HEDG Health, Econometrics and Data Group
THE UNIVERSITY of York

HEDG Working Paper 07/20

Evaluation of the introduction of a pay for performance contract for UK family doctors using participant perceptions

Matt Sutton
Divine Ikenwilo
Diane Skatun

August 2007

EVALUATION OF THE INTRODUCTION OF A PAY FOR PERFORMANCE CONTRACT FOR UK FAMILY DOCTORS USING PARTICIPANT PERCEPTIONS

Matt Sutton[*]
Divine Ikenwilo
Diane Skåtun

Health Economics Research Unit
University of Aberdeen

Abstract

In 2004 the UK government introduced a new 'pay for performance' element into the contract for family doctors (FDs). Its universal introduction with no pre-intervention data is not atypical of system-wide health reform but poses a considerable evaluation challenge. We derive estimates of its impact based on qualitative perceptions of the treatment effect reported by a sample of participants. We exploit variation in the first-year achievements of those participants who thought quality had remained the same to generate pre-intervention estimates for those that perceived a change in quality. The average partnership of 4 FDs was paid £74,000 for achieving 982 of the 1,050 quality points available in the first year. Of these, we estimate the mean net gains attributable to the new contract to be less than 4 quality points. These gains were predominantly made on the clinical criteria and were larger for partnerships facing more competition for patients and with markers of higher quality prior to the introduction of the new contract.

Keywords: Family doctors, pay for performance, perceptional evaluation
JEL Classification: C21, I10, J41

September 2007

Acknowledgements
We are grateful to Hugh Gravelle, Geir Godager and participants in the 2007 Nordic Health Economists' Study Group for comments on a previous version of this paper. The Second National Survey of Doctors in NHSScotland was funded by Scottish Executive Pay Modernisation. HERU receives funding from the Chief Scientist Office of the Scottish Executive Health Department. The views expressed are those of the authors and not necessarily those of the funders.

* Correspondence to Matt Sutton: m.sutton@abdn.ac.uk

Introduction

The UK government's introduction of a new pay for performance element into the contract for Family Doctors (FDs) was a radical and expensive experiment aiming to improve the quality of primary medical care (Roland, 2004). This initiative was not piloted and was introduced universally. The system for measuring achievements was introduced at the same time as the new contract and there is therefore no comparable pre-intervention data on quality with which to evaluate its impact.

This missing data issue is a classic problem in policy evaluation (Blundell and Costa Dias, 2002). The available methods seek to address the problem of not observing the outcome variable when individuals are both in and out of the programme under consideration. Constructing the counterfactuals for the participants and/or non-participants are the central challenges. In our context we wish to construct a counterfactual that is the outcome variable for each participant prior to the introduction of a universally-applied programme. This paper quantifies the changes in quality due to the reform through an innovative use of participants' perceptions of the changes attributable to the new contract.

In the first year, the average partnership in Scotland achieved 982 (94%) of the 1,050 quality points available and was paid £75 for each of these points. Over one-third of the 2,141 participants who responded to our postal survey thought that the quality of patient care had remained the same. We estimate the relationship between the achievements of these partnerships in the first year and a range of population and partnership characteristics. We use this equation to postdict pre-intervention values for the remainder of participants. The mean net gain in performance is estimated to be less than 4 quality points, with a mean decrement for those reporting worsened quality of 5 points and a mean increment for those reporting improved quality of 7 points.

Background

FDs in the UK organise themselves in partnerships and contract with the tax-financed National Health Service. These partnerships tend to be small, with a mean (inter-quartile range) size of 4 (2-6) FDs. They provide continuing care, free at the point of delivery, to the populations that register with them. This care involves diagnosis, advice, minor treatments, prescriptions and referrals to specialist doctors. They are paid an annual capitation fee for each registered patient and, from this gross income, employ nurses, administrators and other health professionals, and extract net income as profit.

This payment system was overhauled in April 2004. The capitation payments for patient registrations were revised but, in practice, partnerships have been protected from any impact on their income. Partnerships were also offered the opportunity to opt-out of providing care out of office hours. The element of this new contract that has attracted most interest was the inclusion of a new 'pay for performance' element designed to improve the quality of care (NHS Confederation and British Medical Association, 2003). This 'Quality and Outcomes Framework' comprises 146 indicators of the quality of care provided. Through performance on these indicators partnerships accrue quality points, up to a maximum of 1,050. Quality is assessed across several domains including clinical care, organisational aspects, patient experience, additional services and an access bonus. The clinical care domain covers 10 chronic conditions.

This scheme has been described as "the boldest such proposal attempted anywhere in the world" (Shekelle, 2003). Moreover, it absorbs a substantial amount of National Health Service resources. In 2005/6, the second year of the QOF, the NHS in Scotland paid its 1,010 general practice partnerships a total of £134m for their QOF achievements.

The QOF was designed and negotiated between the government and the British Medical Association (BMA), the doctors' trade union (Roland, 2004). Participation was officially voluntary but almost all partnerships participated in the first year. The

detailed data on performance are published on a government website (www.isdscotland.org/QOF). The average level of performance was impressive in the first year and increased in the second year. The median percentage of available total points achieved by partnerships in Scotland in March 2006 was 99.2% and 15% of partnerships achieved the maximum 1,050 points (Gravelle et al, 2007).

Concerns have been expressed about the value for money of this reform (NHS Employers and BMA, 2006). The government terms this 'Benefits Realisation' (Scottish Executive, 2004, 2005), by which it means mechanisms for ensuring the delivery of benefits rather than a dawning realisation that benefits will need to be demonstrated to justify the expenditure. The BMA is under criticism for the substantial pay rise that its members have enjoyed and is keen to emphasise that this was value for money (BMA, 2007). It believes that the government underestimated the quality of primary medical services prior to the new contract and claims that these new payments represent 'back-pay' for the quality of service delivered over several years (Buckman, 2007).

The introduction of the QOF is not atypical of system-wide health reform. It is vigorously debated and yet its implementation was not designed to evaluate its impact. With the QOF, a complicated measurement system was introduced to quantify the quality of each partnership. The data generated were not collected prior to the new contract and it was introduced simultaneously and universally across the UK. Therefore all that is apparent is that, after the contract, quality was consistently high across the board. It is not known whether partnerships were already performing to these standards prior to the introduction of the QOF or whether they have responded substantially to the new incentives.

A number of studies have examined changes in quality on a limited range of clinical care indicators in small samples of practices (Campbell et al, 2007; Gulliford et al, 2007; Steel et al, 2007; Tahrani et al, 2007). They indicate that quality was already on an upward trend but that the introduction of the QOF may have coincided with an above trend improvement. Improvements in quality have been found for patients with the ten incentivised conditions but quality does not appear to have changed for other patients.

However, such studies have focused on the clinical elements of the QOF and derived indicators similar to those in the QOF from patient records. Prior to the QOF there were no explicit incentives for recording such information routinely. It still remains unclear therefore whether the new contract has improved the recording of quality or underlying quality. In addition, it is not known if these partnerships are typical of all UK partnerships or if changes in clinical care are representative of changes across all of the aspects of quality that the QOF seeks to capture. Finally, since intentions to introduce, and the content of, the QOF were widely disseminated in advance, it is unclear how to define the pre-intervention period. We adopt a method that avoids these problems using a dataset linking FDs' perceptions of the treatment effect with their achievements in the first year after the policy's introduction.

Data

We undertook a postal survey of FDs working in Scotland in early 2006. The survey included questions on job satisfaction, work commitments and attitudes to workload (French et al, 2006). After two reminders, the response rate was 52% giving a sample size of 2,141 respondents. This response rate is commensurate with previous surveys of FDs in the UK (Scott et al, 2006; Whalley et al, 2006). We compare the characteristics of respondents with those of all FDs and allow for selection on observables in the calculation of the average treatment effect.

As part of the survey, respondents were asked "Has the quality of patient care in your practice changed as a result of the new GMS contract?" They were asked to indicate one of three responses: "Yes, it has improved"; "Yes, it has become worse"; or "No, it has remained the same". 2,062 (96.3%) respondents answered this question. 245 (11%) respondents did not agree to linkage of their data to additional sources. A total of 1,827 respondents answered the relevant question and are linked to external data.

Data on the QOF achievements of partnerships in March 2005 and March 2006 are published by the Information Services Division (ISD) of NHS National Services

Scotland (www.isdscotland.org/QOF). We use total points and points for each of the main headings: clinical care, organisational quality, patient experience, additional services and the access bonus.

The characteristics of all 4,102 FDs were obtained from the General Practitioner Contractor Database held at ISD and are for October 2005. From this dataset we derived the gender composition and mean age of FDs in each partnership. ISD also provided information on three pre-QOF quality markers: whether the partnership included at least one FD who could act as a trainer to doctors in training, and had received Practice Accreditation or a Quality Practice Award from the Royal College of General Practitioners.

Socio-economic characteristics of partnership populations are not collected directly but can be attributed to practices on the geographical distribution of their patient registrations. From the 2001 Census we calculated the age-standardised rate of limiting long-term illness and the proportion from minority ethnic groups for each of 42,604 Output Areas. Partnership values of these variables were calculated as weighted averages of the values for the areas of residence from which the partnership drew its population as at September 2005.

Partnerships may respond to competition for patient registrations by improving the quality of their care. Using the registrations database we calculated a measure of the extent to which each partnership was a monopoly provider for its local population. We calculated a Herfindahl index for each of 6,505 areas (called 'datazones') based on the squared proportions of the area population registered with different partnerships. For the partnership we obtained a weighted average of these values as for the other area characteristics. A partnership receives a value of one if it is the sole provider to each of the areas from which it draws its population. The index approaches zero as the partnership draws patients from areas served by many partnerships.

There are variations between partnerships in the computer systems used to manage patient records. The use of one of these (GPASS) is subsidised by central government. The functionality of this system has been criticised and alternatives are perceived to

offer more accurate information during the year on expected QOF performance. We obtained information on the types of computer system in use by partnerships as at March 2005.

Finally, FD partnerships hold contracts with local Health Boards who are responsible for ensuring access to good quality primary care within their areas. These organisations provide a range of support services to partnerships that may influence performance and we included indicators for Health Boards in our analysis.

Method

In the absence of the intervention, general expressions for the level of outcome (y) at times *t-1* and *t* would be:

$$y_{it-1} = \alpha_{t-1} + x_{t-1}'\beta_{t-1} + \varepsilon_{it-1} \quad (1)$$

$$y_{it} = \alpha_t + x_t'\beta_t + \varepsilon_{it} \quad (2)$$

Over a short time period we can assume that the factors x are fixed, as are the coefficients α_t and β_t. The y_{it} can then be written as the previous period value plus the change over time (Δy_i^*), and we can substitute for y_{it-1} with the time-invariant function of the x factors to give:

$$\begin{aligned} y_{it} &= y_{it-1} + \Delta y_i^* \\ &= \alpha + x'\beta + \Delta y_i^* + \varepsilon_{it-1} \end{aligned} \quad (3)$$

Equation (3) could be estimated directly if estimates of Δy_i^* were available. We assume that only the sign of Δy_i^* is known. This categorical variable is positive for those who thought outcomes have improved, zero for those who perceived no change and negative for those who thought outcomes had got worse.

If the treatment effects are homogeneous with respect to the *x* variables, we can substitute our categorical realisations of the change in outcome into equation (3) to give:

$$y_{it} = \alpha + x'\beta + \gamma \mathbf{1}[\Delta y = -1] + \delta \mathbf{1}[\Delta y = +1] + \varepsilon_{it-1} \qquad (4)$$

in which **1**[.] is an indicator function taking a value of one if the statement is true and zero otherwise. We expect to find $\gamma<0$ and $\delta>0$. Note that if we omit the *x* variables and estimate:

$$y_{it} = a + c\mathbf{1}[\Delta y = -1] + d\mathbf{1}[\Delta y = +1] + e_{it-1} \qquad (5)$$

the estimated coefficients (*c* and *d*) will be subject to omitted variable bias unless the effects of the intervention are uncorrelated with the variables determining the pre-intervention level of outcome.

Equation (4) assumes that the treatment effect takes only three values: γ, 0 and δ. Heterogeneous treatment effects are more probable. To test for this, equation (4) can be augmented with interactions between the two indicator functions and each of the *x* variables. To cater for heterogeneous treatment effects, a parametric approach to the matching estimator can be undertaken (Blundell and Costa Dias, 2002). Equation (4) can be estimated for those participants where $\Delta y = 0$:

$$y_{it} \equiv y_{it-1} = \alpha^0 + x'\beta^0 + \varepsilon_{it-1} \quad if\ \Delta y = 0 \qquad (6)$$

*Post*dicted values of outcome can then be generated for all participants using the estimated coefficients from (6):

$$\hat{y}^0_{it-1} = \hat{\alpha}^0 + x'\hat{\beta}^0 \qquad (7)$$

The difference between the observed and these postdicted values $(y_{it} - \hat{y}^0_{it-1})$ will equal the treatment effect if $E[y_{it-1} \mid x] = E[y_{it-1} \mid x, \Delta y = 0]$. Thus the validity of these

estimates relies on the assumption that the pre-intervention equation linking the outcomes to the *x* factors does not vary with the sign of the treatment effect.

Analysis

Determinants of perceptions of change

We begin by analysing the qualitative perceptions of quality change. We assume that these perceptions are determined by a latent structure:

$$\Delta y_i = j \quad if \quad \mu_j < \Delta y_i^* \leq \mu_{j+1} \tag{8}$$

Setting $\mu_{-1} = -\infty$ and $\mu_2 = +\infty$, an individual will report:

$$\begin{aligned} \Delta y_i &= -1 \quad if \quad \delta + x' + w_i \leq \mu_0 \\ \Delta y_i &= 0 \quad if \quad \mu_0 < \delta + x'\lambda + w_i \leq \mu_1 \\ \Delta y_i &= 1 \quad if \quad \delta + x'\lambda + w_i > \mu_1 \end{aligned} \tag{9}$$

We assume that the w_i are distributed as N(0,1) and estimate an ordered probit regression model.

Determinants of outcome

We estimate equations (5) and (4) for the entire sample of respondents and equation (6) for the subset of respondents that reported no change in outcome.

The primary outcome variable (the number of QOF points achieved) is left skewed (see Figure 1) and has a maximum at 1,050 points. We calculate the number of points 'missed' (=1,050–points achieved) and model this variation using a negative binomial regression model.

We also decompose the primary outcome by the main categories of quality points: clinical care, organisational care, patient experience, additional services and the access bonus. We use the same regression framework and analyse the number of points missed from the maximum available for each of these five dependent variables.

The measures of outcome are available at partnership not individual doctor level. Accordingly, we cluster the standard errors by partnership.

Estimated treatment effects

We calculate the average treatment effect in the survey respondents (ATE^S) using:

$$\begin{aligned} ATE^S &= p_{\Delta y=-1}\overline{(y_{it} - \hat{y}^0_{it-1})}_{\Delta y=-1} + p_{\Delta y=0}\overline{(y_{it} - \hat{y}^0_{it-1})}_{\Delta y=0} + p_{\Delta y=1}\overline{(y_{it} - \hat{y}^0_{it-1})}_{\Delta y=1} \\ &= p_{\Delta y=-1}\overline{(y_{it} - \hat{y}^0_{it-1})}_{\Delta y=-1} + p_{\Delta y=1}\overline{(y_{it} - \hat{y}^0_{it-1})}_{\Delta y=1} \end{aligned} \quad (10)$$

in which p are the observed proportions in each category of perceived quality change.

We also present the ATE for the entire population of FDs (ATE^P), by taking the mean value of differences between observed outcomes at time t with postdicted values of outcomes at time $t\text{-}1$.

$$ATE^P = \overline{(y_{it} - \hat{y}^0_{it-1})} \quad (11)$$

Results

Table 1 shows that the average number of points achieved was high across each of the QOF domains in the first year. Partnerships achieved an average of 982 points out of the 1,050 available. The percentage of points available that were achieved was similar on the clinical domains (93.5%) to that achieved overall (93.6%). Partnerships achieved a slightly lower percentage on the organisational domains (91.1%) and higher percentages on patient experience, additional services and access. Average

performance improved by 50 points in the second year to 1,032 points. Average performance increased across all domains.

Amongst all of the 2,062 survey respondents that responded to this question (Table 2), 7.7% said that the quality of patient care had got worse and 37.2% thought that quality had remained the same. Over half (55.2%) of respondents thought that quality had improved. The 1,827 respondents who consented to further linkage of their data were slightly more likely to report an improvement in quality.

For the QOF, quality is assessed at partnership level. We would therefore expect agreement in the perceptions of FDs within the same partnership. Across the 1,057 respondents for whom there is at least one other respondent from the same partnership, the partnership to which the FD belongs explains 32% of the variation in perceptions. The intraclass coefficient equals 0.082 (SE=0.029).

Summary statistics for the variables used are provided in Table 3. Figures are provided for the survey sample and all FDs. The mean number of points achieved by the survey sample was 986, slightly higher than that for all FDs. The sample has slightly lower morbidity than the entire population. For the average partnership, 2% of the population is from ethnic minority groups. The average age of FDs is 45 years and 45% of FDs are female. The survey sample shows slight indications of being from higher quality partnerships than the population as a whole – the sample has greater proportions with training status, a Quality Practice Award and Practice Accreditation. The average value of the Herfindahl index is 0.38.

Table 4 shows the results of the ordered probit regression of the perception of quality change on partnership characteristics. FDs in partnerships with higher ethnic minority proportions, older FDs and more female FDs were less likely to perceive an improvement in quality. Training partnerships were more likely to report a perceived improvement in quality. Higher Herfindahl indices are associated with a lower probability of perceiving an improvement in quality.

Respondents reporting that quality had worsened missed more of the available points (Table 5, Model A) compared to those that thought quality had remained the same.

Those that reported improved quality missed the fewest number of points. The indicators are jointly significant ($\chi^2(2)=14.31$, $p<0.001$).

The coefficients retain their signs but reduce in magnitude when partnership characteristics are introduced into the model (Model B). They are jointly significant only at the 10% significance level ($\chi^2(2)=5.36$, $p=0.069$). This indicates correlation between the sign of the treatment effect and the *x* variables. Partnerships serving areas with sicker populations missed more points, as did those with older FDs. Fewer points were missed by partnerships with training status, a Quality Practice Award or Practice Accreditation. Practices facing more competition achieved more quality. Partnerships with two of the four alternative computer systems also missed significantly fewer points compared to those using the government-subsidised system.

The log-likelihood for the model with interactions between the perceptions indicators and each of the *x* variables (not shown) equals -8984.03. The perceptions indicators and their interactions are jointly significant ($\chi^2(53)=93.71$, $p<0.001$), as are the interaction terms alone ($\chi^2(51)=79.86$, $p=0.006$). There is therefore strong evidence of treatment effects that are heterogeneous with respect to the *x* variables.

Amongst the sub-sample of participants that reported no change in quality, the variation in quality is significantly associated with population morbidity, the three quality indicators (training, QPA and PA) and one of the non-governmental computer systems (Model C). We interpret this equation as describing the pattern of quality in existence prior to the introduction of the new contract and use it to generate pre-intervention values for the remainder of the participants.

The resulting estimates of the quality improvement associated with the QOF are summarised in Table 6. The respondents who reported an improvement in patient care achieved an average of 990.7 points in 2004/5. The average postdicted value for these respondents is 983.5 points, indicating an increase of 7.2 points between the year before and the year of the introduction of the QOF. By definition, the average number of points achieved by those who thought quality had remained the same (981.2 points) is equal before and after the introduction of the QOF. Those respondents who perceived a worsening of quality achieved an average of 968.5 points in 2004/5. The

mean postdicted value for this group is 973.7 points, representing an average reduction of 5.2 points after the QOF's introduction.

Across all 1,797 respondents included in the analysis, we find an average treatment effect (equation (10)) of +3.7 points. Applying the postdicted values to all FDs and estimating the ATE^P using equation (11) generates an estimate of the average improvement from the QOF for all participants of +2.0 points.

The distribution of estimated treatment effects across individual participants is shown in Figure 2. The variation is moderately negatively skewed and shows considerable variation (SD=67.6). Table 7 shows the average values of the estimated treatment effects for two categories of each of the population and partnership characteristics. Partnerships with lower values of population morbidity have a larger mean treatment effect, as do those with higher ethnic minority population proportions and younger FDs. The mean treatment effect for those partnerships above the median value of the Herfindahl index is negative. There is a clear pattern of larger average treatment effects for those partnerships with the three pre-QOF quality markers (training status and receipt of one of two awards from the Royal College).

Our estimates of the quality gains for each of the main areas within the QOF are shown in Table 8. The clinical area accounts for 62% of the total points available and 83% of the estimated mean net gain in total points. The estimated gains in organisational quality points are proportional to their share in the total points available. These findings suggest that FDs concur with the weightings attached to different dimensions of quality by the QOF points system when they form their perceptions. An exception is the 'patient experience' area, on which those reporting worsened quality experienced a reduction of 1.6 points from the 100 available. These 100 points are earned based on the length of patient consultations and the undertaking of patient surveys. FDs believing that the new targets for ten specific conditions had diverted effort away from other aspects of patient care might have been particularly sensitive to this in forming their perceptions of the change in overall quality.

Discussion

The introduction of the Quality and Outcomes Framework for UK family doctors was expensive and unpiloted. Mean performance observed in the first year following introduction was very high but it is known whether this represented a substantial response to an innovative pay for performance system or high achievement prior to its introduction. We have proposed and implemented a perceptional evaluation of this intervention. This is a method for estimating the gains associated with the new system based on participants' perceptions of the effects of this policy change.

The majority of respondents reported that the quality of patient care had improved but a substantial proportion of respondents also thought that quality remained unchanged. We identified a range of partnership characteristics that were significantly related to levels of achievement in the first year. These characteristics exerted a similar influence in the entire sample and in the sub-sample who reported that the intervention had had no effect on the quality of patient care. The qualitative effects of these variables were generally as expected. Proxy indicators of quality tended to be associated with higher levels of achievement and markers of more complex case-mix tended to be associated with lower quality.

We used variation in these correlates of quality to postdict values for all partnerships. The results satisfied the basic properties expected. Respondents that reported improvements in quality had lower mean postdicted values than they achieved in the first year following introduction. Respondents that reported deteriorations in quality had higher mean postdicted values than they achieved in the first year. Within the sample, we estimated that the mean increment in points achievement was 3.7 points, from a total available of 1,050 points. Extrapolation to the entire population of FDs generated a mean estimate of 2.0 points.

Our finding of moderate improvement attributable to the QOF is consistent with other studies that have charted the progression of specific clinical indicators through the introduction of the QOF. The quality of primary medical care was already showing

trend improvement and the introduction of the QOF was associated with moderate increases in quality above this trend.

Our estimates are based on an identifying assumption and rely on the validity of participants' perceptions. Since responses to the questionnaire were anonymous, and the questionnaire was sent from two agencies with no financial interest in the responses, participants had no direct incentive to bias their responses. The distribution of reported perceptions does not suggest justification bias by participants.

Unlike previous studies, our results are not confounded by improvements in the recording of information rather than underlying improvements in the actual quality of services. Moreover, our assessment of quality applies across all of the 146 criteria captured by the QOF and is independent of the date at which benefits from the intervention are thought to have occurred.

This framework we have proposed for using participant perceptions to evaluate reforms retrospectively has wider potential application. Our foundation assumption is that outcomes at time *t-1* equal outcomes at time *t* for those that perceive that outcomes have remained the same. We exploited the observed variation in outcomes for this group to estimate the outcomes at time *t-1* for those that reported changes in outcome. Identification was provided by assuming that this equation that links outcomes at time *t-1* to a set of *x* variables is independent of perceptions at time *t* that outcomes have changed. The properties of this estimator should be explored further in future research.

Future research on our empirical focus should concentrate on assessing the generalisability of our findings across other parts of the UK. Additionally, our questionnaire offered the opportunity for respondents to add free-text comments to explain their perceptions of whether and how the new contract had changed patient care. Analyses of these comments would provide further explanation of the results that we have obtained here. In particular, this would help us to identify whether the moderate gains from the QOF are attributed to a 'ceiling effect' (because the quality of care was already approaching a maximum) or to a design problem (because the QOF distorted priorities from other activities and patient groups).

Our principal finding that the new pay for performance system for UK family doctors generated only a modest improvement in the quality of patient care is not unexpected given previous research and commentary, but raises serious questions about its efficiency. The expenditure of £75 million (equivalent to £15 per head of population) was incurred because of the mean achievement of 982.3 points, but only 2 of these points were effectively purchased by this new system since 980.3 of these points were being achieved prior to its introduction. In the second year, mean achievement increased by 50 points when the payment per point increased by 67% to £125 per point. This may indicate longer-term benefits through learning or a financial response. Nevertheless, neither the original impact we have estimated nor these marginal gains suggest that this new contract represented value for money.

References

Blundell R, Costa Dias M. *Alternative approaches to evaluation in empirical microeconomics.* Institute for Fiscal Studies/Department of Economics UCL. 2002. cemmap working paper CWP10/02.

British Medical Association. *The value of general practice: the facts.* London 2007.

Buckman L. Is doctors' self interest undermining the National Health Service? *British Medical Journal* 2007; 334: 235.

Campbell S, Reeves D, Kontopantelis E, Middleton E, Sibbald B, Roland M. Quality of primary care in England with the introduction of pay for performance. *New England Journal of Medicine* 2007; 357(2): 181-90.

French F, Geue C, Ikenwilo D, Needham G, Rooke C, Skåtun D, Sutton M. *Second national survey of non-training grade doctors in NHS Scotland: changes in job satisfaction, work commitments and attitudes to workload following contractual reform.* Report to the Scottish Executive Health Department 2006. [Available from www.paymodernisation.scot.nhs.uk/]

Gravelle H, Sutton M, Ma A. *Doctor behaviour under a pay for performance contract: evidence from the Quality and Outcomes Framework*. CHE Research Paper 28. University of York 2007.

Gulliford MC, Ashworth M, Robotham D, Mohiddin A. Achievement of metabolic targets for diabetes by English primary care practices under a new system of incentives. *Diabetic Medicine* 2007; 24(5): 505-11.

NHS Confederation and British Medical Association. *New GMS contract: Investing in General Practice*. London 2003.

NHS Employers and British Medical Association. *Delivering investment in general practice. Revisions to the GMS contract, 2006/07.* London 2006.

Roland M. Linking physician pay to quality of care: a major experiment in the United Kingdom. *New England Journal of Medicine* 2004;351:1448-54.

Scott A, Gravelle H, Simoens S, Bojke C, Sibbald B. Job satisfaction and quitting intentions: a structural model of British General Practitioners. *British Journal of Industrial Relations* 2006; 44(3): 519-540.

Scottish Executive. *Delivering the benefits of pay modernisation in NHSScotland.* NHS HDL (2005) 28. Edinburgh 2004.

Scottish Executive. *Draft Budget 2005-06.* Annex A. Edinburgh 2004.

Shekelle P. New contract for general practitioners. *BMJ* 2003;326:457-458.

Steel N, Maisey S, Clark A, Fleetcroft R, Howe A. Quality of clinical primary care and targeted incentive payments: an observational study. *British Journal of General Practice* 2007; 57: 449-54.

Tahrani AA, McCarthy M, Godson J, Taylor S, Slater H, Capps N, Moulik P, Macleod AF. Diabetes care and the new GMS contract: the evidence for a whole county. *British Journal of General Practice* 2007; 57: 483-5.

Whalley D, Bojke C, Gravelle H, Sibbald B. GP job satisfaction in view of contract reform: a national survey. *British Journal of General Practice* 2006; 56 523: 87 –92.

Table 1 Points available and achieved by QOF domain

QOF domain	Points available	Mean points achieved in 2004/5	Mean points achieved in 2005/6
Clinical	650	608	641
Organisational	214	195	207
Patient experience	100	96	98
Additional services	36	35	36
Access	50	49	50
Total	1,050	982	1,032

Notes: Based on all 4102 FDs in 2004/5 and all 4108 FDs in 2005/6. Clinical total includes 'holistic care' points. Organisational total includes 'quality practice' points.

Table 2 FDs' perceptions of quality of patient care following new contract

Perception of change in quality of care	All surveyed FDs		FDs consenting to linkage	
	Frequency	Percentage	Frequency	Percentage
Worsened	158	7.7	134	7.3
Same	766	37.2	661	36.2
Improved	1,138	55.2	1,032	56.5
Total	2,062	100.00	1,827	100.00

Table 3 Summary statistics

	All FDs		FD respondents	
Variable	Mean	St.Dev. (Min, Max)	Mean	St.Dev. (Min, Max)
Achievements				
Total QOF points	982	79 (294, 1050)	986	74 (486.5, 1050)
Population characteristics				
Standardised illness ratio	97.4	22.4 (50.5, 186.7)	96.3	22.4 (50.5, 186.7)
Ethnic minority proportion	0.019	0.023 (0, 0.282)	0.019	0.021 (0, 0.213)
FD characteristics				
Mean FD age (years)	45.2	4.3 (29, 67)	45.1	4.1 (32.5, 67)
Proportion Female FDs	0.449	0.206 (0, 1)	0.460	0.199 (0, 1)
Practice characteristics				
Training practice	0.389	0.488 (0, 1)	0.421	0.494 (0, 1)
Quality Practice Award	0.064	0.244 (0, 1)	0.067	0.250 (0, 1)
Practice Accreditation	0.470	0.499 (0, 1)	0.503	0.500(0, 1)
Herfindahl Index	0.377	0.253 (0.078, 0.998)	0.385	0.255 (0.078, 0.998)
Computer system				
Egton MIS	0.060	0.238 (0, 1)	0.063	0.243 (0, 1)
IPS	0.058	0.233 (0, 1)	0.062	0.242 (0, 1)
Protechnic Exeter	0.012	0.111 (0, 1)	0.015	0.122 (0, 1)
iSOFT	0.029	0.169 (0, 1)	0.030	0.169 (0, 1)

Table 4 Ordered probit regression of perceived change in quality

Variable	Coefficient	S.E.
Standardised illness ratio	-0.013	0.158
Ethnic minority proportion	-4.057*	1.828
Mean FD age (years)	-0.026**	0.008
Proportion Female FDs	-0.427**	0.163
Training practice	0.202**	0.064
Quality Practice Award	0.202	0.125
Practice Accreditation	-0.032	0.067
Herfindahl Index	-0.426**	0.155
Egton MIS	0.137	0.151
IPS	0.169	0.114
Protechnic Exeter	0.020	0.268
iSOFT	0.031	0.224
N	1827	
Wald statistic	$\chi^2(26)=81.71$	
Log pseudo-likelihood	-1573.79	

* significant at 5%; ** significant at 1%. Standard errors adjusted for clustering by partnership. Includes 14 dummy variables for Health Boards.

Table 5 Negative binomial regression of available QOF points missed

	Model A		Model B		Model C	
Variable	coeff.	s.e.	coeff.	s.e.	coeff.	s.e.
Worsened	0.175*	0.076	0.103	0.069		
Improved	-0.087*	0.042	-0.047	0.040		
Standardised illness ratio			0.850**	0.101	1.019**	0.181
Ethnic minority proportion			-0.218	1.253	-1.606	1.977
Mean FD age (years)			0.012*	0.005	0.010	0.008
Proportion Female FDs			-0.122	0.098	-0.184	0.160
Training practice			-0.268**	0.041	-0.289**	0.071
Quality Practice Award			-0.673**	0.092	-0.545**	0.183
Practice Accreditation			-0.254**	0.041	-0.337**	0.070
Herfindahl Index			0.192*	0.093	0.028	0.160
Egton MIS			-0.560**	0.089	-0.480**	0.154
IPS			-0.300**	0.084	-0.026	0.130
Protechnic Exeter			0.052	0.149	-0.201	0.260
iSOFT			0.110	0.110	0.042	0.207
N	1797		1797		643	
LR test	$\chi^2(2)=13.43$		$\chi^2(28)=503.18$		$\chi^2(26)=199.51$	
Log pseudo-likelihood	-9267.07		-9022.20		-3260.76	

* significant at 5%; ** significant at 1%. Model A contains perception of quality change only. Model B contains perception of quality change and partnership and population characteristics. Model C is estimated for participants who perceived no quality change using partnership and population characteristics only. Models B and C also contain 14 Health Board indicators.

Table 6 Estimated changes in quality points

Group	Number	Mean postdicted points	Mean points in 2004/5	Change in mean points
Survey sample				
Worsened	134	973.7	968.5	-5.2
Same	643	981.2	981.2	0.0
Improved	1,020	983.5	990.7	+7.2
All	1,797	981.9	985.6	+3.7
All FDs				
All	4,102	980.3	982.3	+2.0

Table 7 Mean treatment effects by population and partnership characteristics

	Low value		High value	
Characteristic	N	Mean effect	N	Mean effect
Standardised illness ratio*	900	5.4	897	1.9
Ethnic minority proportion*	900	-1.3	897	8.7
Mean FD age*	917	6.4	880	0.9
Proportion female FDs*	1,179	4.4	618	2.3
Herfindahl index*	900	7.6	897	-0.3
Training status+	1,042	-1.2	755	10.4
Quality Practice Award+	1,677	3.0	120	13.7
Practice Accreditation+	893	1.7	904	5.7

* Continuous variables categorised into two groups with lowest and highest values. + Binary variables for which the high value indicates presence of this partnership characteristic.

Table 8 Estimated changes in mean points by main QOF area

QOF area	Available points	Sample perceiving worsened quality	Sample perceiving improved quality	Sample
Clinical	650	-2.10	+5.70	+3.07
Organisational	214	-0.17	+1.35	+0.75
Patient experience	100	-1.64	+0.20	-0.01
Additional services	36	-0.22	+0.07	+0.02

Figure 1 Total quality points achieved by partnerships in 2004/5

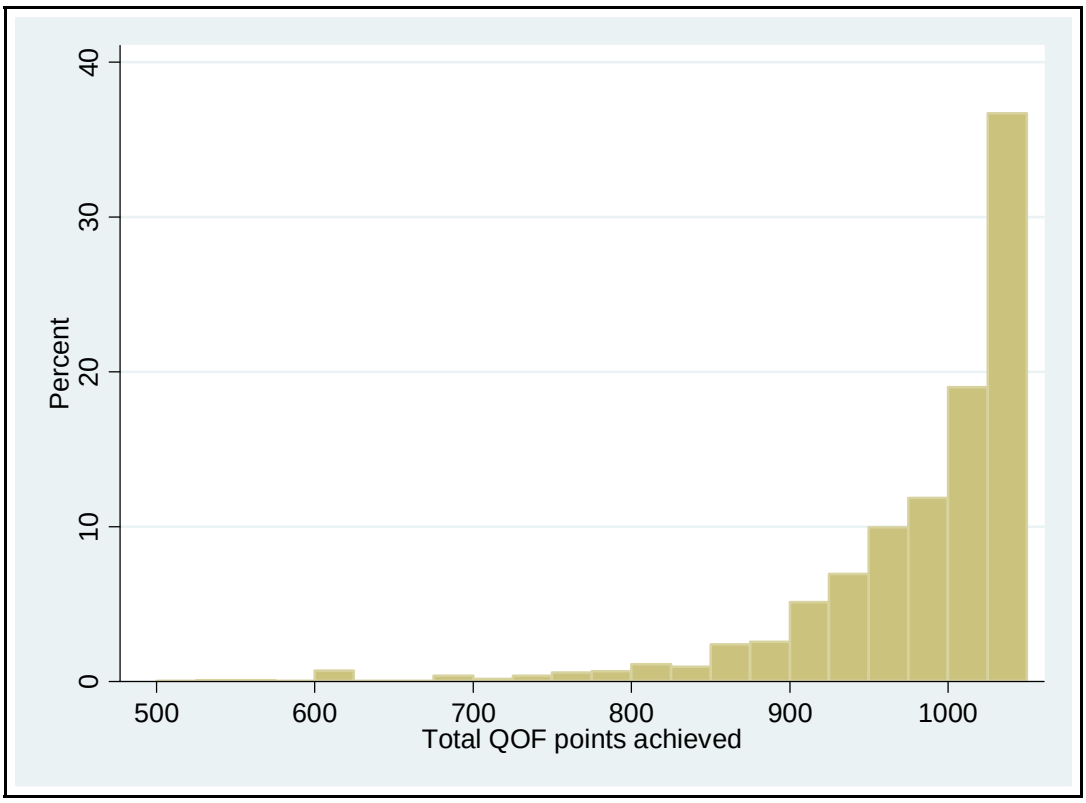

Figure 2 Distribution of estimated treatment effects across participants