The effect of a pay for performance system on GPs' monetary and non-monetary motivations

Yan Feng[1], Shelley Farrar[2], Matt Sutton[3], Ada Ma[4]

HESG and NHESG

Aberdeen

August 2008

Draft not for quotation

1. Introduction

This paper explores whether the use of pay for performance remuneration efficiently motivate GPs. We analyse whether Quality and Outcomes Framework (QOF) has changed GPs' motivation structure and how non-monetary motivation responds to the new payment system. In section two we discuss the background of QOF. Section three explains how GPs in UK are remunerated. Section four describes the model that is developed to explain GPs' behaviours under the QOF setting. Section five introduces the data we used and the econometrics methods we applied in the empirical tests are in section six. Results are summarized in section seven, followed by discussion in section eight and conclusion in section nine.

2. Background

General Practitioners (GPs) are also called family practitioners. They play an important role in the UK National Health Service (NHS) system. Except emergency, GPs are usually the first point of contact for patients to all medical services. As gatekeepers, GPs are responsible for referring patients from primary care sectors to hospitals (secondary care sectors). They are also important for scarce NHS resources allocation. Nearly half of the NHS doctors and dentists in UK are made up by GPs (Review Body on Doctors' and Dentists' Remuneration, 2007).

[1] Correspondence to Yan Feng: y.feng@abdn.ac.uk. Health Economics Research Unit, University of Aberdeen, Polwarth Building, Foresterhill, Aberdeen, AB25 2ZD

[2] Correspondence to Shelley Farrar: s.farrar@abdn.ac.uk

[3] Correspondence to Matt Sutton: Matthew.Sutton@manchester.ac.uk

[4] Correspondence to Ada Ma: ada.ma@abdn.ac.uk

In UK, GPs work in practices either single-handed or as a group with other GPs. On average list size of practice is around 5300 with 4.4 GPs per practice (Gravelle, Sutton and Ma, 2008). As independent contractors, they provide primary care services to registered patients. They bear the cost of care such as the cost for premises and nursing staff, etc. Practices are run as for profit businesses and therefore should be naturally motivated by incomes. However, working extra hours or doing extra work without additional payments are quite common in health care sectors. Such non-monetary motivation also can be called as intrinsic motivation.

The concept of intrinsic motivation stems from social psychology. Its definition varies because researchers made the definition under specific conditions. Deci (1971) initially addressed this question based on laboratory results. He found that tangible rewards may undermine intrinsic motivation. Deci (1975) defined intrinsic motivated behaviours as: *the behaviours which a person engages in to feel competent and self-determining*. Deci et al (1999) concludes from a meta-analysis that tangible rewards may have a substantially negative effect on intrinsic motivation under certain conditions. This question has been discussed in the last few decades in social psychology and become recognised by economists. In recent years, Frey integrated intrinsic motivation from social psychology into economic literature (Frey, 1997). He describes intrinsically motivated activities as: *activities one undertakes because he or she likes to do it or because the individual derives some satisfaction from doing his or her duty* (Frey and Oberholzer-Gee, 1997). His crowding theory accepts price effect as a basic behavioural principle and at the same time examines the change of individual's intrinsic motivation under financial incentives.

Two main mechanisms through which financial incentives influence intrinsic motivation are identified in the literature. The first is the result of satisfaction utility and the second is forgone income.

In terms of satisfaction utility, people may gain lower satisfaction from unpaid altruistic behaviours. This can be explained by two reasons. First exposure to financial incentives leads individuals to implicitly use cost benefit analysis to make decisions about how they employ their time (DeVoe and Pfeffer, 2007). As a result the satisfaction gained from unpaid work will decrease and they may be less likely to do unpaid works which have been undertaken previously (Deci, 1971). Second individuals may feel their altruistic behaviours have been undervalued by money, then decrease or cease participation in such altruistic activities

(Titmuss, 1970). In contrast, Frey (1997) argues that financial incentive may also foster the professional norms under certain conditions. Individuals may feel that their altruistic behaviours became recognised by money awards, which motive them to do more. In health economics, the satisfaction utility from non-monetary motivation could stem from ethics/patients' interests, intellectual satisfaction, autonomy, doctors' reputation, status and etc (Scott, 2000).

A second strand of literature uses Becker's theory of forgone income to explain how the introduction of financial incentives undermines intrinsic motivation. Becker (1965) argues that indirect cost should be treated on the same footing with direct cost and likewise when discussing non-working hours. Here the non-working hours refers to time spent on activities which cannot directly make money. In this paper, individuals maximize their utility by consuming a set of commodities under budget constraint and time constraint. The solution is that marginal utility from last unit consumed should equal to the summed utility of material cost and cost of time (forgone income). Forgone income is measured as the income lost from undertaking unpaid work rather than paid work. Applying this idea into GPs' non-monetary motivation, the resources spent on free service will not be remunerated but the value can be reconstructed by adding together the cost of material costs and the forgone income. In health economics, the costs for GPs' non-monetary motivated behaviours has not been fully discussed, especially the effect of forgone income.

Although there is no consistent answer about how non-monetary motivation responds to financial incentives, two things are quite sure here. First GPs have altruistic motivation which may make their motivation structure more complicated than individuals in private sectors. Second government uses financial incentives to motivate GPs based on the assumption that GPs respond to monetary mechanisms.

3. Paying GPs

The GP contract is the mechanism through which government attempt to incentivise GPs. One of the most significant changes in GPs' contract over the past two decades was the introduction of pay for performance system (Roland, 2004). Pay for performance is an incentive tool widely used in public agencies and organisations (Dixit, 2000). It creates a direct connection between payment to individuals and the quality and quantity of the work

they produced. With *valid measuring*[5] of the quality of health services became possible, the new General Medical Service (GMS) contract was introduced to UK NHS on 1st April 2004. As part of the new contract, QOF made performance payment be an important part of GPs' income. *QOF is a system to remunerate general practices for providing good quality care to their patients, and to help fund work to further improve the quality of health care delivered. The core philosophy underpinning the QOF is that incentives*[6] *are the best method of funding work, driving up standards, and recognising Practices' achievements*[7].

The introduction of QOF emphasis the management strategy of NHS that financial incentive is a good way to motivate GPs. It also reinforces the assumption about GPs' motivation from NHS management's perspective that GPs can be strongly motivated by monetary rewards. Assuming that GPs are only monetary motivated, the incentive scheme would be appropriate for GPs. However apart from monetary motivation, GPs are also motivated by non-monetary factors. The question here is whether the introduction of pay for performance will undermine or foster GPs' non-monetary motivation. There is no performance data before QOF introduced, therefore a test on the effect of introducing pay for performance system itself is not possible. However a policy adjustment was made in the following year which makes the empirical test possible.

QOF measures achievement against a range of evidence-based indicators, with points and payments awarded according to the level of achievement. The unit of assessment is practices. This paper only focuses on practices' performance in clinical domain which contains 10 disease areas in both financial years 2004/5 and 2005/6.[8] Each disease area has a number of indicators with numbers of points allocated. In total, both of the two financial years have 76 indicators with 550 points.

QOF payment is the sum of achieved payments from every indicator. The payment calculation for all indicators is identical: achieved points multiplied by the value of pounds per point. First the achieved points are determined by practices' performance, measured by performance index (α_i), which is a nonlinear function of the proportion of patients treated. They are not awarded any point if performance index is no more than the minimum payment

[5] See details from Brook R.H, McGlynn E.A, Cleary P. Quality of health care. Part 2: measuring quality of care. New England Journal of Medicine 1996; 335: 966-70.
[6] It refers financial incentives here.
[7] What is QOF? General Practice - Quality & Outcomes Framework; ISD; http://www.isdscotland.org/isd/3365.html.
[8] Financial year 2004/5 started from 1st April 2004 to 31st March 2005. The same rule applies to financial year 2005/6.

threshold (α_l). Full points will be given if the performance index reaches the maximum payment threshold (α_u). Achieved points became linearly correlated to performance index if the performance index is located between the maximum and minimum payment thresholds. Second, the value of pounds per point is adjusted by the Contractor Population Index (CPI) as well as Adjusted Disease Prevalence Factor (ADPF) before it used to calculated payments. After the adjustments, each general practice will have unique value of pounds per point for every disease.

QOF policy has been adjusted on 1st April 2005. The unadjusted value of pounds per point was raised from £75 to £124.6. As a result, income earned by practices with every patient treated increases. If forgone income from patient treated without payment could be equalized as the income from patient treated with payment as suggested in Becker (1965), then the policy adjustment suggests that forgone income of being altruistic has risen in 2005/6.

$$\begin{aligned}
&QOF\ income\ \ from\ every\ \ patient\ treated\ \ with\ \ payment\ \ for\ indicator\ i \\
&= \frac{0}{number\ of\ \ eligible\ \ patients \times (\alpha_u - \alpha_l)} = 0\ \ if\ \ \alpha_i \le \alpha_l \\
&= \frac{unadjusted\ \ value\ of\ \ pounds\ \ per\ \ point \times ADPF \times CPI \times full\ \ available\ \ points\ \times \dfrac{\alpha_i - \alpha_l}{\alpha_u - \alpha_l}}{number\ of\ \ eligible\ \ patients \times (\alpha_u - \alpha_l)} \\
&if\ \ \alpha_l < \alpha_i < \alpha_u \\
&= \frac{unadjusted\ \ value\ of\ \ pounds\ \ per\ \ point \times ADPF \times CPI \times full\ \ points}{number\ of\ \ eligible\ \ patients \times (\alpha_u - \alpha_l)}\ \ if\ \ 1 \ge \alpha_i \ge \alpha_u
\end{aligned}$$

4. Theoretical model

In this section, we develop theoretical model to provide a framework for empirical test on whether GPs' non-monetary motivation is crowed out by QOF.

We made a few assumptions about GPs' utility function. We assume the following. Cost function is convex and differentiable. Technology and satisfaction utility from providing free services are constant overtime. Practices where its performance index α_{gi} is located in $[0, \alpha_{ui}]$ are motivated by money only. Practices with GPs who are motivated by satisfaction of providing free services would provide treatments up to levels where their α_{gi} are located in $(\alpha_{ui}, 1]$.

4.1 Theoretical analyses

Practice g' utility from indicator i could be generalised as:

$$U_{gi} = I_{gi}(N_{gi1}, MI_{gi}) - C_{gi1}(N_{gi1}, c_{gi}) + S_{gi2}(N_{gi2}, s_{gi}) - C_{gi2}(N_{gi2}, fi_{gi}, c_{gi}) \quad (1)$$

where U_{gi} is the utility for practice g from indicator i. I_{gi} is the monetary income function which is dependent on two elements N_{gi1} and MI_{gi}. N_{gi1} is the number of patients treated with remuneration from QOF scheme. MI_{gi} is the marginal income from treat N_{gi1} patients. C_{gi1} is the resource costs function which is decided by two elements N_{gi1} and c_{gi}. c_{gi} is the marginal resource costs. $S_{gi2}(N_{gi2}, s_{gi})$ is the satisfaction utility function from having free service to patients. It is determined by two elements N_{gi2} and s_{gi}. N_{gi2} is the number of patients treated without remuneration under QOF scheme. s_{gi} is the marginal satisfaction from having free service to N_{gi2} patients. C_{gi2} is the resource costs function for providing free service, which is dependent on N_{gi2}, fi_{gi} and c_{gi}. fi_{gi} is the marginal forgone income for having free service to N_{gi2} patients.

GPs' monetary motivation is decided by the first two arguments I_{gi} and C_{gi1}. The non-monetary motivation is decided by the last two arguments S_{gi2} and C_{gi2}. As stated above MI_{gi} could be a proxy for fi_{gi}. Substitute the marginal forgone income into equation (1),

$$U_{gi} = I_{gi}(N_{gi1}, MI_{gi}) - C_{gi1}(N_{gi1}, c_{gi}) + S_{gi2}(N_{gi2}, s_{gi}) - C_{gi2}(N_{gi2}, MI_{gi}, c_{gi}) \quad (2)$$

If individuals motivated by monetary income only, then utility function could be rewritten as,

$$U_{gi} = I_{gi}(N_{gi1}, MI_{gi}) - C_{gi1}(N_{gi1}, c_{gi}) \quad (3)$$

Maximizing the utility by changing number of patients treated with payment gets,

$$\frac{\partial U_{gi}}{\partial N_{gi1}} = MI_{gi} - c_{gi} = 0 \quad (4)$$

Equation (4) suggests that if GPs are only monetary motivated, then they should provide services at the level where marginal income equals to marginal resource costs. If a practice's performance index is below the maximum payment threshold, increased marginal income will encourage GPs to increase their output till marginal income equals to marginal cost again.

If individuals are motivated by both monetary and non-monetary motivations, then the utility function takes the form of equation (2). However, under QOF scheme the monetary income became fixed when a practice's performance index has reached the maximum payment threshold. In other words, the monetary motivation arguments can be replaced by a constant. Rewriting equation (2) gives,

$$U_{gi} = \alpha_{gi} + S_{gi2}(N_{gi2}, s_{gi}) - C_{gi2}(N_{gi2}, MI_{gi}, c_{gi}) \quad (5)$$

Maximizing the utility with reference to the number of patients treated without remuneration,

$$\frac{\partial U_{gi}}{\partial N_{gi2}} = s_{gi} - (MI_{gi} + c_{gi}) = 0 \quad (6)$$

Equation (6) shows that the non-monetary motivated output will be produced at the level where the marginal satisfaction from providing free service equals the sum of marginal utility from forgone income and resource costs. The output of GPs who are motivated by both monetary and non-monetary motivations is the sum of pre-assigned number of patients treated with payment (N_{gi1}) and optimal number of patients treated without payments (N^{*}_{gi2}). The higher is the payment, the more expensive the free service becomes, as the unit cost of provision increased ($s_{gi} < (MI_{gi} + c_{gi})$). As a result, GPs will reduce their non-monetary motivated output and the performance index will decrease and move towards the maximum payment threshold.

The above discussion suggests that GPs may have adverse responses towards financial incentives. It depends on the performance index they have achieved. If a practice's performance index is located below the maximum payment threshold, then increasing marginal income will encourage it to increase treatment thus raising the probability of achieving maximum payment. As a result, the overall output will increase as well. If the performance index achieved is above the maximum payment threshold, then increasing the marginal income will undermine the provision of free service. It may lead to reduction of the overall output.

4.2 Hypothesis for empirical section: Marginal income from patients treated is expected to negatively correlate to non-monetary motivated output but positively correlated to monetary motivated output. How it affects GPs' overall motivation cannot be predicted by theoretical analysis. Empirical tests are needed.

5. Data

5.1 Performance indicators

We use GPs' performance data for financial years 2004/5 and 2005/6. Our research only focuses on the performance of GPs who were employed on GMS contract in Scotland.

GPs' performance data and QOF policy data are downloaded from the Information Services Division (ISD) website (http://www.isdscotland.org/isd/3305); GPs characteristics data comes from GP Contractor Database held at ISD. Patients' characteristics data comes from 2001 Census and Scottish Index of Multiple Deprivation 2004 (http://www.scotland.gov.uk/stats/simd2004). Standardised Illness Ratio (SIR), patients' age and proportion of patients who have ethnic minority status are come from 2001 Census. Patients' deprivation status comes from Scottish Index of Multiple Deprivation 2004.

Thirty nine indicators from QOF are eligible for this research (see table A3 for details).The eligible indicators must have a maximum and a minimum payment thresholds, which are used to distinguish monetary and non-monetary motivations. Second, time length for eligible indicators must be identical. This paper chooses indicators with 15 months time exposure. For example, eligible indicator CHD3 is defined as the percentage of eligible patients with coronary heart disease whose notes recorded smoking status in the past 15 months, except those who have never smoked where smoking status need be recorded only once. The selected indicators must appear in both years (2004/5 and 2005/6). Taking away observations with missing values there remain 64999 observations and 892 practices in the final dataset. The unit of observation is the performance data at practice level for a specific indicator under one financial year.

5.2 Dependent variables

Three dependant variables are modelled. *1) Non-monetary motivation:* number of patients freely treated is used as a proxy of non-monetary motivation because practices cannot obtain remuneration from those services. It is the difference between number of patients treated and number of patients needs to be treated in order to achieve the maximum payment. *2) Monetary motivation:* a dummy variable measures whether practices reach the maximum payment threshold. It is equal to one if practice has reached the maximum payment threshold, otherwise it takes zero. *3) Overall motivation:* number of patients treated. An example of

clinical indicators from the Coronary Heart Disease (CHD) disease domain and summary statistics of those three variables can be found in tables A1 and A2 in the appendix.

5.3 Independent variables (see summary statistics in table 1)

1. Income from every patient treated with payment. It is equal to forgone income from every patient freely treated. Their calculation has been discussed in section 3.

2. GPs' characteristics

Ex-Fund holding status: Fund holding program was introduced in 1991 and stopped in 1999. Fund-holders were asked to use fixed budget to purchase secondary care and selected surgeries for patients. They were encouraged to use the budget as efficiently as possible, so they may have became more capable in lowering costs. However, choosing to be a fund-holder could be a sign of being monetary motivated because it is an opportunity to make extra money. Fund-holders were allowed to spend any portion of their surplus to invest on their practices to improve their care of patients. This investment are likely to be within the budget they are allocated as if they have gone over the budget the practice would have to pay out of their own pocket. It is difficult to predict the effect of being an Ex-Fund holder on non-monetary motivation. Empirical tests are needed.

Dispensing status in 2004: This is a dummy variable with value of one for practices eligible to dispense drugs and appliances. Dispensing practices are usually located in remote areas or they have small quantities of drugs and appliances needs to buy. There are two variables included as proxy of rural status. One is a dummy variable *rural* measured by population mode in settlements. It has a value of one if there are fewer than 3000 people living in the settlement. The other is the *inverse of population density* which is measured by hectares per person.

Training status in 2004: This is a dummy variable about whether practices are eligible to provide training to qualified doctors who want to become GPs. Training practices are expected to have stronger motivations.

Practice herfindahl index: This index indicates the intensity of competition among practices. A low herfindahl index suggests high competitions. As a result, high working motivations are expected from practices with low herfindahl indices.

GPs characteristics also include *number of GPs per every thousand patients, percentage of qualified GPs who were not born in UK, average age of GPs, percentage of female GPs* and *number of GPs in one practice*.

3. Patients' characteristics

Standardised Illness Ratio (SIR) is used as proxy for patients' health status with 100 as the baseline. It measures the proportion of people who self assessed as having a limiting long-term illness and it is indirectly standardised by age. Higher ratios are associated with bad health status. It also suggests high material costs and as a result may lead to low working motivations from GPs.

Patients' health status may also be reflected by age, which proxied by *proportion of patients aged above 75* and *proportion of patients aged below 15.* Those patients are more vulnerable and are likely to have higher needs for primary care. The proportion of patients who are relatively vulnerable is expected to be negatively correlated with GPs' motivation, because providing services to them may be costly.

Scottish Index of Multiple Deprivation 2004 (SIMD2004) is used as proxy of patients' deprivation status. This measurement covers six individual deprivation domains: education, access, income, house, health and employment. A high SIMD score suggests patients have high deprivation status. Patients' characteristics also include a variable which measures *the proportion of patients with ethnic minority status*.

Table1: Summary statistics about independent variables in 2004/5 and 2005/6

Variables	Observations	Mean	SD	Min	Max
Income per head	64999	16.59	26.50	0.93	470.22
Average age	64999	44.85	5.94	20	67
Prop of female GPs	64999	0.42	0.25	0	1
Number of performers	64999	4.45	2.51	1	15
Prop of qualified and non UK born GPs	64999	0.09	0.21	0	1
Ex-Fund Holders	64999	0.47	0.50	0	1
Herfindahl Index	64999	0.34	0.24	0.08	1.00
GP1000	64999	0.91	0.53	0.26	13.79
Training 2004	64999	0.29	0.45	0	1
Dispensing 2004	64999	0.09	0.29	0	1
SIR	64999	101.55	23.94	52.14	172.17
SIMD	64999	23.30	12.21	3.77	65.44
Rural	64999	0.18	0.39	0	1
Megrate	64999	0.02	0.03	0	0.41
prop75plus	64999	0.07	0.02	0	0.16
prop15less	64999	0.16	0.03	0.02	0.27
Prac population density	64999	3.16	10.56	0.01	136.38

6. Methods

Logit method is used to model monetary motivation. Negative binominal model is used to model overall motivation. And non-monetary motivation is modelled by Zero Inflated Negative Binominal (ZINB) method. All the three models are estimated with observations clustered by practice and indicator. The software used is STATA10.

6.1 Monetary motivation

Logit model is used because dependent variable has binary outcomes:

$$\Pr(y_{igt1}=1 \mid x_{igt}) = f_{\text{logit}}(\alpha + I_{igt}\beta + \gamma_g^{'}\chi + k_g^{'}\delta + T\phi + \varepsilon_{igt})$$

y_{igt1} is dummy variable about whether performance index for practice g for indicator i in year t got the maximum payment threshold. x_{igt} is the matrix which include all the independent variables for observation y_{igt}. I_{igt} is the income from every patient treated with payment. $\boldsymbol{\gamma}_{\mathbf{g}}^{'}$ is the matrix which contain all the practice characteristics variables. $\boldsymbol{\kappa}_{\mathbf{g}}^{'}$ is the matrix which contain all the patients characteristics variables. T is the time dummy. ε_{igt} is the error term for all the unobserved heterogeneity.

6.2 Overall motivation

Negative binominal model is used because dependant variable takes count data. Poisson model usually be the starting point for count data analysis. It based on the assumption that observations should have equal dispersion (mean value equals to variance). However, in practice data usually have over dispersion (variance is greater than mean) as result of unobserved heterogeneity from observations. Negative binominal regression model is the most frequently cited alternative for Poisson regression model. Results in the following section show that negative binominal regression model is preferred for the data we have. Model specification takes:

$$\Pr(y_{igt2}=k \mid x_{igt}) = \frac{\Gamma(\theta+k)}{\Gamma(\theta)\Gamma(k+1)}\left(\frac{\theta}{\theta+\lambda}\right)^{\theta}\left(\frac{\lambda}{\theta+\lambda}\right)^{k}$$

where $E(y_{igt2}) = \lambda = \exp(\alpha + I_{igt}\beta + \gamma_g^{'}\chi + k_g^{'}\delta + T\phi + \varepsilon_{igt})$

and $Var(y_{igt2}) = \lambda + \theta^{-1}\lambda^2$

y_{igt2} is the number of patients treated in practice g for indicator i at year t. Rest of the variables have the same explanations like in the monetary motivation model.

6.3 Non-monetary motivation

We give zero value to observations which did not reach the maximum payment threshold. It made the dataset have more zero observations than if the observations come from negative binominal distribution. The excess zero observation problem could be solved by using zero inflated negative binominal regression if negative binominal model is preferred than Poisson model by test (Winkelmann, 2003). STATA will use a logit model first to predict the probability of having given-zero observations. It is followed by a negative binominal regression for observations which come from negative binomial distribution. The independent variables for logit model are the same like in the monetary motivation model. It is explained in the theoretical part that whether GPs could get the maximum payment threshold is determined by the marginal material cost and marginal income. The model is a two-part model with probability of belong to the inflated zero observations specified by a logit link function, and all other observations specified by a negative binomial function. The dependent is y_{igt3}, which is the number of patients freely treated for indicator i by practice g in year t. Rest of the variables have the same explanations like in the monetary motivation model.

$$E(y_{igt3}) = \lambda = \exp(\alpha + I_{igt}\beta + \gamma_g^{'}\chi + k_g^{'}\delta + T\phi + \varepsilon_{igt})$$

For observations under the same practice or same indictor, error terms are possibly correlated with each other. Again it is because the unobserved independent variables are included in error term. In regression, practice and indictor are clustered to allow error terms within the same practice and indicator correlated. However, correlation among repeated groups (defined as practice and indicator) may happen. For example, error terms may correlate with each other for observations which belong to the same practice but not the same indicator or belong to the same indicator but not the same practice. This problem may bias the result. Also, error term and independent variables may correlate with each other because error term includes all the unobserved variables. For example, unobserved satisfaction utility may correlate with GPs' characteristics. This problem may also bias the result.

7. Results

7.1 Monetary motivation model (table 2)

The relationship between income per head and the possibility of getting the maximum payment threshold is significantly positive like expected. One pound income increase from

every patient treated with payment increased the probability of getting maximum payment threshold by 0.26%.

Table 2: Determinants of monetary motivation model for 2004/5 and 2005/6

Dependants variable	Probability of getting the maximum payment threshold	
	Odd Ratio	Z value
Income per head	1.0026	4.38**
Average age	1.0003	0.13
Prop of female GPs	1.0890	1.45
Number of performers	0.9727	-3.84**
Prop of qualified and non-UK-born GPs	0.5950	-8.17**
Ex-Fund Holders	1.1521	5.22**
Herfindahl Index	0.6384	-5.79**
GP1000	1.1666	4.30**
Training 2004	1.1786	4.40**
Dispensing 2004	0.7025	-6.62**
SIR	0.9941	-8.20**
SIMD	1.0048	3.90**
Rural	1.0955	1.78
Megrate	16.1755	3.48**
Prop75plus	0.7370	-0.45
Prop15less	18.3580	5.45**
Practice pop density	0.9928	-4.71**
Year dummy	3.0459	54.75**

1. * significant at 5%; ** significant at 1%
2. Standard Error adjusted for 33212 clusters
3. The Wald chi2 (18) takes value of 3573.66 with Probability>chi2 = 0.0000; the log pseudo likelihood takes value of -25812.942; Pseudo R2 takes value of 0.0545

7.2 Overall motivation (table 3)

The coefficient for income per head tells that one pound increase for every patient treated with payment reduced the overall output by $0.22\%\,(1-e^{\beta}=1-e^{-0.0022}=1-99.78\%=0.22\%)$. The same calculation method applies to the rest of the discussions.

Table 3: Determinants of overall motivation model for 2004/5 and 2005/6

Dependants variable	Overall number of patients treated	
	Coef	Z value
Income per head	-0.0022	-26.75**
Average age	-0.0002	-1.32
Prop of female GPs	0.0079	2.45*
Number of performers	-0.0026	-6.58**
Prop of qualified and non-UK-born GPs	-0.0263	-6.02**
Ex-Fund Holders	0.0057	3.92**
Herfindahl Index	-0.0211	-4.76**
GP1000	0.010	4.37**
Training 2004	-0.0044	2.29*
Dispensing 2004	-0.0136	-3.85**
SIR	-0.0003	-8.16**
SIMD	0.0003	5.11**
Rural	-0.0001	-0.05
Megrate	0.1576	5.10**
Prop75plus	-0.0868	-2.35*
Prop15less	0.0785	2.85**
Practice pop density	-0.0005	-3.82**
Year dummy	0.0583	78.75**
Constant	-0.0847	-7.26**

1. * significant at 5%; ** significant at 1%
2. Standard Error adjusted for 33212 clusters
3. Exposure takes the number of eligible patients
4. The Wald chi2 (18) takes value of 6501.14 with Probability>chi2 = 0.0000. Also the log pseudo likelihood takes value of -256631.49
5. Alpha has coefficient of 0.0101531 with robust standard error 0.0002791. 95% confidential level is (0.0096204, 0.0107152)

7.4 Non-monetary motivation model (table 4)

STATA gives the result of likelihood ratio test. It tests the equal dispersion assumption from Poisson distribution against the over dispersion assumption from negative binominal distribution. The result rejects null hypothesis of being equal dispersion. STATA gives the result of Vuong test for zero inflated negative binominal against standard negative binomial. The result shows that zero inflated negative binominal is preferred.

Table 4: Determinants of non-monetary motivation model for 2004/5 and 2005/6

Dependants variable	Patients freely treated		Inflated zero value of freely	
	Coef	Z value	Coef	Z value
Income per head	-0.0048	-18.40**	-0.0087	-8.88**
Average age	-0.0006	-1.03	0.0006	0.17
Prop of female GPs	0.0100	0.85	-0.0735	-1.08
Number of performers	-0.0183	-13.06**	0.0597	7.59**
Prop of qualified and non-UK-born GPs	-0.0372	-2.58*	0.5807	7.85**
Ex-Fund Holders	0.0227	4.35**	-0.1424	-4.76**
Herfindahl	-0.0872	-5.46**	0.3673	4.31**
GP1000	0.0522	6.26**	-0.3330	-6.50**
Training 2004	-0.0006	-0.08	-0.1455	-3.55**
Dispensing 2004	-0.0180	1.46	0.2964	4.76**
SIR	-0.0007	-5.45**	0.0069	8.62**
SIMD	0.0015	6.59**	-0.0050	-3.62**
Rural	0.0019	0.18	0.0041	0.07
Megrate	0.6601	6.62**	-3.0618	-3.26**
Prop75plus	-0.4949	-3.63**	1.1959	1.57
Prop15less	0.3333	3.30**	-2.7010	-4.43**
Practice pop density	-0.0011	-2.59*	0.0097	4.63**
Year dummy	0.1115	30.48**	-1.1259	-47.65**
Constant	-0.4073	-9.67**	-1.4921	-6.09**

1. * significant at 5%; ** significant at 1%
2. Standard Error adjusted for 33212 clusters
3. Exposure takes the number of patients can be freely treated
4. The non zero observations is 52711 and the zero observations is 12288
5. The Wald chi2 (18) takes value of 1510.84 with Probability>chi2 = 0.0000. Also the log pseudo likelihood takes value of -182005.3
6. Likelihood-ratio test of alpha=0: chibar2(01) = 8.3e+04 Pr>=chibar2 = 0.0000
7. Vuong test of zinb vs. standard negative binomial: z = 69.46 Pr>z = 0.0000

The results on the left panel are the negative binomial portion of the ZINB estimation and those on the right panel are the zero inflation estimation. The two portions of the empirical model are estimated jointly. A positive coefficient on the left panel shows that the independent variable tends to have a positive influence on the number of patients treated freely, and the positive coefficient on the right panel shows that the independent variable tends increase the probability of having no patients treated for free. Income per head has a negative coefficient of -0.0048. This result suggests that one pound payment increase for every patient treated by GP is related with a reduction in free service provided of 0.48%.

One more GP a practice has lead to 1.81% free service dropped. One percent increase in the percentage of qualified non UK born GPs made 0.0365% free service decrease. Practices with Ex-Fund Holder status provide 2.3% more free service on average. Practice herfindahl index has negative coefficient valued -0.0872. It suggests practices with high competition

with others provided more free service. One more GP for every thousand patients lead to the provision of free service to increase by 5.36%. Coefficient of SIR tells that one percent more patients self assessed as having limiting long-term illness made the amount of free service provided by GP to drop by 0.07%. SIMD score has a significant positive coefficient of 0.0015. That means GPs who work with patients with high deprivation status provided more free service. Proportion of patients with ethnic minority status has positive relation with GPs' non-monetary motivation. One percentage more patients belong to ethnic minority made the free service increased by 0.93%. One percentage more patients belong to age group above 75 years old made the amount of free service dropped by 0.39%. In contract one percentage more patients belong to age group below 15 years old, 0.40% more free service provided. Also population density has significant negative coefficient. One unit increase in the hectares per person decreased the free service by 0.11%. The significant positive coefficient from year dummy tells that GPs have 11.80% free service increase in 2005/6.

8. Discussions

The main results from the empirical tests are consistent with the hypothesis developed in the theoretical work in section 4.2. An increase in forgone income, as measured by an increase in the amount paid per patient treated is shown to have had a negative effect on GPs' non-monetary motivation. In contrast, an increase in the financial incentive associated to the value of pounds per point in QOF has a positive effect on GPs' monetary motivation. The overall model provides evidence that an increase in the level of payment per patient had a significant and negative effect on GPs' overall motivation as measured by the total number of patients treated. If it is correct than perhaps we can say that the negative effect of the non-monetary motivation has outweighed the positive effects of the monetary motivation. This has important implications for policy as using strong financial incentives to further motivate well-motivated professionals may not be an efficient way of influencing behaviour and performance.

When GPs get one pound more from patient treated with payment, the non-monetary output will decreased to 99.52%. The result tells that cross sectional, GPs are willing to provide a larger volume of free treatments if the forgone income is low. Furthermore, increasing the income from every patient treated with payment over time will decrease the free services. One pound increase from every patient treated with payment made the probability of getting the maximum payment threshold increased 0.26%. This result is consistent with the existing

literature that financial incentives have positive effects on individuals' monetary motivation. In overall, the effect of increasing monetary gain from treat one patient by one pound has a negative effect on output. It made the overall motivation dropped to 99.78%. This result is co-determined by the effects of financial incentive on monetary motivation and non-monetary motivation. The result suggests that the negative effect from decreased non-motivation overcomes the positive effect on monetary motivation.

The potential problem for the pay for performance policy from the result we have got above is that increasing payment per patient may be associated with GPs working to converge towards the maximum payment thresholds. Those practices with below maximum payment threshold performance have a greater incentive to increase activity levels. However, those above the maximum threshold perform to reduce the number of patients they treat because of the increase in forgone income. This interpretation of the results is dependent on the assumption that GP satisfaction utility remains constant over time. Future research will examine evidence of such a convergence.

All three models have significant positive year dummies. That means GPs are more motivated for all types of motivations after policy adjustments. It also suggests that for non-monetary motivation and overall motivations, there should be other arguments which positively respond to the financial incentives. Because as the result explained in above, financial incentive itself has negative relations with these two types of motivations.

Whether GPs' QOF income from every patient treated can motivate their monetary motivation depends on the relationship between income and cost of providing services. The empirical results suggests that the more GPs there are in a practice the less they are responded to financial incentives. It can be explain that the marginal income from providing an additional unit of services is lower for individual GP is in large practices as the QOF payment is awarded at practice level and is shared across all GPs.

The number of GPs in the practice is also negatively associated with non-monetary motivation. This could be explained by self-selection of GPs who are more monetary motivated or business-minded joining larger and more efficient practices. Also the less motivated result could be explained as free-ride effect. In larger practices non-altruistic

behaviours may be less easily observable and hence the peer pressure or professional norms may be less influential on GPs' behaviour.

The variables measuring the degree of competition between practices suggests that in areas characterised by higher (potential) competition, GPs are motivated to provide higher quality services to attract patients. Results from number of GPs for every thousand patients, population density and practice herfindahl index are good examples.

The motivations of GPs who work for patients with high deprivation status are higher on average. This may be explained by GPs' self selection: GPs who choose to work in areas which patients have high deprivation status may be more strongly non-monetary motivated. However, the results suggest that GPs in such areas are more highly motivated both in monetary and non-monetary ways.

From the theoretical analysis the material costs are important determinants for all types of GP motivations. The more expensive is the treatment, the less motivated are the GPs. In the absence of a direct measure of the cost we have used proxies. Costs of providing care have been proxied by using SIR and patients' age, whereby higher costs are associated with treating patient who are older and patients with a higher SIR. The empirical results provide evidence to support the idea that GPs have strong motivations toward relatively healthy or young patients but not the sick or the old. However, these proxies are unlikely to capture sufficiently the costs of providing care.

Non-monetary motivation has been included in the model only in the form of 'forgone income'. Satisfaction utility as a form of non-monetary motivation has not been explicitly included in the model. However, if there has been an increase in the satisfaction utility over the two year period, this may contribute to the positive sign on coefficient for the year dummy. For example, assume GPs in general recognise the effect of financial incentive on their satisfaction utility is positive. If the effect is big enough to offset the negative effect from the increased forgone income on non-monetary motivation, then non-monetary motivation can be positive related with financial incentive. As a result the year dummy should be positive because QOF income increased in 2005/6.

A possible weakness of the approach is that the provision of patient services without payment may not be the only altruistic activities taken by GPs. We may be under-estimating the extent of intrinsic motivation and therefore, misestimating the effect of increased financial incentives on non-monetary motivation. If financial incentive undermines the motivation for GPs to provide free service, it may discourage other types of altruistic behaviours as well. Another possible weakness is on the measurement of forgone income. It is arguable that there are alternative methods to calculate forgone income. For example, it can be measured by the forgone income from indicators which available points have not been fully achieved.

9. Conclusion

The new GMS contract was introduced in 2004 to drive up health service standard and making resource allocation be more efficient in primary care. If driving up standard is only judged by whether GPs achieve the upper thresholds they set for the clinical indicators, then the government has succeed. However one should also take the effect of income changes on GPs' non-monetary motivation into account. Results from the empirical section show that one pound increase in payment for every patient treated decreases the overall output by 0.22% in 2004/5 and 2005/6.

This article demonstrates both in theory and empirical analysis why GPs' non-monetary utility should not be ignored. In terms of the non-monetary utility, forgone income utility is important. The importance of satisfaction utility has been showed in theoretical section and also been suggested by the empirical result. We demonstrated that it is important for policy makers to understand GPs' motivation structure so that the chosen incentives would match with motivations of GPs.

10. References

1. Avinash Dixit. Incentives and Organisations in the Public Sector: An Interpretation Review. Presented at the National Academy of Sciences conference on Devising Incentives to Promote Human Capital, Irvine, CA, December 17-18, 1999. The first version – 16th November 1999 and this version - 10th May 2000.

2. Anthony Scott. Chapter 22: Economics of General Practice. Handbook of Health Economics. Culyer and Newhouse (Eds). Elsevier, Amsterdam, 2000.

3. Brook R.H, McGlynn E.A, Cleary P. Quality of health care. Part 2: measuring quality of care. New England Journal of Medicine. 335: 966-70, 1996.

4. Edward L. Deci. Effect of externally mediated rewards on intrinsic motivation. Journal of Personality and Social Psychology. 18: 105-115, 1971

5. Edward L. Deci. Intrinsic motivation. Plenum Press, New York, 1975.

6. Edward L.Deci, Koestner Richard, Richard M.Ryan. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. Psychological Bulletin. 125(6): 627-668, Nov 1999.

7. Frey B.S. Not Just For the Money: An Economic Theory of Personal Motivation. Edward Elgar: Cheletenham, 1997.

8. Frey B.S and Felix Oberholzer-Gee. The cost of price incentives: an empirical analysis of motivation crowding out. The American Economic Review. 87(4): 746-755, Sep 1997.

9. Gary S.Becker. A theory of the allocation of time. Economic Journal. 75: 493-517, 1965.

10. Hugh Gravelle, Matt Sutton, Ada Ma. Doctor Behaviour Under a Pay for Performance Contract: Further Evidence from the Quality and Outcomes Framework. CHE Research Paper 34, Feb 2008.

11. Rainer Winkelmann. Econometric analysis of count data (fourth edition). Spring-Verlag Berlin.Heidelberg, 2003.

12. Review Body on Doctors' and Dentists' Remuneration. Thirty-Six Report 2007, March 2007.

13. Richard Titmuss. The Gift Relationship. Allen and Unwin, London, 1970.

14. Roland M. Linking Physicians' Pay to the Quality of Care - A Major Experiment in the United Kingdom. New England Journal of Medicine. 315: 1448-1454, 2004.

15. Sanford E.DeVoe and Jeffrey Pfeffer. Hourly payment and volunteering: the effect of organizational practices on decisions about time use. Academy of Management Journal. 50:783-798, 2007.

11. Appendix

Table A1: Indicators allocation, maximum and minimum payment thresholds, points allocation and GPs' performance - Example from CHD in 2004/5

Indicator name	Minimum payment threshold	Maximum payment threshold	Full available points	Mean of achieved points (Variance)	Mean value for number of patients freely treated (variance)	Mean value for number of patients treated (variance)
Secondary Prevention in Coronary Heart Disease (CHD)						
CHD3	25%	90%	7	6.91 (0.49)	14.56 (12.03)	232.02 (145.96)
CHD4	25%	70%	4	3.84 (0.45)	11.87 (9.56)	50.27 (35.56)
CHD5	25%	90%	7	6.92 (0.35)	14.32 (12.00)	231.09 (145.23)
CHD6	25%	70%	19	18.84 (0.91)	36.38 (28.57)	201.70 (128.21)
CHD7	25%	90%	7	6.58 (0.86)	6.50 (8.51)	212.16 (134.44)
CHD8	25%	60%	16	15.17 (2.35)	26.87 (28.41)	159.45 (104.36)
CHD9	25%	90%	7	6.74 (0.46)	4.87 (7.15)	213.98 (135.18)

Table A2: Indicators allocation, maximum and minimum payment thresholds, points allocation and GPs' performance - Example from CHD in 2005/6

Indicator name	Minimum payment threshold	Maximum payment threshold	Full available points	Mean of achieved points (Variance)	Mean value for number of patients freely treated (variance)	Mean value for number of patients treated (variance)
Secondary prevention in Coronary Heart Disease (CHD)						
CHD3	25%	90%	7	6.99 (0.14)	17.76 (12.02)	236.15 (146.51)
CHD4	25%	70%	4	3.97 (0.19)	3.07 (2.83)	53.92 (36.94)
CHD5	25%	90%	7	6.99 (0.11)	17.93 (12.04)	235.44 (146.24)
CHD6	25%	70%	19	18.98 (0.42)	44.79 (29.02)	209.74 (130.23)
CHD7	25%	90%	7	6.92 (0.36)	11.15 (9.65)	223.04 (138.74)
CHD8	25%	60%	16	15.88 (0.83)	44.87 (33.19)	177.37 (112.58)
CHD9	25%	90%	7	6.98 (0.20)	12.47 (10.55)	227.48 (141.74)

Table A3: Eligible Indicators included in empirical analysis for year 2004/5 and 2005/6

Clinical Disease Domains	Eligible Indicators
Secondary Prevention in Coronary Heart Disease (CHD)	CHD3, CHD4, CHD5, CHD6, CHD7, CHD8, CH9
Stroke and Transient Inschaemic Attacks (STROKE)	STROKE3, STROKE4, STROKE5, STROKE6, STROKE7, STROKE8
Diabetes Mellitus (DM)	DM2, DM3, DM4, DM5, DM6, DM7, DM8, DM9, DM10, DM11, DM13, DM14, DM16, DM17
Chronic Obstructive Pulmonary Disease (COPD)	COPD4, COPD5
Epilepsy	EPILEPSY 2, EPILEPSY3, EPILEPSY4
Hypothyroid (THYROI)	THYROI2
Mental Health (MH)	MH2, MH4
Asthma	ASTHMA3, ASTHMA4, ASTHMA5, ASTHMA6