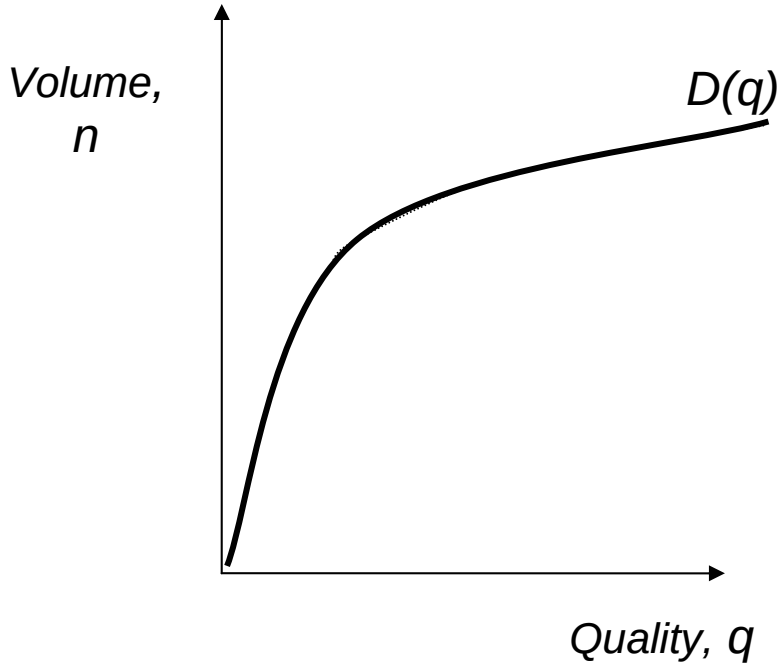

Figure 1. The demand for treatments, *D*, is increasing with the quality of treatments, *q*.

Supplier behaviour

Suppliers are considered to gain utility from two sources: satisfaction of intrinsic motivation *I*, and discretionary spending *W*.

$$U = U(I, W)$$

Satisfaction of intrinsic motivation represents the utility gained from curing patients, maintaining a high standard of ethics and complying with targets and other professional requirements. Satisfaction of intrinsic motivation, *I*, is an increasing function of the number of treatments, *n*, and their quality, *q*. Here we use a simple functional form for *I* where *n* and *q* enter multiplicatively[2].

$$I(qn) = qn \qquad I_{qn} \geq 0$$

Discretionary spending refers to the diversion of expenditure from the most efficient patient care to fund activities that directly satisfy the provider. These may include spending on personal projects, overspending on interesting care, tolerance of problems that reduce efficiency and other rent-extraction behaviours. Discretionary spending is defined as the gap between the supplier's income and their total effective spending on treatment:

$$W = y - qn$$

Discretionary spending is also equal to the total waste, *nw*.

The supplier utility function is therefore

$$U = U(qn, y - qn)$$

Suppliers will trade off satisfaction of intrinsic motivation and discretionary spending. Convex preferences over these sources of utility can be simulated with the Cobb-Douglas form:

$$U = (qn)^{a}(y - qn)^{1-a} \qquad 0 \leq a \leq 1$$

Here *a* is a measure of the weighting the provider gives to satisfying intrinsic motivation, against discretionary spending.

[2] Here providers are considered, for convenience, to be indifferent between combinations of *n* and *q* with the same product. This is clearly not true in reality.

Equilibrium output under direct provider budgets (before Payment by Results)

We now consider a health market with a single provider, and calculate the quantity and quality of treatments supplied before Payment by Results, when providers received direct budget B (i.e. income $y = B$) regardless of their activity.

We consider that the provider chooses a level of quality, and the quantity of treatments is determined by demand at that level of quality[3]. That is,

$$n = D(q)$$

The provider is assumed to know the demand schedule of patients, so for any level of q chosen, it knows the level of n that will result.

The provider will choose the level of quality that maximises utility, subject to the constraint that their total spending cannot exceed the budget B. That is, the quantity of treatments multiplied by the effective spending, or quality, of treatments cannot exceed their available budget.

$$B \geq en = qn = qD(q)$$

We can now write out the provider's utility maximisation problem:

$$\max_q U = (qD(q))^a (B - qD(q))^{1-a} \quad s.t.\ qD(q) \leq B$$

First and second order conditions (see Appendix B) indicate that utility is maximised, at q_o, where

$$q_o D(q_o) = aB$$

or, where n_o represents the number of treatments

$$q_o n_o = aB$$

[3] It would be more realistic to consider that providers choose both n and q, subject to the constraint that n must not exceed demand at that quality. Any excess demand would be eliminated by formation of waiting lists, which would reduce demand at that level of quality and clear the market (as observed in the NHS before PbR). However our choice of supplier utility function does not give solutions for providers choosing over both n and q so we simply assume that there is no excess demand, without formally demonstrating why this is so. This does result in a restriction of the possible combinations of n and q that can arise before PbR, but it does not affect the key results. It would be desirable to find a utility function that enables suppliers to choose both n and q, but still permits solutions.

Result: a provider placing any weight on discretionary spending will not supply at maximum efficiency

The provider divides the budget *B* between effective spending and waste. The division is determined by the weighting *a* the provider places on intrinsic motivation against discretionary spending. When $a = 0$, all the budget goes to discretionary spending or waste ($B = n_o w_o$), and when $a = 1$ all the budget is spent at maximal efficiency, and there is no waste ($B = n_o q_o$). For all values of *a* less than 1, corresponding to suppliers putting some weighting on discretionary spending, output will be less than the maximum possible efficiency – that is, for a given volume, the budget could be used for higher quality of treatment, or for a given quality of treatment, more patients could be treated if efficiency was improved. This is shown graphically in Figure 2.

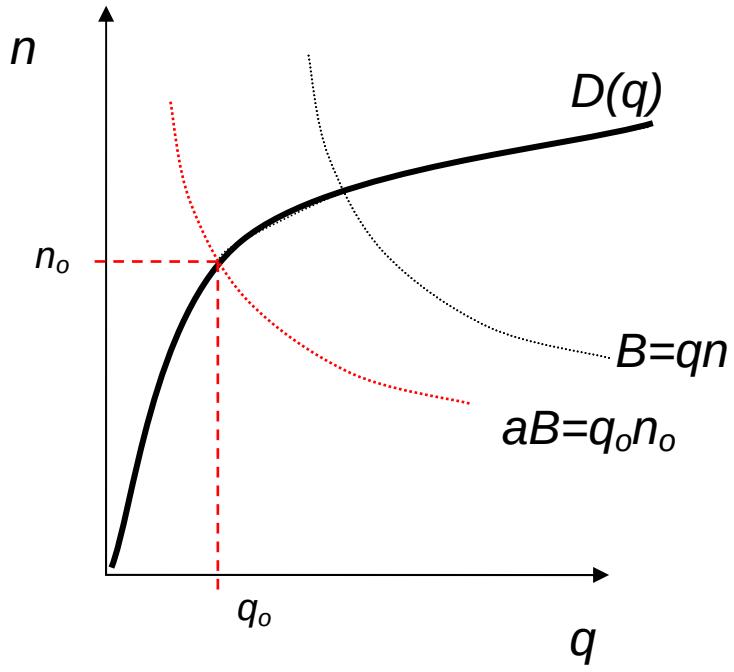

Figure 2. A hospital faces demand for treatments *D(q)*. It receives budget *B*, which constrains the combinations of quality *q* and quantity *n* that can be afforded. It chooses quality q_o to maximise utility, resulting in patients demanding n_o treatments. If *a* equals 1, then the total spent on quality, *qn*, is equal to the budget *B*. If *a* is less than 1, *B* will not all be spent on quality *qn*, and there will be some waste. If *a* were zero, providers would choose to spend nothing on quality, no treatments would be demanded or supplied and the entire budget would be diverted to waste, or discretionary spending.

Under Payment by Results the budget passes to commissioners, leading to an activity constraint

We now consider the situation under Payment by Results, where the hospital no longer receives a direct budget. Funds are instead given to the commissioner who purchases care on behalf of patients. The tariff of transaction, *t*, is fixed at the average cost of treatment before Payment by Results:

$$t = \frac{B}{n_o}$$

The commissioner now holds the budget constraint. If there is only one type of treatment, then the fixed tariff implies a limit on activity affordable under the fixed budget, n_{max}.

$$n_{\max} = \frac{B}{t}$$

But tariff is the previous average cost of treatments (above), so

$$n_o = \frac{B}{t} \qquad \text{and} \qquad n_{\max} = n_o$$

so the commissioner is unable to afford treatments beyond the volume provided before PbR.

We have here considered a market with a single supplier, but the analysis and results are the same for multiple suppliers, as the activity limit is derived from the commissioner's constraint which is independent of the number of suppliers. Total affordable activity will always be equal to the total budget divided by the tariff. If the tariff is set at previous average cost, then the total activity constraint must always equal the previous total activity level.

Result: activity cannot increase after PbR

Equilibrium output under Payment by Results with supplier monopoly

We now consider the behaviour of the unique supplier after the implementation of Payment by Results.

Before Payment by Results the provider generated n_o treatments at quality q_o with budget B. The commissioner, constrained to budget B, can afford to purchase a maximum volume of treatments $n_{\max}=n_o$.

A perfect commissioner will represent patients' needs (i.e. the demand curve $D(q)$) up to n_{max} but will not buy any further treatments beyond this level. The demand curve facing the provider will therefore now be bi-phasic (see Figure 3).

Under PbR, provider income y is equal to activity times tariff, tn. This is the key change in their incentives: income, and discretionary spending, is now dependent on activity. Their utility is now

$$U = (qn)^a (tn - qn)^{1-a}$$

or, by rearrangement

$$U = nq^a (t - q)^{1-a}$$

It is immediately clear that while $t > q$ the provider will choose arbitrarily large volumes of activity, as utility is strictly increasing in n. The utility function can therefore be rewritten replacing n with $\hat{D}(q)$, the greatest possible value of n for a given value of q, as defined by the constraints of

- patient demand $D(q)$ (binding for values of $q < q_o$)
- the commissioner's budget constraint $n_{\max}$ (binding for values of $q > q_0$

This gives the problem

$$\max_q U = \hat{D}(q) q^a (t - q)^{1-a}$$

Where

$$\hat{D}(q) = \begin{cases} D(q), q \le q_o \\ n_o, q_o \le q \end{cases}$$

Solving first and second order conditions (see Appendix B) gives the result that providers will choose quality $q_m = q_o$, leading to activity levels of $n_m = n_{max} = n_o$. This is shown graphically in Figure 3.

Result: Quantity and quality of output after PbR will be unchanged in a market with a single monopoly provider.

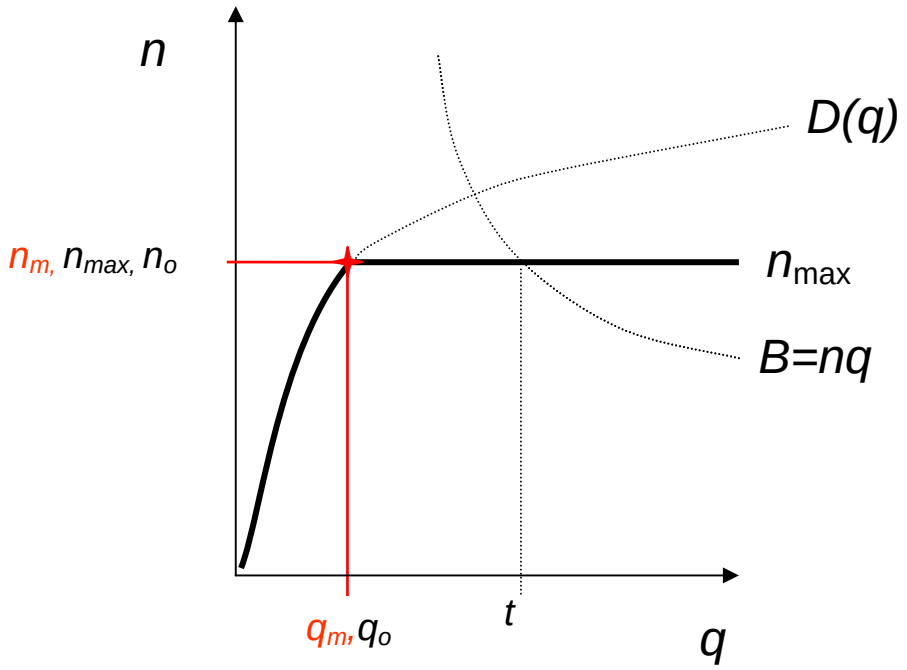

Figure 3. Under PbR, the commissioner inherits budget *B* with which they purchase treatments at tariff *t* from a monopoly provider. The budget constrains activity at $B/t = n_{max} = n_o$, i.e. at the same level as before PbR. The demand schedule is therefore biphasic, being determined by patient demand up to the commissioner's activity limit n_{max}. The provider chooses quality q_m to maximise utility, and it is shown that this is equal to q_o, the quality level before PbR. Therefore in a market with a single provider the same number of treatments are supplied as before PbR, at the same level of quality.

We now consider the results if the aggregate health market comprises multiple providers, all with local monopolies. If all providers operated at the same efficiency before PbR - that is treatments cost the same at all providers - then the result observed

for a single provider will be simply multiplied up to the level of the aggregate market, resulting in unchanged quantity and quality before and after PbR.

If providers before PbR had varying costs, and the tariff is set at the average of these, commissioners still inherit an activity constraint set at the total budget divided by the tariff, and aggregate volumes cannot exceed the level before PbR. However individual providers will be allowed (or compelled) to increase or decrease activity, depending on whether they operated at low or high cost before PbR. The potential increases and required decreases will exactly offset each other but it can be shown that, for some demand schedules, the potential increases will not be fully realised, and there will be a net reduction in activity after PbR – though never an increase. It is also possible that quality supplied will change, but the direction and magnitude again depend on the demand schedule. These effects are discussed more fully in Appendix A.

Competitive equilibrium

We now consider the effect of introducing competition between providers. For convenience, we assume that the single provider previously analysed is replaced by multiple providers competing for patients, and calculate the aggregate output of the market.

When patients are able to choose between providers, they will always select the highest quality offering (assuming they can determine which this is). Providers able to increase their quality beyond that of competitors will take all the market, and all others will receive no revenue. Providers will therefore be induced to increase q until it equals the tariff, t. At this point they are diverting all the tariff to driving up quality through effective spending, while waste is totally eliminated. This result is depicted graphically in Figure 4.

Result: competition under patient choice results in the maximum possible quality, for the tariff available.

This Bertrand style competition represents the best possible outcome for society, in that all the health budget is spent on maximising the quality of treatments[4].

Note that volumes are still limited by the commissioner's activity constraint, determined by the budget and tariff, so overall levels will be the same as before PbR.

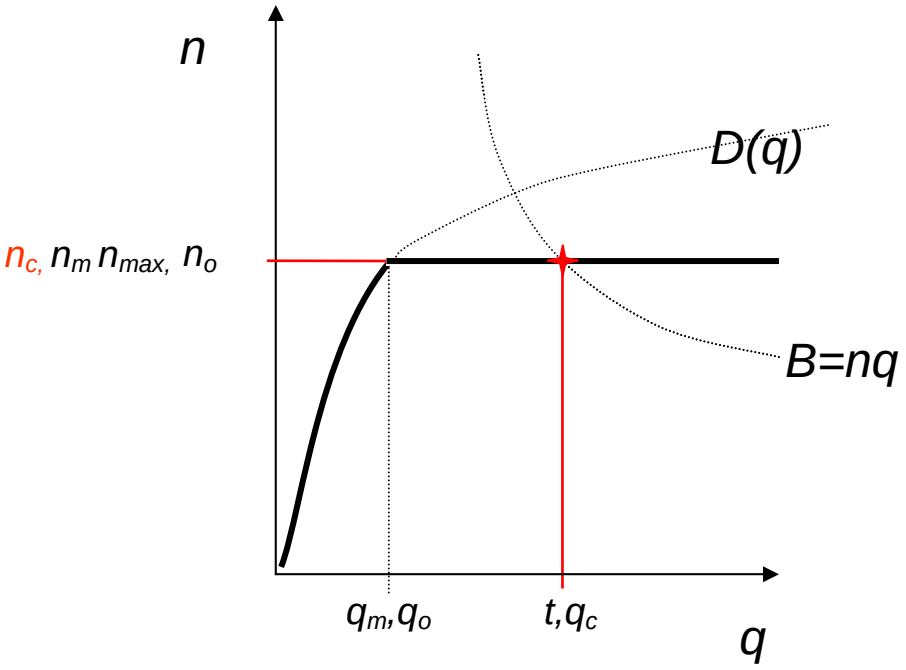

Figure 4. With perfect competition, providers gain no revenue unless they offer the maximum possible quality, so tariff is all diverted to effective spending, and quality $q_c = t$. Activity is unchanged at n_{max}.

[4] It is still possible that a Pareto improvement could be made by varying the tariff, so that the budget is spent on a different combination of *n* and *q*, but this model does not indicate where the overall optimum would lie.

Extensions

The model presented shows the base-case effect of PbR and Choice, but there are many important extensions possible, including analysis of the following situations:

- Use of alternative means of setting the tariff, to achieve desired quantity and quality levels
- The behaviour and impact of hospitals with high and low costs before PbR
- Variation in the weighting (*a*) put on satisfying intrinsic motivation against discretionary spending for different types of provider. For instance acute trusts may have a higher value of *a* than independent sector providers who can extract surplus directly as profit
- The implementation of practice-based commissioning, and the possibility of trading between different treatments

Market failures and distortions

The model defines the optimum outcome for society, given the rules of PbR and Choice, and shows that a perfectly functioning market will reach this outcome. However there are many significant market failures which will prevent this outcome being realised:

- Information problems

  Incomplete understanding of patient needs
  Incomplete understanding of supplier offering

- Agency problems

  Commissioning: providers influence commissioning decision
  Commissioning: inefficient patient prioritisation
  Commissioning: failure to maintain budget constraint
  Suppliers: failure to control labour

- Natural monopoly

  Local monopoly
  Speciality monopoly
  Monopoly through ineffective choice

- Factor market failures

  Capital market failures
  Labour market failures

- Regulatory failures and distortion of supplier incentives

  Imperfect financial governance of supplier organisations
  Barriers to entry of suppliers

Barriers to exit of suppliers

- Externalities

  Public Health externality
  Basic Research externality

The role of government is to identify and mitigate these failures, and this model may be used to understand them, and design policies to optimise the function of the NHS after Health Reform.

Conclusions

The model shows that introduction of Payment by Results cannot increase the aggregate volume of activity in the NHS, because the calculation of tariff based on previous average costs fixes the amount of activity affordable with a given budget at the previous volume. Under the conditions analysed, quality will also be unchanged as a result of PbR.

The major benefit of the policies analysed is the effect of Patient Choice in driving competition based on quality. It is shown that perfect competition will result in the whole tariff being used at maximal efficiency, to produce the greatest possible quality. This insight focusses attention on the imperative for addressing failures in the newly created market for health, in order to deliver the maximum benefit for patients and taxpayers.

Acknowledgements

We are grateful to Andrew Street and Marisa Miraldo for discussions of the model and paper.

Appendix A - Effects of cost variation at providers

The analysis of the transition to PbR has been carried out for a single provider, showing that its activity and quality will be unchanged after PbR, if commissioners are to spend no more than the provider's budget before PbR. The results presented are true of the entire system if all providers have the same costs before PbR, as they would behave exactly as the unique suppliers described (before Patient Choice). However if providers have variable costs before PbR, they will experience effective increases or decreases in the price of treatment, as tariffs based on national average costs are adopted. This has implications for the activity constraints of commissioners dealing with these providers, and can affect the aggregate activity and the average quality in the system as a whole. This section presents a brief informal analysis of these effects.

Providers with low costs before PbR will experience an effective increase in the price of their services, as the commissioners who inherit their budget now have to pay the tariff level for treatments, which is greater than the previous implicit unit price. At this higher effective price, commissioners can afford to purchase fewer treatments with the budget they inherit than were previously supplied. This activity limit will reduce the maximum volume that can be supplied to below the previous activity. It can be shown that activity levels will not be reduced any further than this new maximum when providers choose new levels of quality to maximise their utility.

Conversely, providers with high costs before PbR will experience an effective reduction in the price of services, enabling commissioners to purchase greater quantities of treatment with the previous level of budget. It can be shown that this increase exactly offsets the decrease at low cost trusts. Although providers' utility is strictly increasing in volume, they must raise quality in order to increase demand for their services to realise this potential growth in activity. Whether utility is maximised at the commissioner's maximum activity or at a lower level depends on the exact nature of the demand schedule. Therefore activity may not reach the commissioner's maximum.

It can therefore be seen that while cost variation will lead to exactly offsetting maximum activity levels, such that the total maximum activity is equal to the activity

before Payment by Results, and that low cost providers will be forced to move to their new, lower maximum level, high cost providers may not choose to supply the full activity permitted by commissioners. Therefore, while activity cannot increase under PbR, as shown previously, it is possible that it may decrease.

The net effect of cost variation on quality provided has not been calculated, but it is not inevitable that the effects for high and low cost trusts will offset each other and there will be a variation in quality as a result – though the direction of change is not determined.

Appendix B – Solution to Provider's Utility Maximisation problem, before and after PbR.

Pre PbR:

Stage 1 – Reimbursement method is determined (exogenously)
Stage 2 – Providers have perfect knowledge of patient preferences and set quality, q.
Stage 3 – Patients observe q and consume treatments n=D(q).
Outcome – (q,n*)*

Stage 1

y=B

Stage 2

Providers know that patients will demand D(q) treatments for a quality level of q. The providers' utility function is therefore:

$$U = [qD(q)]^{a}[B - qD(q)]^{1-a} \qquad \text{where } 0 \leq a \leq 1$$

They are only able to set q at a level where the total cost of activity, qD(q), is affordable. That is they cannot overspend their budget, B.

Providers will therefore face the following maximisation problem:

$$\max_{q} U = [qD(q)]^{a}[B - qD(q)]^{1-a}$$
$$\text{st } qD(q) \leq B$$

The first order condition for this maximisation problem is:

$$\frac{\partial U}{\partial q} = a[qD(q)]^{a-1}[B - qD(q)]^{1-a}[qD'(q) + D(q)] - (1-a)[B - qD(q)]^{-a}[qD(q)]^{a}[qD'(q) + D(q)] = 0$$

Solving the first order condition leads to 3 stationary points, where the constraint is satisfied:

1. q*=0
2. q*D(q*)=aB
3. q*D(q*)=B

Evaluating $\frac{\partial^2 U}{\partial q^2}$ at the stationary points shows us that 2 is a maximum and 1 and 3 are minima.

Stage 3

Patients observe the level of quality offered, q*, and consume n*=D(q*).

Outcome

We have a unique solution, (q*,n*) = (q*, D(q*)) = (q_0, n_0), where q*D(q*)=aB

Post Introduction of PbR:

Stage 1 – Reimbursement method is determined (exogenously)
Stage 2 – Providers have perfect knowledge of patient preferences and set quality, q.
Stage 3 – Patients observe q and consume treatments n=D(q).
Outcome – (q_m,n_m)

Stage 1

y=nt

Stage 2

Providers know that patients will demand D(q) treatments for a quality level of q. The providers' utility function is therefore:

$$U = [qD(q)]^a [tD(q) - qD(q)]^{1-a}$$

which simplifies to:

$$U = D(q)q^a [t - q]^{1-a}$$

Commissioners hold the budget, B. The imposition of the tariff means that commissioners will only allow activity levels which are affordable under the fixed budget, B. Hence the provider maximisation problem becomes:

$$\max_q D(q)q^a [t - q]^{1-a}$$
$$\text{st } tD(q) \leq B$$

we can rewrite the maximisation problem as:

$$\max_q \hat{D}(q)q^a [t - q]^{1-a}$$

where, $$\hat{D}(q) = \begin{cases} D(q) & q \leq q_0 \\ n_0 & q_0 \leq q \end{cases}$$

This is solved in 2 stages, finding stationary points for each of the phases of $\hat{D}(q)$.

Fist Order Condition:

$$\frac{\partial U}{\partial q} = \left[q^a \hat{D}'(q) + aq^{a-1}\hat{D}(q)\right](t - q)^{1-a} - (1 - a)(t - q)^{-a}\left[q^a \hat{D}(q)\right] = 0$$

i. $\hat{D}(q) = D(q)$ *for* $0 \le q \le q_0$

$$\frac{\partial U}{\partial q} = \left[q^a D'(q) + aq^{a-1} D(q)\right](t-q)^{1-a} - (1-a)(t-q)^{-a}\left[q^a D(q)\right] = 0$$

$$\Rightarrow \quad \frac{D'(q)}{D(q)}(t-q) = 1 - \frac{q_0}{q}$$

When $0 \le q \le q_0$, the LHS is always positive, due to the assumption on D(q). However the RHS is always either zero or negative, and so a maximum is never found on the interior of $[0,q_0]$ when $\hat{D}(q) = D(q)$. Utility is increasing in quality over this range of q and so the maximum is located where $q=q_0$.

ii. $\hat{D}(q) = n_0$ *for* $q_0 \le q$

$$\frac{\partial U}{\partial q} = \left[q^a(0) + aq^{a-1} n_0\right](t-q)^{1-a} - (1-a)(t-q)^{-a}\left[q^a n_0\right] = 0$$

$$\Rightarrow \quad q_m = at = q_0 \quad \text{or} \quad q_m = t$$

Evaluating $\frac{\partial^2 U}{\partial q^2}$ at q=at and q=t shows us that $q_m = q_0$ is the maximum.

Hence providers will set $q_m = at = q_0$

Stage 3

Patients observe the level of quality offered, q_m, and consume $n_m = \hat{D}(q_m) = n_0$.

Outcome

We have a unique solution, (q_m, n_m), $(q_m, \hat{D}(q_m)) = (q_0, n_0)$.