Work in progress: Please do not circulate without the authors' discretion

Impact of Hospital Pay-for-Performance Structures on Quality Improvement: Evidence from the Advancing Quality programme in England

Alex Turner, Silviya Nikolova, Matt Sutton
Health Economics, Health Sciences Research Group, University of Manchester
Address for correspondence: alexander.turner@postgrad.manchester.ac.uk

Abstract

Introduction

There is mixed evidence as to whether Pay-for-Performance (P4P) schemes introduced into the healthcare sector have had their desired effect of improving quality of care. There is particular uncertainty over two issues about how these payments should be structured: as bonuses or fines; and as tournaments or benchmarks. This study investigates this with respect to the Advancing Quality (AQ) programme; a P4P scheme introduced in all acute hospital Trusts in the North-West of England in October 2008 as a pure tournament scheme. After the first twelve months of its operation the programme adopted a mixed design with tournament, improvement, and minimum threshold components. From March 2010, AQ became part of a regional Commissioning for Quality and Innovation (CQUIN) scheme. The CQUIN payment structure is a fines structure with locally negotiated thresholds. Relatively little is known as to why present P4P programs fail to meet expectations. Disentangling the impact of different design choices is of major interests to both economists and policymakers. This is the primary question of interest in this paper.

Data and Methods

We use quarterly data on 28 performance indicators reported by all 24 Trusts in the North West of England. The data comes from the Quality Measures Reporter dataset; a patient level data set which links each patient to quality measures based on health conditions.

We adopt different estimation methods to study the effect of different design structures on the distribution of performance. We first regress quarterly changes in performance scores on the quartile ranking of the Trust in each year. We also formulate a regression specification which allows us to test whether the mean performance of trusts differs under a 100% tournament regime and a 100% performance improvement regime, relative to a regime in which all trusts face a common minimum threshold. This model also allows us to test the effect of setting the negotiated threshold in the CQUIN scheme above its minimum value. We test this using a wide range of dependent variables to account for downward pressures on improvement caused by the ceiling of the performance measure.

Results

We estimate that 100% tournament system and 100% improvement system do not differ significantly from a regime based solely on a common minimum threshold. We find similar results for the 25th percentile. We estimate that mean performance for the 75th percentile under a pure improvement system is -4.752 lower compared to a common fixed threshold regime, although this disappears after controlling for ceiling effects. We find that, irrespective of scheme design, P4P is associated with greater performance improvements for poor performing trusts, but can make no conclusions on which design is optimal to maximise improvement for the lowest performers.

Conclusion

This study provides little evidence to suggest that any one incentive scheme design is optimal for generating the greatest performance improvements across all Trusts and also generate improvements for Trusts of initial poor quality.

……..

Introduction

Both in the UK and internationally, concerns over the quality of care being provided in hospitals have been well documented[1,2,3]. Pay-for-performance (P4P) schemes have emerged as the main tool used by policy makers to tackle this problem, and its popularity is growing[4]. Under its Value-Based Purchasing Program, Medicare is set to expand pay-for-performance across many US hospitals in 2013[5]. In a healthcare setting, P4P schemes involve directly relating a proportion of the remuneration of providers to achieved results on quality indicators[4]. In particular, P4P schemes seek to improve quality by incentivising providers to allocate additional effort towards specific elements of care which are either rewarded with bonus payments or subject to fines[6]. Economic theory suggests that linking payment to performance in this way will induce providers of care to change their behaviour and improve performance[7,8]. However justification of P4P's popularity in published studies is hard to find. Success at a physician level is mixed, with the success of the Quality and Outcomes Framework (QOF) in the UK[9,10,11], contrasting with the failure of similar programs in the US to consistently motivate the same improvements in processes of care[12,13,14,15]. According to a recent review[16], only 3 hospital-based pay-for-performance programs have been evaluated, and only 1 with published evidence of sufficient quality from which to draw conclusions. Evidence on the extent to which this program, the Premier Hospital Quality Incentive Demonstration (PHQID) adopted by the Centers for Medicare and Medicaid Services in 2003, led to the desired improvements in processes of care is also mixed[17,18,19]. Later studies assessing the success of other hospital-level P4P programs don't help to clarify the picture[20,21]. Also, evidence from the P4P literature in the US shows little indication that P4P schemes have motivated improvements in patient outcomes, in particular, improvements in patient mortality[19,22,23,24]. However, a newly published study in the UK[25] finds that the Advancing Quality (AQ) program, a P4P scheme introduced in all acute hospital Trusts in the North-West of England in 2008, led to significant reductions in risk-adjusted mortality over its first 18 months. This is surprising given that AQ exhibited the same 'tournament-style' design and covered the same quality of care measures as the PHQID, where no such reductions were found.

Relatively little is known as to why some P4P programs fail to meet expectations. Design choices could be a major factor in determining the effectiveness of such schemes[4]. In terms of incentive design there is particular uncertainty over two issues about how payments should be structured: as bonuses or fines; and as tournaments or benchmarks. The majority of hospital-based P4P programs have had a tournament-style structure with predominantly positive incentives permeated through bonus payments. Prospect theory

suggests that larger gains could be created using financial penalties rather than rewards as people are generally more motivated by the desire to avoid losing what they consider theirs, compared to gaining something of equal value that is perceived as extra, such as bonus payment[26].

We use the Advancing Quality program to tackle this issue directly. During its first 30 months, AQ employed three different design structures. For the first 18 months the scheme was based on a tournament-style system, with adjustments to this system occurring after 12 months aimed at rewarding improvement in performance rather than just its level. After this 18 month period, AQ became part of a regional Commissioning for Quality and Innovation (CQUIN) scheme, which employed a fines structure with locally negotiated thresholds.

Although other P4P schemes have experienced changes in their design (take the Blue Cross State of Michigan scheme for example), these changes have rarely encompassed the whole spectrum of design possibilities. Also, given differences in the "institutional context" in which different P4P schemes are applied, results from different schemes which vary by design are rarely comparable. AQ solves both of these issues by allowing us to test how the drastic and numerous differences in incentive scheme design within AQ affects performance, within the same group of hospital Trusts.

We use the variety of designs used within AQ to formulate a regression model which allows us to test whether the mean performance of trusts differs between a 100% tournament regime, identical to the incentive scheme in year 1, a hypothetical 100% regime based on performance improvement, and a regime in which there is a common fixed threshold for all trusts. In addition this model also allows us to test the effect of setting the negotiated threshold in the CQUIN scheme above its minimum value. We test this using a wide range of dependent variables, which allows us to account for downward pressures on improvement caused by the 100% ceiling of the performance measures. Findings from this will then provide a recommendation to policy makers on the optimal design of incentive schemes going forward.

Another politically charged issue in healthcare, is the gap in quality performance between trusts. As a result, of major concern for policy makers is whether P4P can help to close the gap in performance by giving greater incentives for initially poor performing trusts to improve. Previous studies assessing the distribution of performance based on initial quality at a physician level, have found evidence to suggest a negative relationship between baseline performance and performance improvement; with providers of lower initial quality improving more[12,13,9,27]. However others have found no such relationship[14,15]. At a hospital level the evidence is weaker. Although both of the papers studying this issue found that P4P motivated greater performance improvements for initially poor performing hospitals compared to initially high performing hospitals[19,18], one of these found that this difference in improvement was not significant at any strata[18].

AQ gives us the opportunity to assess which incentive scheme design leads to the greatest improvement in performance for the poorest performing Trusts, and thus which design should be employed if this is the main aim of the policy maker.

To do this, we first derive a theoretical model to help us form predictions of whether changes in incentive scheme design at 12 and 18 months lead to changes in how performance improvements were distributed across trusts based on initial performance, and thus form predictions of whether performance gaps between bad-performing and good-performing Trusts become narrower or wider. These predictions are then tested by regressing quarterly changes on the performance scores on the quartile rankings of the Trusts at the beginning of each year.

To further assess the overall impact of P4P design on distribution of performance, we estimate the regression model described above using quantile regression.

Results will help us to assess which design generates the greatest improvements in performance for different parts of the initial performance distribution, and thus which design should be implemented if the policy makers main aim to improve performance of trusts with the poorest quality.

The AQ Program

The AQ scheme was introduced for all 24 National Health Service (NHS) acute trusts in the North West of England providing emergency care, and was the first trust-level P4P scheme to be introduced in the UK[25]. Trusts are required to collect and submit data on 28 Clinical Indicator Measures (CIMs), covering five Clinical Focus Areas (CFAs); namely Acute Myocardial Infarction (AMI), heart failure, community-acquired pneumonia, coronary artery bypass grafting surgery, and hip and knee replacement surgery[25]. These measures comprise both of process of care measures and measures of patient's outcomes. Clinicians were responsible for ensuring the clinical process measures were followed and that data was collected (manually and/or through existing electronic systems) and outcomes monitored. The data is assured by the Audit Commission and results are publically reported annually on the Advancing Quality website[28].

Under AQ, performance was measured using the composite quality score (CQS), which is calculated using an opportunity model. This is the same score used in the PHQID and aggregates scores on each of the individual measures by summing the total number of processes care measures adhered to (successes) and dividing this by the total number of possible successes[29].

Bonuses were paid to trusts based on their performance on these quality scores. Hospital Chief Executive Officers agreed at the outset of the scheme that any bonuses attained must be directed towards those clinical teams which won the hospital the bonus. Members of the teams could not take these bonuses as personal income, instead having to invest these funds to further improve care[25].

In order to qualify for payments in any of the schemes (or avoid fines), Trusts were required to meet AQ data assurance requirements which entails clinical coding completeness of 95%, measure data completeness of 95%, and data accuracy checks of 80%.

Incentive scheme design

Year 1: October 2008 – September 2009

Advanced Quality began as a pure tournament system. This system ran for 12 months from October 2008 to September 2009. At the end of this first year, trusts were ranked according to their Composite Quality Scores for each of the 5 CFAs. The top quartile (top 25%) of trusts in each CFA received a 4% bonus over and above the national tariffs for the relevant condition. The second quartile also received a bonus, which was 2% of revenue; the bottom 50% received no bonus. The Ambulance Service contributed to the delivery of two of the eight AMI measures (aspirin and thrombolysis) and therefore they received 25% of any bonuses paid out for AMI. Unlike PHQID, no penalty payments were introduced to penalize the poorest performing hospitals[25]. Despite the tournament style system throughout this period, staff from all the participating hospitals met face-to-face at regular intervals to share problems and learning. This was especially the case in relation to community-acquired pneumonia and heart failure, where targeted measures presented particular challenges. Bonuses amounting to approximately £3.2million were paid to hospitals over the first 12 month period of the AQ scheme. However, these payments did not represent "new money", with the scheme being funded through contributions from participating PCTs. Data from this period was released in June 2010.

Year 2: October 2009 – March 2010

For the next 2 quarters of the scheme, from October 2009 to March 2010, the incentive reward structure changed. The reward structure in this period consisted of three components. The first involved an award for attainment: bonuses were attained for each CFA by all trusts which performed above the median composite quality score from the previous year. Meeting this median target was a pre-requisite for payment on the other two components. The second involved a reward for top performance and was identical to that in year 1: Trusts in the top 2 quartiles for each CFA received an additional incentive payment, on top of that gained for surpassing the median score. The third component was based on top improvement: Trusts were ranked based on percentage improvement in each clinical area. Trusts performing in the top quartile for quality improvement for that clinical area received an additional incentive payment. The share of payments corresponding to each CFA was set to be roughly equivalent to that in the first year: AMI: 13%, CABG: 10%, Heart Failure: 10%, Hip & Knee: 42% and Pneumonia: 25%. Within each CFA, payment was split equally between the three components of the incentive scheme. 2/3 of payments assigned to reward top performance were awarded to those in the top quartile with the

remaining 1/3 awarded to those in the 2nd quartile. Bonuses amounting to approximately £1.6million were paid to hospitals over this period.

Year 3/CQUIN: April 2010 – March 2011

After this 6 month period, AQ became part of a regional Commissioning for Quality and Innovation (CQUIN) scheme. The CQUIN payment system is based on a fines structure with locally negotiated thresholds based on the CQS. The initial thresholds were set centrally by AQ, and were based on median yearly scores from the year 1 scheme. These thresholds differed across CFAs and also across trusts. If trusts failed to meet these thresholds then they lost a percentage of their contract value. In 2010/11, total CQUIN payments amounted to 1.5% of providers total contract value. To receive the full 1.5%, trusts had to meet targets on national goals (set by the Department of Health), regional goals (set by Strategic Health Authorities) and local goals (set by Primary Care Trusts (PCTs)). 0.3% of contract value related to achieving the national goals with the remaining 1.2% for regional and local goals. Targets relating to each CFA included in AQ were included in the regional goals and each accounted for 0.01% of the total contract value. "Local" CQUINs were also in place which allowed Primary Care Trusts (PCTs) to increase (but never reduce) both the thresholds required to receive payment and the % contract value stipulated by the regional AQ thresholds. This allowed PCTs to increase rewards for performance on conditions which were of more importance in their local area. Even after these adjustment, a maximum threshold of 95% was set, as stipulated by Premier.

A model for the distribution of performance improvement

Following a currently unpublished paper[30], we assume that there is a set N of i risk-neutral Trusts and that each Trust produces a single non-verifiable, Z-dimensional credence good called patient treatment quality in a set M of m Clinical Focus Areas (CFA). Each area has Z_m quality measures that sum up to Z. Trusts' quality performance in each CFA is measured using the Composite Quality Score (CQS). The score aggregates performance on each individual measure z over its target patient population into one percentage value. We denote Trust i's chosen treatment vector by t_i that consists of M area-specific subvectors $t^{i,m} = \left(t_1^{i,m}, \dots, t_{Z_m}^{i,m}\right)$ and $t_z^{i,m} \in [0,100], i \in N, m \in M$ and $z \in Z$ where Z is finite. We refer to an element of M as a clinical focus area and an element z_m as a measure in the set of Z_m measures for the m^{th} CFA.

The quality of care is an unobservable function $q_i(t_i, \gamma_i)$, where γ_i is a Trust's innate cost type of producing quality defined over the treatment space $[0,100]^{M \times Z_m}$. Here we replace the interpretation of innate cost type with demonstrated performance at the beginning of each year of the AQ programme.

Although the bonus schemes in the three years of AQ differed, they all implemented payoffs to Trust i as

$$u_i(t_i, \gamma_i) = q_i^l(t_i, \gamma_i)P^l - c_i(t_i, \gamma_i)$$

where, for year 1 of the scheme, P^l is the prize of being in the l^{th} quartile, for the year 2 scheme P^l is the combined prize from its 3 components and, for year 3, P_l is the prize for meeting the CQUIN threshold. Costs $c_i(t_i, \gamma_i)$ are weakly convex in each dimension. This cost type is private information to the Trust concerned. For simplicity, we assume the simple asymmetric marginal cost function

$$MC_i(t_z^{i,m}, \gamma_z^{i,m}) = \frac{1}{(100 - t_z^{i,m})^{\gamma_z^{i,m}}}$$

Note that cost asymmetry is the only source of asymmetry in the model. We assume that $\gamma_i \in (0, \infty)$. Total cost function for $t_z^{i,m}$ is defined as $(t_i, \gamma_i) = c_{i0} + \frac{1}{(\gamma_i - 1)} \frac{1}{(100 - t_i)^{\gamma_i - 1}}$. For $\gamma_i = 1$, total cost function is $c_i(t_i, \gamma_i) = c_{i0} - \ln(100 - t_i)$. For the case of $0 < \gamma_i < 1$, $c_i(100, \gamma_i) < \infty$, while $c_i(100, \gamma_i) = \infty$ for $\gamma_i \geq 1$. The total cost function for area m and trust i is

$$C\left(t_i^m, c_0^{i,m}, \gamma^{i,m}\right) = \sum_{z=1}^{Z_m} c_{0,z}^{i,m} + \sum_{z=1}^{Z_m} \frac{1}{(\gamma_z^{i,m} - 1)(100 - t_z^{i,m})^{\gamma_z^{i,m} - 1}}$$

Where $c_0^{i,m}$ and $\gamma^{i,m}$ are Z_m – dimensional vectors of corresponding parameters. The aggregate performance for the m^{th} area is

$$T^{i,m} = \sum_{z=1}^{Z_m} t_z^{i,m}(\gamma_i, c_{0,i})$$

Where $t_z^{i,m}(\gamma_i, c_{0,i})$ is the solution of the maximization problem for the area m for trust i and quality indicator z. Define the probability of being in the first quartile of a distribution as $p_{1:4}(T^{i,m})$ and the probability of being in the second quartile of a distribution as $p_{2:4}(T^{i,m})$. Then expected payoff for area m for the first period would be:

$$p_{1:4}(T^{i,m})P^1 + p_{2:4}(T^{i,m})P^2 - c(t_i, \gamma_i, c_{0,i})$$

The model implies that Trusts will only increase quality if their expected bonus exceeds the cost of quality improvements.

Payoff structure changes in the second year. It is conditional on Trust attaining the median score for year 1 for $T^{i,m}$, $T_{t=1,50}^{i,m}$. For individual trust i expected payoff is

$$I\left(T_{t=2}^{i,m} \geq T_{t=1,50}^{i,m}\right)P_{i,1} + p_{1:4}\left(T_{t=2}^{i,m} | T_{t=2}^{i,m} \geq T_{t=1,50}^{i,m}\right)P_{i,2}^1 + p_{2:4}\left(T_{t=2}^{i,m} | T_{t=2}^{i,m} \geq T_{t=1,50}^{i,m}\right)P_{i,2}^2$$
$$+ p_{1:4}\left(\Delta\% T_{t=2}^{i,m} | T_{t=2}^{i,m} \geq T_{t=1,50}^{i,m}\right)P_{i,3} - c(t_{t=2,i}, \gamma_i, c_{0,i})$$

Where $I\left(T_{t=2}^{i,m} \geq T_{t=1,50}^{i,m}\right)$ is an indicator function, $p_{s:4}(.|.)$ is s^{th}conditional quartile of the corresponding distribution. In the second year the amount of money reserved for total hospitals' payments $P_{.,1}$is the same as $P_{.,2}^{1} + P_{.,2}^{2}$ and the same as $P_{.,3}$.

For the third year of AQ (CQUIN), there are no probabilities involved. A trust will get a reward P^m for area m if $T^{i,m} > \bar{T}^{i,m}$, where $\bar{T}^{i,m}$ is an agreed threshold. Thus, a trust will participate as long as

$$c\left(T^{i,m}, \gamma_i, c_{0,i}\right) \leq P^m$$

Predictions

First consider the pure tournament system in year 1. Predictions regarding improvements in performance measured by mean performance scrores are ambiguous. On one hand, given initial quality is observed, trusts can infer their probability of winning a bonus and thus their expected bonus. All else equal, Trusts that are initially good-performing have a greater probability of winning a bonus and, therefore, are expected to improve more. On the other hand, our theoretical model implies that good-performing Trusts have higher marginal cost which makes each additional increase in performance more costly; reducing the expected payoff. As a result, the gap between good- and poor-performing Trusts could become smaller or larger.

Under the mixed incentive scheme in year 2, the "attainment" component of the bonus creates a new incentive for poor-performing trusts to improve quality and for good-performing trusts to maintain their performance. The "top improvement" component introduces a further incentive for poor-performing Trusts to improve performance. Moreover, these Trusts have enough room to increase quality as opposed to good-performing Trusts which scores are already close to the maximum. The third, "top performance", component generates incentives similar in nature to the incentives in year 1. Given that for poor-performing Trusts the marginal cost for an additional improvement is lower, our model predicts that performance improvement, on average, will be larger for these Trusts. The gap in performance is hence expected to decrease.

In year 3 the attainment nature of the reward schedule provides all Trusts with an incentive to achieve a minimum improvement in performance and the step change for low-performing trusts is larger. As the reward is conditional only on attaining the threshold and further gains are costly especially for Trusts at the high end of the performance distribution, the model predicts that change in performance will be, on average, larger for low-performing Trusts and the gap between the good- and poor-performing Trusts will slightly decrease.

Data

This study uses data from the Quality Measures Reporter dataset; a patient level data set which links each patient to quality measures based on health conditions. The data consists of trust-level measures for each indicator on the incentivized conditions. We decided against including hip & knee replacement surgery in the analysis as we did not have outcome data covering the last 2 quarters of the first year of CQUIN. CABG was also not considered as only 4 out of the 24 Trusts in North West of England performed this procedure. We omitted stroke from the analysis as it became an incentivized condition only under the CQUIN and, therefore, the impact of different payment structures could not be analysed. Time inconsistencies in the receipt of data precluded us from considering AMI, but will be incorporated in future analysis.

Data for the 2 studied conditions, pneumonia and heart failure, was available at a quarterly level for the first 10 quarters of the AQ program spanning the period from October 1st, 2008 to March 31st, 2011. There are 24 Trusts that collected information on the heart failure condition and 23 on pneumonia as the Liverpool Heart and Chest Trust do not treat pneumonia patients. In the analysis we use pooled quarterly data on pneumonia and heart failure.

Methods

We first test how performance differed under different incentive schemes. We define the following regression specification:

$$y_{izmt} = \beta_0 + \beta_1\, quarter_t + \beta_2 D1 + \beta_3 D2_{imt} + \beta_4 D3 + X_{it}\gamma + Z_i\theta + c_{iz} + \varepsilon_{izmt} \qquad (1)$$

Where y_{izmt} is the value of the performance measure for trust *i*, on quality indicator *z*, for condition *m*, in quarter *t*. $D1$ is a proxy for pure tournament scheme. It takes a value of 1 in year 1, 1/3 in year 2, and 0 year 3. *D1* values reflect the contribution of the pure tournament scheme to the incentive design in the respective year: a pure tournament in year 1, a mixed scheme in year 2 where the tournament forms a third of the scheme and a full fixed threshold system in year 3. $D3$ is a proxy for improvement with value of a third in year 2 and 0 in years 1 and 3 to capture the contribution of the improvement component to the design schemes in the different years. We define

$$D2_{imt} = CQUINthreshold_{imt} - minCQUINthreshold_m$$

which measures the difference between the CQUIN threshold set for trust *i*, for condition *m* in year 3, relative to the minimum CQUIN threshold over all trusts for that condition. $D2_{imt}$ is set to 0 for years 1 and 2. Hence, our "base design" is a regime where each Trust faces a common threshold, which capture the minimum attainment award in year 2. X_{it} is a set of time-varying covariates which consists of the proportion of patients treated within 18 weeks, and the number of beds available for day-only procedures. Z_i is a set of time-invariant covariates which consists of a measure of the size and type of trust, a measure

of the quality of financial management, foundation trust status, and measures of commitment to core standards and national priorities published by the Care Quality Commission. c_{iz} is the unobserved trust-indicator effect for trust *i* and indicator *z*. ε_{izmt} are idiosyncratic errors.

β_2 measures whether the average outcome under a pure tournament-style incentive scheme is greater relative to a scheme in which trusts face a common minimum threshold. β_3 measures whether outcomes differ between an incentive regime consisting solely of a payment for performance improvement, as extrapolated from the small role it played in the mixed scheme in year 2, and that of the same common threshold scheme. β_4 measures the effect of unit increase the CQUIN threshold above its minimum value. We also test whether there is any difference in average outcomes under the pure-tournament and pure-improvement incentive schemes.

We define y_{izmt} in 4 ways. Firstly, the achievement score on each indicator which allows us to measure how the variables described above effect mean performance, and is defined as follows:

$$P_{izmt} = \frac{T_{izmt}}{E_{izmt}} * 100$$

where T_{izmt} is total amount of patients treated on indicator *z* in quarter *t*, and E_{izmt} is the number of patients eligible for treatment on indicator *z* in quarter *t*.

Secondly it is defined as the quarterly improvement in the achievement which allows us to test how the design structure affect the rate of performance improvement. It is defined as follows:

$$dP_{izmt} = P_{izm,t} - P_{izm,t-1}$$

However, performance on these measures is likely to be affected greatly by the 100% ceiling on the score, with its effect larger for the later schemes. Thus we also define y_{izmt} as the both the level and change in the log odds ratio:

$$\log odds_{izmt} = ln\left(\frac{P_{izmt}}{100 - P_{izmt}}\right)$$

which takes account of the ceiling at 100.

The choice of either pooled OLS, random effects (RE) or fixed effects (FE) estimation was carried out using Hausman tests and the Wald test produced as a by-product of RE estimation.

To test the predictions on how the incentive design impacted the distribution of performance, we estimated:

$$dP_{izmt} = \beta_1(y1 * yrstartq1)_{mt} + \beta_2(y1 * yrstartq2)_{mt} + \beta_3(y1 * yrstartq3)_{mt} + \beta_4(y1 * yrstartq4)_{mt} + \beta_5(y2 * yrstartq1)_{mt} + \beta_6(y2 * yrstartq2)_{mt} + \beta_7(y2 * yrstartq3)_{mt} + \beta_8(y2 * yrstartq4)_{mt} + \beta_9(y3 * yrstartq1)_{mt} + \beta_{10}(y3 * yrstartq2)_{mt} + \beta_{11}(y3 * yrstartq3)_{mt} + +\beta_{12}(y3 * yrstartq4)_{mt} + X_{it}\gamma + Z_i\theta + u_{izmt} \ldots (2)$$

$yrstartq *$, $(*)$=1,2,3,4, is a dummy variable which equals 1 if a trust belongs to quartile $(*)$ at the beginning of each AQ period where quarter 1 is the baseline period for the first year of AQ, quarter 4 is the baseline for the second AQ period, and quarter 6 is the baseline for the third AQ period. Quartiles are created for each condition in each period using ranks of Trusts based on the level of CQS in the baseline quarter.

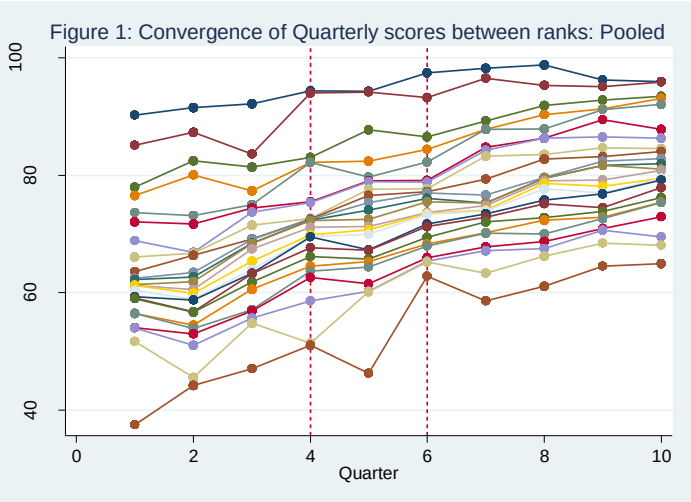
The coefficient estimate on each dummy captures the average quarterly change in CQS for trusts in a given quartile in a given period. Post-estimation tests on these coefficients are used to test predictions. The tests are run on both the unadjusted mean quarterly changes, and on these changes after being adjusted for covariate effects. To test whether each incentive scheme lead to significantly different improvements for poor performing trusts relative to good performing trusts we test the difference in mean quarterly improvement between trusts in quartile 4 relative to quartile 1. To test which year generated the greatest performance improvements for top-performing trusts we test the equality of β_1 and β_5, β_1 and β_9, and β_5 and β_9, and for initially poor performing trusts, the equality of β_4 and β_8, β_4 and β_{12}, and β_8 and β_{12}.

In addition to this regression we also estimate the model (1) using quantile regression. Given data is at a trust-indicator-quarter level, this does not measure how the effects of incentive scheme design differs based on the initial performance of trusts. It does however measure the impact on the distribution of performance in general. We assess the impact of incentive scheme design on poor performance using a regression at the 25th percentile, and on the impact on good performance with a regression at the 75th percentile. FE estimates will be presented here.

Robust standard errors were employed in the presence of heteroskedasticity

Results

Figure 1 illustrates how the distribution of achievement changed throughout the studied period. Each line corresponds to the rank of a trust, 1-24, and represents the trend in mean score over all indicators for that rank. The whole distribution follows and upward trend indicating AQ was associated with a general improvement in adherence to recommended processes of care. The mean score for the top ranked trust increases but only marginally, with the trend for the bottom ranked trust increasing at a much faster rate. This indicates that AQ is correlated with a narrowing of the performance gap between good and poor performing trusts. The figure suggests that the rate at which this gap narrowed slowed considerably when AQ became part of the wider CQUIN scheme.



Estimates for the mean regressions and the quantile regressions at the 25th and 75th percentiles, for all dependent variables are reported in Table 1.

Heteroskedasticity was present when estimating model (1) using all dependent variables and when estimating the conditional mean and conditional quantiles. As a result, robust standard errors where used in all specifications. Hausman tests rejected equality for FE and RE estimates indicating correlation between unobserved trust-indicator effects and the regressors. This was consistent across all dependent variables for the mean regressions. Thus, FE estimation was employed across the board. As no simple code exists for random effects estimation in a quantile regression setting, these models were only estimated by pooled OLS and fixed effects. Given substantial evidence of heterogeneity bias in the mean regressions, only fixed effect estimates are reported.

Table 1: Mean regression and Quantile regression FE estimates for the 4 dependent variables

	Score					
	Mean		**25th percentile**		**75th percentile**	
Quarter	2.180***	(-6.92)	1.566***	(-3.49)	0.920***	(-3.86)
D1	-2.264	(-1.02)	-2.955	(-1.21)	-1.301	(-0.89)
D2	-0.473***	(-3.97)	-0.276***	(-4.50)	-0.286***	(-5.05)
D3	-2.04	(-0.75)	2.092	(-0.85)	-4.752*	(-2.50)
Totaldaybeds	0.0456	(-1.42)	0.0303	(-1.35)	0.00259	(-0.12)
P18weeks	0.280*	(-2.16)	0.308*	(-2.36)	0.0942	(-1.02)
N	2104		2104		2104	

	logodds					
	Mean		**25th percentile**		**75th percentile**	
Quarter	0.132***	(7.86)	0.164***	(6.95)	0.100***	(5.17)
D1	-0.237*	(-2.10)	-0.0334	(-0.21)	-0.129	(-1.11)
D2	-0.0300***	(-4.41)	-0.0187**	(-2.98)	-0.0186***	(-3.82)
D3	-0.177	(-1.13)	0.138	(0.64)	-0.124	(-0.73)

Totaldaybeds	0.00159	(0.89)	0.00137	(0.83)	0.000883	(0.67)
P18weeks	0.0173*	(2.27)	0.0234**	(2.80)	0.00719	(0.90)
N	2096		2096		2096	

	dscore					
	Mean		**25th percentile**		**75th percentile**	
D1	0.935	(1.04)	-0.372	(-0.49)	1.767*	(2.55)
D2	0.0912	(1.19)	0.0288	(0.46)	0.0521	(1.15)
D3	2.499	(0.98)	0.215	(0.14)	2.358	(1.31)
Totaldaybeds	0.0176	(0.90)	0.0128	(0.72)	0.0306	(1.59)
P18weeks	0.233*	(2.11)	0.126	(1.43)	0.145	(1.63)
N	1893		1893		1893	

	dlogodds					
	Mean		**25th percentile**		**75th percentile**	
D1	0.0529	(1.18)	-0.00236	(-0.04)	0.0619	(0.74)
D2	0.00821	(1.96)	0.00584	(0.76)	0.00457	(0.64)
D3	0.285*	(1.98)	0.341	(1.49)	0.282	(1.34)
Totaldaybeds	0.000996	(0.83)	0.000316	(0.20)	0.00166	(0.88)
P18weeks	0.00813	(1.28)	0.00676	(0.75)	0.0137	(1.41)
N	1879		1879		1879	

t statistics in parentheses

* p<0.05, ** p<0.01, *** p<0.001

First consider the mean regression estimates when indicator scores are used as the dependent variable. The coefficient of the quarter dummy is positive and significant (p<0.001). It indicates that, ceteris paribus, mean indicator scores across all trusts increase, on average, by 2.18 percentage points per quarter. This is consistent with economic theory which suggests P4P should motivate general improvements in performance. The coefficients of the D1 and D3 dummies are negative but insignificant, indicating that mean performance under a 100% tournament system or a 100% improvement system doesn't differ significantly from that of a regime consisting solely of a common minimum threshold. The coefficient of the D2 dummy is negative and significant (p<0.001) indicating that, all else equal, a 1 unit increase in the CQUIN threshold relative to its minimum value, is correlated with a 0.473 percentage point reduction in mean scores, on average. This is surprising given that we would expect a higher threshold to generate higher performance given trusts need to perform better in order to avoid fines. However this result may indicate that thresholds were too challenging to meet, deterring Trusts from attempting to achieve them. This argument is supported by the fact that a third of Trusts (8 out of 24) for heart failure, and over a fifth of trusts (5 out 18) for pneumonia failed to meet their target thresholds. Post estimation test reveal no difference in performance between the pure tournament and pure improvement systems. Results for the 25th percentile are nearly identical to that of the mean regressions, with similar magnitude and identical significance of estimates effects. This indicates the effect of incentive design on poor performance is

similar to that on mean levels of performance. However, the D3 dummy is negative and significant when estimating at the 75th percentile. This implies mean performance for the top 25% of the performance distribution (or good performance) would be on average 4.752 percentage points lower under a pure improvement regime, in comparison to a common fixed threshold regime. This is intuitive given that good performance is subject to a large ceiling effect which limits the degree to which it can improve. This result is driven by the fact that good performance is concentrated in trust which, on average, perform better. Given that improvement for these trusts is limited by ceiling effects, and given they have greater costs of improvement relative to other trusts, attainment of the improvement award very unlikely, and thus does little to motivate improvements.

When taking ceiling effects into account, through the use of the log odds ratio as the dependent variable, signs and significance remains similar. However, the negative effect of the improvement award for top-performing trust disappears, which is intuitive given it is likely that this effect was largely driven by ceiling effects. The negative coefficient of the D1 dummy also becomes significant for the mean regression indicating that on average across all trusts, performance under a pure tournament regime is lower than that with common threshold regime.

Across both the score and log odds specifications, a higher proportion of patient treated within 18 weeks significantly increases mean performance across all trusts. For the score specification, where interpretation is easier, a 1 percentage point increase in the percentage of patients treated within 18 weeks, is associated with a 0.28 percentage point increase in performance. This is intuitive given lower waiting times could be an indicator of general Trust quality. The magnitude of this effect is greater for poor performing trusts, indicated by greater t-statistics for the waiting time coefficient in the regression on the 25th percentile. However this effect is insignificant for trusts in the top quartile of the performance distribution. It is unknown why this effect differs so distinctly based on initial performance of the trust.

For the dscore specification estimated both at the mean and at the percentiles, the majority of the effects are insignificant, indicating that incentive scheme design and the degree to which CQUIN threshold exceed its minimum value, has little impact of performance improvement, irrespective of initial performance. However, a regime consisting of a 100% tournament system, is correlated with a significant increase in performance improvement of 1.77 percentage points per quarter, relative to a scheme in which there is only a common fixed threshold to meet, for trusts in the top 25% of the performance improvement distribution. However, when ceiling effects are controlled for using the difference in the log odds ratio this effect disappears. Also, after taking this account, the effect of a pure improvement scheme becomes significant for the mean regressions, indicating it generates greater performance improvement relative to a common threshold scheme.

Table 2 reports the average quarterly change in the CQS in years 1, 2 and 3 for trusts in quartile 1, 2, 3 and 4. Table 3 reports the F-statistics for tests of equality between mean changes from the post-estimation tests when no covariates are accounted for. Table 5 reports these statistics after controlling for the effects of

covariates. A test for heteroskedasticity rejects homoscedasticity and thus robust standard errors are used in estimation of model (2).

Table 2: Average quarterly change in CQS for each quartile in each year

Year	Quartile			
	Quartile 1	**Quartile 2**	**Quartile 3**	**Quartile 4**
Year 1	1.206	0 .426	2.32	5.10
Year 2	-1.112	1.251	1.81	7.69
Year 3	0 .748	1.518	0.90	3.191

All values are to 3.d.p

Table 3: F tests of the equivalence between the average quarterly mean changes in achievement for each quartile: within years Non-adjusted estimates

	Year 1				**Year 2**			
	Quartile 1	**Quartile 2**	**Quartile 3**	**Quartile 4**	**Quartile 1**	**Quartile 2**	**Quartile 3**	**Quartile 4**
Quartile 1	**.**	0.15	0.37	5.74*	**.**	1.13	1.35	10.70**
Quartile 2	**.**	**.**	0.70	4.96*	**.**	**.**	0.06	6.55*
Quartile 3	**.**	**.**	**.**	2.12	**.**	**.**	**.**	4.48*
Quartile 4	**.**	**.**	**.**	**.**	**.**	**.**	**.**	**.**

	Year 3			
	Quartile 1	**Quartile 2**	**Quartile 3**	**Quartile 4**
Quartile 1	**.**	0.71	0.01	4.58*
Quartile 2	**.**	**.**	0.17	1.80
Quartile 3	**.**	**.**	**.**	1.98
Quartile 4	**.**	**.**	**.**	**.**

*p<0.05, **p<0.01, *** p<0.001

Table 5: F tests of the equivalence between the average quarterly mean changes in achievement for each quartile across years: unadjusted estimates

	Quartile 1			**Quartile 2**		
	Year 1	**Year 2**	**Year 3**	**Year 1**	**Year 2**	**Year 3**
Year 1	**.**	1.31	0.14	**.**	0.14	0.34
Year 2	**.**	**.**	1.08	**.**	**.**	0.03
Year 3	**.**	**.**	**.**	**.**	**.**	**.**

	Quartile 3			**Quartile 4**		
	Year 1	**Year 2**	**Year 3**	**Year 1**	**Year 2**	**Year 3**
Year 1	**.**	0.05	0.53	**.**	1.16	1.46
Year 2	**.**	**.**	0.16	**.**	**.**	3.80
Year 3	**.**	**.**	**.**	**.**	**.**	**.**

*p<0.05, **p<0.01, *** p<0.001

Table 4: F tests of the equivalence between the average quarterly mean changes in achievement for each quartile within years: Adjusted estimates

	Year 1				Year 2			
	Quartile 1	**Quartile 2**	**Quartile 3**	**Quartile 4**	**Quartile 1**	**Quartile 2**	**Quartile 3**	**Quartile 4**
Quartile 1	.	0.11	0.76	5.18*	.	1.31	1.74	11.63***
Quartile 2	.	.	1.07	5.00*	.	.	0.13	7.07**
Quartile 3	.	.	.	1.47	.	.	.	4.28*
Quartile 4	.	.	.	.	.	.	.	.

	Year 3			
	Quartile 1	**Quartile 2**	**Quartile 3**	**Quartile 4**
Quartile 1	.	2.46	0.26	5.00*
Quartile 2	.	.	0.48	0.74
Quartile 3	.	.	.	1.83
Quartile 4	.	.	.	.

*p<0.05, **p<0.01, *** p<0.001

Table 6: F tests of the equivalence between the average quarterly mean changes in achievement for each quartile across years

	Quartile 1			Quartile 2		
	Year 1	**Year 2**	**Year 3**	**Year 1**	**Year 2**	**Year 3**
Year 1	.	1.43	0.25	.	0.20	1.06
Year 2	.	.	0.91	.	.	0.36
Year 3	.	.	.	.	.	.

	Quartile 3			Quartile 4		
	Year 1	**Year 2**	**Year 3**	**Year 1**	**Year 2**	**Year 3**
Year 1	.	0.04	0.53	.	1.45	0.84
Year 2	.	.	0.17	.	.	3.84
Year 3	.	.	.	.	.	.

*p<0.05, **p<0.01, *** p<0.001

The raw averages for pure tournament system in year 1 indicate that performance improvement was negatively related with baseline performance in this year, with the poorest performing trusts (quartile 4) improving the most (5.10 per quarter) and the top performing trusts (quartile 1) improving the least (1.206 per quartere). The difference in these improvements are significant at a 5% level (F=5.74), and remains so after controlling for covariates (F=5.18). This indicates that the greater incentives for initially top performing trusts to improve as a result of the higher expected bonus is outweighed by the relatively higher

marginal costs of improvement. These differential rates of improvement have caused the performance gap between poor and top trusts to narrow in year 1.

Results for the mixed incentive scheme in year 2 are consistent with our predictions. In this year, the poorest performing trusts improved, on average, by 7.6 percentage points. The relatively greater incentives to improve as a result of the improvement award, along with lower marginal cost of improvements mean that this is considerably higher than the improvements for initially top performing trusts which who saw average falls in performance of 1.1 percentage points per quarter. When covariates are not controlled for, the difference in improvement is significant at a 1% level (F=10.70), with this significance increasing to a 0.1% level after controlling for covariates. As predicted, this indicates that there was a further narrowing of the performance gap in year 2.

Also as predicted, the year 3 scheme leads to a further narrowing of the performance gap, with the average improvement for initially poor performing trusts (3.19 percentage points per quarter) being significantly higher (5%) than that of initially top performing trusts (0.748 percentage points per quarter).

These results indicate that irrespective of the design of the incentive scheme, pay-for-performance motivates greatest improvements in performance for trusts who are initially poor performing.

Tables 4 and 6 report the F-statistics from the testing equality of mean performance improvement for each quartile over time. For each quartile, both when covariates are controlled for and when there not, there is no significant difference in performance improvement across the incentive schemes. Thus, based solely on these results, we cannot make a recommendation as to which design to employ if policymaker's main aim is to increase the performance of the initially poor performing.

Discussion

Pay-for-performance as a tool to improve the quality of healthcare provision is growing in popularity despite scant evidence of its success. Little is known as to why some schemes fail to motivate the desired improvements. The paper aimed to shed some light as to what design structure is optimal to generate the greatest level of performance, and motivate improvements for initially bad performing trusts. We also aimed to assess whether design structures employed within AQ favoured initially good or initially poor performing trusts. Although we find that AQ is associated with a substantial increase in performance, no one incentive scheme design generated either significantly greater levels or improvements in performance relative to other designs. Surprisingly we find little evidence from the quantile regressions that incentive scheme designs affected poor performance differently compared to the mean level of performance. Evidence suggests that a common threshold may be optimal to generate higher performance for top performing trusts, relative to a pure improvement scheme, although this association became insignificant after ceiling effects were accounted for. Also, results indicate that a pure improvement scheme is preferable to a common threshold scheme to generate performance improvements. Results suggest that irrespective on incentive scheme design, P4P generates greater improvements for initially poor performing

trusts relative to initially good performing trusts. However, tests of performance improvement over time for different quartiles of the performance distribution did not lead to conclusive conclusions about which design to employ to increase the performance of trust which are initially poor. We also find that setting of the CQUIN threshold above it minimum values is associated with lower mean performance.

However, there are aspects of the AQ scheme, which cloud the effect of incentive scheme designs. Unlike P4P schemes in the US, public reporting of process of care measures didn't begin before financial incentives were introduced. Given that public reporting and P4P began simultaneously within AQ, we cannot disentangle the effects of incentive design from incentives created by public reporting. Public reporting gives greater incentives for initially poor performers to improve, given trusts don't want the public "embarrassment" of being seen as poor quality, although there is doubts over whether patients in the UK use this information inform the choice of location of treatment. It may be this which is driving the greater performance improvement for poor performing trusts found in this study.

Also given the lack of availability of a control group when analysing process measures for both AQ and CQUIN, results merely provide an indication of the effects, rather than information on causality.

Concluding remarks

The aim of this study was to assess the impact of design structure on mean performance and the distribution of performance. Results provide no indication about and an optimal structure to generate either greater performance or greater improvements in performance, although results suggest P4P may generate greater improvements for initially lower performers irrespective of design. However, these conclusions must be taken with caution, given the effects of public reporting are uncertain.

---

[1] Institute of Medicine (2001). Crossing the quality chasm: a new health system for the 21st century. Washington, DC: National Academies Press.

[2] Jha, A. K., Li, Z., Orav, E. J., & Epstein, A. M. (2005). Care in US hospitals—the Hospital Quality Alliance program. New England Journal of Medicine, 353(3), 265-274.

[3] Darzi, A. (2008). High quality care for all: NHS next stage review final report. London: Department of Health.

[4] Van Herck, P., De Smedt, D., Annemans, L., Remmen, R., Rosenthal, M. B., & Sermeus, W. (2010). Systematic review: effects, design choices, and context of pay-for-performance in health care. BMC Health Services Research, 10(1), 247.

[5] Centers for Medicare and Medicaid Services. (2011). Medicare program: inpatient value-based purchasing program. Final rule. Fed Regist; 76:26490-547.

[6] Nicholas, L. H., Dimick, J. B., & Iwashyna, T. J. (2011). Do Hospitals Alter Patient Care Effort Allocations under Pay-for-Performance?. Health services research, 46(1p1), 61-81.

[7] Prendergast, C. (1999). The provision of incentives in firms. Journal of economic literature, 37(1), 7-63.

[8] Asch, B., and Warner, J. (1996) "Incentive Systems: Theory and Evidence," in D. Lewin, D. Mitchell, and M. Zaidi (eds), Handbook of Human Resource Management, ed. Stamford, Connecticut: JAI Press, 175-215

[9] Doran, T., Fullwood, C., Gravelle, H., Reeves, D., Kontopantelis, E., Hiroeh, U., & Roland, M. (2006). Pay-for-performance programs in family practices in the United Kingdom. New England Journal of Medicine, 355(4), 375-384.

[10] Campbell, S., Reeves, D., Kontopantelis, E., Middleton, E., Sibbald, B., & Roland, M. (2007). Quality of primary care in England with the introduction of pay for performance. New England Journal of Medicine, 357(2), 181-190.

[11] Gravelle, H., Sutton, M., & Ma, A. (2007). Doctor behaviour under a pay for performance contract: evidence from the Quality and Outcomes Framework. Centre for Health Economics, University of York Working Papers.

[12] Rosenthal, M. B., Frank, R. G., Li, Z., & Epstein, A. M. (2005). Early experience with pay-for-performance. JAMA: the journal of the American Medical Association, 294(14), 1788-1793.

[13] Coleman, K., Reiter, K. L., & Fulwiler, D. (2007). The impact of pay-for-performance on diabetes care in a large network of community health centers. Journal of health care for the poor and underserved, 18(5), 966-983.
[14] Mullen, K. J., Frank, R. G., & Rosenthal, M. B. (2010). Can you get what you pay for? Pay-for-performance and the quality of healthcare providers. The Rand journal of economics, 41(1), 64-91.
[15] Damberg, C. L., Raube, K., Teleki, S. S., & dela Cruz, E. (2009). Taking stock of pay-for-performance: a candid assessment from the front lines. Health Affairs, 28(2), 517-525.
[16] Mehrotra, A., Damberg, C. L., Sorbero, M. E., & Teleki, S. S. (2009). Pay for performance in the hospital setting: what is the state of the evidence?.American Journal of Medical Quality, 24(1), 19-28.
[17] Grossbart, S. R. (2006). What's the return? Assessing the effect of "pay-for-performance" initiatives on the quality of care delivery. Medical Care Research and Review, 63(1 suppl), 29S-48S.
[18] Lindenauer, P. K., Remus, D., Roman, S., Rothberg, M. B., Benjamin, E. M., Ma, A., & Bratzler, D. W. (2007). Public reporting and pay for performance in hospital quality improvement. New England Journal of Medicine, 356(5), 486-496.
[19] Glickman, S. W., Ou, F. S., DeLong, E. R., Roe, M. T., Lytle, B. L., Mulgund, J., & Peterson, E. D. (2007). Pay for performance, quality of care, and outcomes in acute myocardial infarction. JAMA: the journal of the American Medical Association, 297(21), 2373-2380.
[20] Ryan, A. M., & Blustein, J. (2011). The Effect of the MassHealth Hospital Pay-for-Performance Program on Quality. Health services research, 46(3), 712-728.
[21] Werner, R. M., Kolstad, J. T., Stuart, E. A., & Polsky, D. (2011). The effect of pay-for-performance in hospitals: lessons for quality improvement. Health Affairs, 30(4), 690-698.
[22] Ryan, A. M. (2009). Effects of the Premier hospital quality incentive demonstration on Medicare patient mortality and cost. Health services research, 44(3), 821-842.
[23] Jha, A. K., Joynt, K. E., Orav, E. J., & Epstein, A. M. (2012). The long-term effect of premier pay for performance on patient outcomes. New England Journal of Medicine, 366(17), 1606-1615.
[24] Flodgren, G., Eccles, M., Shepperd, S., Scott, A., Parmelli, E., & Beyer, F. (2011). An overview of reviews evaluating the effectiveness of financial incentives in changing healthcare professional behaviours and patient outcomes. Cochrane Database Syst Rev, 7.
[25] Sutton, M., Nikolova, S., Boaden, R., Lester, H., McDonald, R., & Roland, M. (2012). Reduced Mortality with Hospital Pay for Performance in England. New England Journal of Medicine, 367(19), 1821-1828.
[26] Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica: Journal of the Econometric Society, 263-291.
[27] Vaghela, P., Ashworth, M., Schofield, P., & Gulliford, M. C. (2009). Population intermediate outcomes of diabetes under pay-for-performance incentives in England from 2004 to 2008. Diabetes Care, 32(3), 427-429.
[28] http://www.advancingqualitynw.nhs.uk
[29] Premier Inc. (2004) CMS HQI Demonstration Project: Composite Quality Score Methodology Overview. Accessed 19/08/2012. Available at: https://www.premierinc.com/quality-safety/tools-services/p4p/hqi/resources/composite-scoring-overview.pdf
[30] Nikolova, S., Schweinzer, P., and Sinko, A. (2010), Quality Incentives and Competition Among English Health Care Trusts, University of Manchester Working Paper.