Social Science & Medicine 235 (2019) 112343

Contents lists available at ScienceDirect

Social Science & Medicine

journal homepage: www.elsevier.com/locate/socscimed

Details matter: Physician responses to multiple payments for the same activity

Yan Feng[a,*], Hugh Gravelle[b]

[a] Centre for Primary Care and Public Health, Queen Mary University of London, United Kingdom
[b] Centre for Health Economics, University of York, United Kingdom

ARTICLE INFO

Keywords:
United Kingdom
Pay for performance
Physicians
Incentives
Contract design

ABSTRACT

The UK Quality and Outcomes Framework rewards general practices for achieving quality indicators for chronic disease management. Some indicators are multi-rewarded. For example, there are indicators for controlling blood pressure for patients with diabetes and for patients with chronic heart disease. Thus if a patient has diabetes and heart disease the practice is rewarded twice for controlling her blood pressure. Other indicators are singly rewarded: the incentivised activity is only for patients with single specific condition. We compare general practice performance on single and multi-reward indicators. We use a 2005/6–2012/13 panel of over 800 Scottish general practices, control for practice characteristics, practice fixed effects, indicator characteristics (whether the indicator was for measurement, treatment, or intermediate outcome, maximum payment, upper thresholds), condition, and year and cluster on indicators. We find that the proportion of patients with a given condition for whom a quality indicator was achieved was higher, and the proportion who were exception reported was lower, for multi-reward indicators than for single reward indicators. We also exploit the replacement of multi-reward smoking indicators by single reward indicators in 2006/7. Compared to indicators which were always single or always multi-reward, the proportion of the relevant patients for whom the smoking indicators were achieved fell when the smoking indicators were no longer multi-reward. Fine details of pay for performance schemes matter: they affect physician behaviour and patient outcomes.

1. Introduction

Pay for performance (P4P) schemes linking health care provider financial rewards to quality have been introduced in many healthcare systems. These include Australia (Scott, 2007), Canada (Pink et al., 2006), Germany (Greb et al., 2006), the Netherlands (Custers et al., 2006), New Zealand (Perkins and Seddon, 2006), Spain (Gené-Badia et al., 2007), the UK (Meacock et al., 2014), and the US (Rosenthal et al., 2006). They have had mixed success (Guthrie and Tang, 2016; Gillam et al., 2012; Scott et al., 2011; Van Herck et al., 2010; Greene and Nash, 2009; Christianson et al., 2007; Petersen et al., 2006; Town et al., 2005; Armour et al., 2001). Both standard and behavioural economic models suggest that the design of P4P schemes matters in the sense that it will change the behaviour of those incentivised, not just the rewards they receive (Eijkenaar et al., 2013; Kristensen et al., 2013; Mehrotra et al., 2010). Thus the success of the Advancing Quality scheme in England (Sutton et al., 2012) but its failure in the US (Ryan, 2009) has been attributed to differences in coverage and the magnitude of rewards on offer (Epstein, 2012). In this paper we investigate if implicit incentives created by rather more subtle features of a major UK P4P scheme affected the behaviour of health care providers and the outcome for patients.

We use data from the UK Quality and Outcomes Framework (QOF) to investigate whether and how General Practitioners (GPs) in Scotland responded to the implicit financial incentives created by the fine details of the scheme. The QOF rewards general practices for achievements on a large number of quality indicators for management of chronic conditions and creates complex multi-tasking incentives (Holmstrom and Milgrom, 1991). Some groups of indicators produce multiple rewards for the same activity. For example, there are indicators for controlling blood pressure for patients with diabetes and for patients with chronic heart disease (CHD). Thus if a patient has both conditions the practice is rewarded twice for controlling the patient's blood pressure. Other indicators are single reward: the incentivised activity only counts towards for one indicator. For example, the indicator for retinal screening applies only to diabetic patients. We compare general practice performance and quality for single reward and multi-reward indicators and find that, controlling for other features of the indicators and for practice

* Corresponding author. Yvonne Carter Building, 58 Turner Street, Whitechapel, London, E1 2AB, United Kingdom.
E-mail addresses: yan.feng@qmul.ac.uk (Y. Feng), hugh.gravelle@york.ac.uk (H. Gravelle).

https://doi.org/10.1016/j.socscimed.2019.05.048
Received 22 August 2018; Received in revised form 25 May 2019; Accepted 29 May 2019
Available online 03 June 2019

characteristics, practices have greater achievement on multi-reward indicators. We also exploit the replacement of six separate multi-reward smoking indicators for patients with six conditions with a single reward indicator for patients with any of the six conditions. We compare the changes in practice achievement for these indicators with changes for indicators which were always single or always multi-reward and find that the change from multiple to single rewards for an activity worsened practice performance.

Our paper contributes to the empirical literature on how physician agents, who are likely to be both intrinsically and extrinsically motivated (McGuire, 2000; Rebitzer and Taylor, 2011), behave under multitask P4P schemes. Our results reinforce the argument that, whether we view the design of incentive schemes as engineering (Roth, 2002) or plumbing (Duflo, 2017), fine details matter: they affect the behaviour of incentivised providers and hence outcomes for patients.

2. Background

2.1. NHS primary care

Patients face no charges in the National Health Service (NHS) which is funded almost entirely out of general taxation. Patients register with a general practice which also acts as the gatekeeper for non-emergency hospital care. On average, general practices have around 5800 patients and 5 GPs and most are partnerships owned by their GPs.

Around 82% of Scottish practices have a General Medical Services (GMS) contract with the NHS under which they are paid by a mixture of capitation, quality incentives, and items of service. Capitation accounts for about 60% of practice revenue and varies with the demographic and morbidity mix of the practice (ISD Scotland, 2016a). Around 11% of revenue is from the QOF and increases with practice list size for given quality. Practice payments for providing specific services including vaccinating and screening target proportions of the relevant practice population also increase with list size. Practices are reimbursed for the costs of their premises but have to fund all other expenses, such as hiring practice nurses and clerical staff, from their revenue. The other 18% of practices have contracts which are similar to the GMS contract but they are also rewarded for providing additional services for specific patient groups.

2.2. Quality and Outcomes Framework

The QOF was introduced in April 2004. Although participation was voluntary almost all general practices in Scotland took part. The QOF linked payment to indicators for activities intended to improve the management of patients with chronic conditions. We focus on the clinical quality indicators which were measured as the ratio of patients for whom an indicator was achieved to the number of patients declared eligible for the indicator. For example, indicator CHD6 was the proportion of eligible patients with Chronic Heart Disease (CHD) whose blood pressure was 150/90 or less. Practices were awarded points which increased linearly with achievement between a lower threshold (40% for most indicators from 2006/7 onwards) and an upper threshold (varying between 50% and 90% across indicators and years). The upper thresholds for the clinical indicators were set at less than 100% to allow for patients for whom the indicator could not be achieved but for whom there was no appropriate exception reporting code (Coulson, 2011). Practices were paid a price per point (on average around £125) which varied with the practice list size and the number of patients with the relevant condition. Initially (2005/6 in our data) there were 65 ratio clinical indicators (out of a total of 146 indicators) covering 11 conditions and accounting for 47% of the total points available. There were subsequent changes to the number of indictors, indicator definitions, the conditions covered, upper and lower thresholds, and the maximum points available for an indicator. In 2012/13, there were 76 ratio clinical indicators for 22 conditions potentially accounting for 59% of total points.

Some ratio clinical indicators are for activities which are specific to patients with a particular condition. For example, in 2011/12 the indicator DM21 was the percentage of patients with diabetes who had a record of retinal screening in the previous 15 months. Similarly, CHD10 was the percentage of patients with CHD being treated with a beta blocker. With these *single reward* indicators the incentivised activity (record of retinal screening, prescribing beta blocker) counted for achievement only on one indicator and was thus rewarded only once even for a multi-morbid patient with both diabetes and CHD.

Other activities count for more than one indicator. For example in 2010/11, there were indicators for the percentage of patients with coronary heart disease (CHD8) and the percentage of patients with diabetes (DM17) whose total cholesterol was 5 mmol/l or less. CHD8 and DM17 are examples of *multi-reward* indicators where an activity (achieving a cholesterol target) for a multi-morbid patient with both CHD and diabetes counts for achievement on more than one indicator and thus is rewarded more than once.

3. Methods

3.1. Incentives under the QOF

The number of points, and hence revenue, earned on a ratio clinical indicator is determined by reported achievement $N/(P-E)$, where N is the number of patients for whom the indicator is achieved, P the number with the relevant condition, and E the number with the condition who are exception reported by the practice. P–E is the number declared eligible by the practice for the indicator. Patients can be exception reported on clinical grounds (for example they are allergic to the treatment, or terminally ill, or very frail) or because they have refused to attend, or because the relevant investigative or secondary care service is not available. Thus a practice can increase reported achievement by increasing N and E.

We investigate whether an indicator being single or multi-reward affects practice behaviour as measured by *population achievement* (N/P), *exception reporting* (E/P), and *reported achievement* ($N/(P-E)$). We interpret population achievement (N/P) as a measure of quality of care because, unlike reported achievement, it is not affected by exception reporting (Doran et al., 2006).

If an activity, say recording blood pressure, is incentivised by indicators for more than one condition (say CHD and diabetes), then a practice will earn more revenue from recording blood pressure for a patient with both CHD and diabetes than for a patient with only CHD or only diabetes. Other things equal, this incentive will imply that a higher proportion of multimorbid patients will have blood pressure recorded than patients with only CHD or diabetes. This in turn will imply, other things equal, that population achievement (N/P) for the multi-reward indicators for recording blood pressure for patients with CHD and for patients with diabetes will be greater than population achievement for single reward indicators.

However, not all other things are equal when comparing achievement on different indicators. Marginal revenue from increasing achievement on an indicator depends on the maximum number of points available for the indicator, the upper threshold beyond which greater achievement has no effect on points and revenue, and the number of practice patients with the relevant condition. The marginal cost of achieving an indicator may differ across indicators because they apply to patients with different conditions and different levels of multimorbidity. The cost of achieving an indicator may vary with patient age and socio-economic characteristics. Costs will vary with the type of incentivised activity. For example, recording blood pressure is easier than controlling it. And finally there may also be differences in cost and revenue functions and GP preferences across practices.

In our empirical modelling we control for indicator points and thresholds, as well as practice, and patient mix characteristics and so

we expect that population achievement N/P will be greater for multi reward indicators than for single reward indicators.

Exception reporting more patients for an indicator will also increase reported achievement $N/(P-E)$ and so the marginal revenue from gaming exception reporting will be greater for multi reward indicators than for single reward indicators. But GPs are intrinsically as well as extrinsically motivated (McGuire, 2000; Bénabou and Tirole, 2006). Increasing population achievement (N/P) on a quality indicator benefits patients whereas gaming exception reporting and thereby increasing E/P does not and offends against professional norms. Thus E/P may not be higher for multi-reward indicators. Indeed, if the incentives from multiple rewards lead the practice to choose a high enough N to take reported achievement ($N/(P–E)$) above the upper threshold for the indicator, then the practice could reduce gaming of E without financial penalty. Being multi rewarded could increase or reduce or have no effect on E/P.

3.2. Estimation: panel models

Using data from eight years (2005/6–2012/13) we estimate models to examine if practice performance on an indicator depends on whether the indicator activity is multi or single reward:

$$y_{igt} = \beta_0 + \beta^M M_{it} + \mathbf{x}_{it}^{\prime Q}\boldsymbol{\beta}^Q + \sum_{c=1}^{22} D_i^c r_{cgt}\beta_{rc} + \mathbf{x}_{gt}^{\prime p}\beta^p + \mathbf{D}_t^{\prime T}\beta^T + \alpha_g + \varepsilon_{igt} \quad (1)$$

We use three measures of performance (y_{igt}): *population achievement* N_{igt}/P_{igt}, *exception reporting* E_{igt}/P_{igt}, and *reported achievement* $N_{igt}/(P_{igt}-E_{igt})$.

M_{it} is the extent to which the indicator is multi-reward. We use three measures of multi-reward, each entered in separate regressions. The simplest is a 1/0 dummy for the indicator being multi-reward. The other two allow for the possibility that the effect of being multi-reward depends on the number of other indicators (0–5) which can be rewarded for the same activity to the same patient. We include this as a continuous variable and, more flexibly, as a set of five indicator for 1, 2, …, 5 other indicators rewarded for the same activity.

$\mathbf{x}_{it}^Q$ is a vector of characteristics of indicator i in year t (which might affect the revenue from an indicator, or the effort required to achieve it, or the gain to patients from achieving the indicator): the maximum number of points for the indicator; the upper threshold; dummies for indicator type (treatment indicators, such as prescribing beta blockers for CHD patients, intermediate outcome indicators, such as blood pressure 140/80 or less for diabetes patients, versus the baseline category of record indicator, such as measuring blood pressure); and 21 dummies for the condition targeted by the indicator i (diabetes is the baseline condition). We also include the prevalence rate r_{cgt} (percentage of patients in practice g at time t with condition c) for the condition for which the indicator is a quality measure. This allows for the number of patients with a condition to affect costs and for the fact that, until 2009/10 the marginal financial reward for achieving an indicator varied with the number of patients with the relevant condition (Guthrie et al., 2006; Gravelle et al., 2010).

$\mathbf{x}_{gt}^p$ is a vector of practice characteristics including numbers of GPs, proportion of female GPs, list size, patient age and gender mix, dummies for rurality, and whether the practice can dispense as well as prescribe medicines. It also includes the percentage of the practice's patients who live in datazones (small areas with 500–1000) which are in the top 15% of the distribution of the Scottish Index of Multiple Deprivation (Scottish Government, 2018). $\mathbf{D}_t^T$ is a vector of year dummies, α_g is a practice fixed effect, and ε_{igt} is a zero mean error term.

Whether an indicator is a multi-reward or single-reward is a characteristic of the indicator, not the practice. Since different practices' performances on a given indicator may be correlated because of unobserved indicator characteristics which affect the ease or difficulty of the incentivised activity we cluster the errors on the indicator (Moulton, 1986).

We believe that the specification (1) enables us to identify the effect of an indicator being multi or single rewarded because it allows for characteristics of indicators (maximum points, upper threshold, whether they are for recording, treatment or intermediate outcomes, the type of condition, and the prevalence of the condition) which may affect performance and which might differ systematically between multi and single reward indicators. In particular, the two specifications with the number of other indicators rewarded for the same activity as the multi-reward measure are persuasive tests of the effect of multi versus single reward.

3.3. Estimation: difference in differences models

The panel pooled model uses a rich set of covariates and a large number of indicators over eight years. But the specification cannot allow for unobserved characteristics of indicators, such as GPs' perceptions of their effects on patient health, which might affect performance and might differ between multi and single reward indicators. We can however exploit the replacement of two sets of multi-reward smoking indicators with two single reward indicators for the same activities in 2006/7. We estimate difference in differences (DID) models of the effect of an activity switching from being multi to single reward. As controls we use indicators of the same activity type (recording, treatment) which were always multi-reward or always single reward before and after the change to the smoking indicators.

In 2005/6 there were six multi-reward indicators for recording smoking status for patients with asthma (aged 20 or more), high blood pressure, CHD, COPD, diabetes, and stroke. In 2006/7 and 2007/8 these six indicators were replaced by a single reward indicator (SMOKING1) for recording the smoking status of a patient with any, or any combination of the six conditions. Similarly, six multi-reward indicators for offering advice on stopping smoking to smokers with the six conditions were replaced by a single reward indicator (SMOKING2) for smoking cessation advice for patients who smoke and have one or more of the six conditions. We cannot extend the DID analysis to later years because in 2008/9 two further conditions were added to the six covered by the single reward smoking indicators and there was no corresponding single reward indicator in previous years for the two additional conditions. Appendix A1 has further details on the smoking indicators.

We measure practice population achievement in recording smoking in 2005/6 as the total number over the six indicators for whom smoking status was recorded divided by the total number with the six conditions. For reported achievement the denominator is the total number with the conditions minus the total exception reported. For exception reporting the numerator is the total exception reported over the six conditions. For 2006/7 and 2007/8 we measure the recording of smoking status for these patients as the achievement on the replacement single reward record indicator SMOKING1. Thus we regard recording of smoking status as one indicator for which multiple rewards were removed in 2006/7–2007/8. We proceed similarly for the smoking cessation advice indicators.

To test for the effect of replacing multi-reward smoking indicators with a single reward indicator we estimate

$$y_{igt} = \beta_0 + \beta_1 D_t^1 + \beta_2 D_t^2 + \beta_3 S_i + \beta_4 A_i + \delta_1 S_i D_t^1 + \delta_2 S_i D_t^2 + \gamma_1 A_i D_t^1 + \gamma_2 A_i D_t^2 + \mathbf{x}_{it}^{\prime Q}\beta^Q + \mathbf{x}_{gt}^{\prime p}\beta^p + (1-S_i)\sum_c \beta^{rc} r_{cgt} D_i^c + S_i \sum_c \beta^{rcs} r_{cgt} + \alpha_g + \varepsilon_{igt} \quad (2)$$

where D_t^1, D_t^2 are dummy variables for 2006/7 and 2007/8. S_i is a dummy variable for the smoking indicator, and A_i is a dummy variable for indicators which were single reward in all three years 2005/6–2007/8. The baseline indicator type is indicators which were multi-reward in all three years.

Table 1
Summary statistics for multi and single reward indicators.

						Multi-reward indicators				
	Indicators	Conditions	Record	Treatment	Outcome	Upper threshold	Max points	Population achievement (N/P)	Exception rate (E/P)	Reported Achievement (N/(P–E))
2005/6	40	8	17	16	7	0.8088	8.08	0.8616	0.0660	0.9208
2006/7	28	9	10	10	8	0.8196	9.14	0.8472	0.0854	0.9257
2007/8	28	9	10	10	8	0.8196	9.14	0.8524	0.0776	0.9231
2008/9	29	9	10	11	8	0.8190	8.97	0.8549	0.0764	0.9249
2009/10	29	9	10	11	8	0.8190	8.90	0.8574	0.0736	0.9247
2010/11	29	9	10	11	8	0.8190	8.90	0.8522	0.0758	0.9213
2011/12	30	11	10	11	9	0.8027	8.67	0.8150	0.0997	0.9047
2012/13	34	12	10	13	11	0.8179	7.38	0.8122	0.1046	0.9061
All yrs	31	10	11	12	8	0.8153	8.59	0.8447	0.0818	0.9189
						Single reward indicators				
	Indicators	Conditions	Record	Treatment	Outcome	Upper threshold	Max points	Population achievement (N/P)	Exception rate (E/P)	Reported Achievement (N/(P–E))
2005/6	25	10	18	4	3	0.8300	6.76	0.8268	0.0840	0.9000
2006/7	34	16	25	6	3	0.8265	9.56	0.8182	0.0991	0.9060
2007/8	34	16	25	6	3	0.8265	9.56	0.8346	0.0914	0.9162
2008/9	33	15	25	5	3	0.8273	9.58	0.8385	0.0899	0.9192
2009/10	39	18	29	6	4	0.8128	9.36	0.8142	0.1028	0.9033
2010/11	39	18	29	6	4	0.8128	9.36	0.8143	0.1010	0.9024
2011/12	39	18	29	6	4	0.8051	8.38	0.7968	0.1126	0.8959
2012/13	42	19	30	8	4	0.8071	8.07	0.8031	0.1047	0.8950
All yrs	36	16	26	6	4	0.8174	8.88	0.8171	0.0991	0.9044

Notes. Indicators: number of indicators. Conditions: number of conditions covered. Record, Treatment, and Outcome are the numbers of indicators for recording patient details (e.g. blood pressure), for treating patients (e.g. prescribing statins), and achieving intermediate outcomes (e.g. blood pressure below specified limits). Upper threshold, Maximum points, Population achievement (N/P), Exception rate (E/P), Reported Achievement rate (N/(P–E)) are means over indicators, where N is number of patients for whom indicator achieved, P is the number with the relevant condition, E is the number of patients exception reported for the indicator. Statistics for all years are means over 2005/6 to 2012/13 over indicators.

The always single reward and always multi-reward control groups are sets of indicators for the same type of activity as the smoking indicator: recording indicators were the control group for SMOKING1 (recording smoking status) and treatment indicators were the control group for SMOKING2 (smoking cessation advice).

The time invariant unobserved idiosyncratic characteristics of indicators which affect outcomes via will be picked up in β_0 (for always multi-reward indicators), β_3 (for the smoking indicator) and β_4 (for always single reward indicators). x_{it}^{Q} is a vector of two indicator characteristics (maximum points, upper threshold). For the smoking record (cessation advice) indicator we use the total maximum points on the six smoking record (cessation advice) indicators and the national indicator prevalence weighted mean values of the upper thresholds of the six smoking record (cessation advice) indicators for 2005/6 and the values for SMOKING1 (SMOKING2) in 2006/7 and 2007/8. x_{gt}^{p} is a vector of practice characteristics. r_{cgt} is the practice condition prevalence rate for the condition covered by the indicator. The specification allows the effect of say diabetes prevalence to vary depending on whether the indicator is a control group indicator or a smoking indicator.

δ_1 is the 2006/7–2005/6 DID: the change in performance from 2005/6 to 2006/7 for the smoking indicator for recording (cessation advice) relative to the change in a control group of always multi-reward record (treatment) indicators. δ_2 is the 2007/8–2005/6 DID relative to always multi-reward indicators. $\delta_t - \gamma_t\,(t = 1,2)$ are the analogous DIDs for the smoking indicator relative to control groups of always single indicators.

The DID models identify the effect of a change from multi to single reward only if the smoking and control indicators would otherwise have had parallel trends in the absence of the change in QOF incentives. Comparison of the change in smoking and control indicators between 2004/5 and 2005/6 would not be useful because 2004/5 was the first year of the QOF when practices were still learning the system and the price per QOF point increased from £75 to £125 in 2005/6. Instead we estimate two DIDs: the first comparing the 2006/7–2005/6 change in the smoking indicator with the 2006/7–2005/6 change in the control indicators (δ_1 or $\delta_1 - \gamma_1$ depending on whether the controls are always multi reward or always single reward). The second DID compares the 2007/8–2005/6 change in the smoking indicator with the 2007/8–2005/6 change in the controls (δ_2 or $\delta_2 - \gamma_2$). If there was an underlying downward trend in the smoking indicators then the 2007/8–2005/6 DID will be more negative than the 2006/7–2005/6 DID: $\delta_2 - \delta_1 < 0$ or $(\delta_2 - \gamma_2) < (\delta_1 - \gamma_1)$. We test for an underlying trend by comparing these DIDs.

3.4. Data

We use practice level QOF data extracted by the Information Services Division (ISD) from Scottish general practice electronic patient records for eight financial years (2005/6 to 2012/13) and GP practice information from the ISD General Practitioner Contractor Database (ISD Scotland, 2016b). We do not use data for the first transitional year of the QOF in 2004/5. In 2004/5 practices may have been unfamiliar with the complexities of a new incentive scheme with 146 quality indicators. In 2004/5 the average price per QOF was £75 and was increased to £125 in subsequent years. The total number of points achieved by practices between 2004/5 and 2005/6 increased by 6%, the largest change by far between 2004/5 and 2016/17. There is also no information on exception reporting for 2004/5. There were other major changes to the GP contract in 2004/5 including the introduction of a new capitation formula which led to marked changes in the total practice revenue.

We drop practices with non-GMS contracts (10.8% of the initial 513,629 observations) because they were rewarded differently for QOF achievement and paid directly for some activities counting towards QOF indicators. We also drop practices with fewer than 1000 patients (2.2% of the initial sample) because they may be in the process of opening or closing. ISD does not report numbers of patients for whom an indicator was achieved, the number declared eligible for the indicator, and number with the condition, if these are less than five patients. We drop these observations (5.5% of the initial sample). Two practice year observations with reported prevalence for a condition greater than the list size were also dropped. The final data set is an unbalanced panel of 815–857 practices with 418,720 practice-year-indicator observations. Details of the multi-reward indicators between 2005/6 and 2012/13 are in Appendix A2.

Ethical approval is not required for this study. All data used in this study is publicly available. No data was directly collected from human subjects.

4. Results

Table 1 has summary statistics on the characteristic and outcomes of single and multi-reward indicators. Multi-reward indicators were less likely to be for recording patient characteristics than single reward indicators. Multi-reward and single reward indicators had similar upper thresholds and maximum points. On average, multi-reward activities had higher population achievement rates and lower exception rates than single reward indicators. Reported achievement varies positively with population achievement and negatively with exception reporting. The higher population achievement of multi-reward indicators more than offsets their lower exception rate, so that reported achievement $(N/(P-E) = (N/P)/(1-E/P))$ is higher for multi-reward indicators.

Notice that for multi-reward indicators reported achievement $(N/(P-E))$ is on average above the upper threshold beyond which increase in N or E do not increase practice revenue. For single reward indicators overall reported achievement is very close to the upper threshold. There are three possible reasons for over achievement. First, some indicators may apply only to a small number of patients. Thus a practice with ten patients eligible for an indicator with an upper threshold of 65% will have to treat 7 of them (70% reported achievement) to get the maximum points. Second, to reduce the risk of losing points when treated patients leave before the financial year end date at which reported achievement is counted, practices may treat more than enough to achieve the upper threshold. Third, GPs may also care about patient health as well as income and so wish to achieve an indicator for patients even when this has no effect on practice revenue from the QOF. The strength of the first two of these motives is, other things equal, greater for multi-reward indicators.

4.1. Panel models

Table 2 reports key results from the models for three outcome measures (*panels a* to *c*) with three measures (columns (i) to (iii)) of the extent to which an indicator was multi-reward (full results are in Appendix A3).

In *panel a* population achievement (N/P) is higher for multi-reward indicators (column (i)) and in columns (ii) and (iii) where population achievement increases with the number of other indicators for which the activity is rewarded. In *panel b* exception reporting (E/P) is smaller for multi-reward indicators. In *panel c* reported achievement $(N/(P-E))$ is greater for multi-reward indicators and increases with the number of other indicators rewarded for the same activity.

The coefficients on the covariates (Appendix A3) show that treatment and outcome indicators have smaller reported achievement and population achievement rates than record indicators and treatment indicators have higher exception rates. The coefficient on maximum available QOF points is statistically insignificant for all models. Reported and population achievement are smaller in larger, rural (except for very remote small towns), and dispensing practices. Practices with a higher proportion of the young (less than 85 years old) have higher reported achievement and population achievement. The exception rate is higher in larger practices, those with a higher proportion of female

Table 2
Practice QOF performance and multi-reward indicators 2005/6–2012/13.

	Measure of multiple reward					
	(i)		(ii)		(iii)	
	Indicator is multi-reward		Number other indicators rewarded for same activity		Number other indicators rewarded for same activity	
	coef	(SE)	coef	(SE)	coef	(SE)
a. Population achievement rate (N/P)						
M dummy	0.0456***	(0.0123)				
n other			0.0165***	(0.0037)		
1 other					0.0297	(0.0158)
2 other					0.0440**	(0.0146)
3 other					0.0131	(0.0166)
4 other					0.0762***	(0.0162)
5 other					0.0993***	(0.0159)
R^2	0.3297		0.3354		0.3443	
b. Exception rate (E/P)						
M dummy	−0.0263**	(0.0098)				
n other			−0.0088**	(0.0028)		
1 other					−0.0191	(0.0124)
2 other					−0.0342**	(0.0112)
3 other					0.0057	(0.0141)
4 other					−0.0433***	(0.0121)
5 other					−0.0604***	(0.0104)
R^2	0.2152		0.2167		0.2294	
c. Reported achievement (N/(P–E))						
M dummy	0.0246***	(0.0062)				
n other			0.0095***	(0.0020)		
1 other					0.0133	(0.0085)
2 other					0.0151*	(0.0075)
3 other					0.0226**	(0.0068)
4 other					0.0410***	(0.0092)
5 other					0.0484***	(0.0108)
R^2	0.2112		0.2155		0.2160	

Notes. Specification (1) linear practice fixed effects, SEs clustered on indicators. R^2: overall. Total indicator-year-practice observations: 418,720, 815 to 857 practices, 8 years. *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

Table 3
Smoking indicators 2005/6–2007/8.

Indicator	Definition	Max points	Lower threshold	Upper threshold	Population achievement (N/P)	Exception rate (E/P)	Reported achievement (N/(P–E))
				2005/6			
ASTHMA4	Record	6	25	70	0.9289	0.0251	0.9529
BP2	Record	10	25	90	0.9722	0.0121	0.9840
CHD3	Record	7	25	90	0.9586	0.0181	0.9763
COPD4	Record	6	25	90	0.9412	0.0321	0.9724
DM3	Record	3	25	90	0.9718	0.0148	0.9864
STROKE3	Record	3	25	90	0.9372	0.0309	0.9672
All smoking record indicators					0.9572	0.0181	0.9749
ASTHMA5	Cessation advice	6	25	70	0.8590	0.0524	0.9073
BP3	Cessation advice	10	25	90	0.9744	0.0074	0.9816
CHD4	Cessation advice	4	25	90	0.9285	0.0316	0.9589
COPD5	Cessation advice	6	25	90	0.9246	0.0385	0.9617
DM4	Cessation advice	5	25	90	0.9318	0.0334	0.9638
STROKE4	Cessation advice	2	25	70	0.8984	0.0421	0.9380
All smoking cessation advice indicators					0.9280	0.0293	0.9560
				2006/7 and 2007/8			
SMOKING1	Record	33	40	90	0.9587	0.0047	0.9632
SMOKING2	Cessation advice	35	40	90	0.9283	0.0097	0.9374

Notes. Record: % patients with record of smoking status in the previous 15 months, except those who have never smoked where smoking status need only be recorded once since diagnosis. Cessation advice: % patients who smoke whose notes contain a record that smoking cessation advice or referral to a specialist service, where available, has been offered within the previous 15 months. ASTHMA4, ASTHMA5: patients aged 20 and over. SMOKING1, SMOKING2 are for patients with any or any combination of asthma (age 20 and above), hypertension, CHD, COPD, diabetes, stroke. Numerators and denominators are sums of numerators and denominators over all practices with data.

Table 4
Difference in differences in mean performance.

	Smoking indicator				Control group of indicators				DIDs (2006/7-07/08 vs 2005/6)		
	n obs	Population achievement (N/P)	Exception rate (E/P)	Reported achievement (N/(P–E))	n obs	Population achievement (N/P)	Exception rate (E/P)	Reported achievement (N/(P–E))	Population achievement (N/P)	Exception rate (E/P)	Reported achievement (N/(P–E))
	Smoking status record				*Single reward indicators*				*Smoking status record*		
2005/6	646	0.9570	0.0187	0.9752	4214	0.8619	0.0581	0.9141	−0.0020	−0.0192***	−0.0209***
2006/7-07/08	1292	0.9571	0.0050	0.9619	8428	0.8640	0.0637	0.9218	(0.0055)	(0.0030)	(0.0040)
	Smoking status record				*Multi reward indicators*				*Smoking status record*		
2005/6	646	0.9570	0.0187	0.9752	4250	0.9352	0.0299	0.9639	−0.0040*	−0.0146***	−0.0185***
2006/7-07/08	1292	0.9571	0.0050	0.9619	8500	0.9392	0.0309	0.9692	(0.0015)	(0.0009)	(0.0016)
	Smoking cessation advice				*Single reward indicators*				*Smoking cessation advice*		
2005/6	795	0.9291	0.0286	0.9564	2295	0.7502	0.1441	0.8726	−0.0176	−0.0208**	−0.0396**
2006/7-07/08	1590	0.9282	0.0099	0.9375	4590	0.7669	0.1462	0.8934	(0.0084)	(0.0032)	(0.0068)
	Smoking cessation advice				*Multi reward indicators*				*Smoking cessation advice*		
2005/6	795	0.9291	0.0286	0.9564	10964	0.7812	0.1017	0.8709	−0.0124*	−0.0265***	−0.0395***
2006/7-07/08	1590	0.9282	0.0099	0.9375	21928	0.7928	0.1096	0.8915	(0.0057)	(0.0030)	(0.0047)

Notes. n obs: number of observations. Control groups: for smoking status record: 7 record indicators which were single reward in 2005/6, 2006/7–2007/8, 7 record indicators which were multi reward in 2005/6, 2006/7 and 2007/8; for smoking cessation advice: 3 treatment indicators which were single reward in 2005/6, 2006/7–2007/8, 14 treatment indicators which were multi-reward in 2005/6, 2006/7–2007/8. DID: change in mean performance on smoking indicator minus change in mean of performance on controls. SEs on DID coefficients are clustered on indicators. *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

patients, a higher proportion of very old (85+), a higher proportion of male GPs, in urban areas, and in less deprived areas.

Since the number of eligible patients is small in some cases we also estimated negative binomial count data regressions as robustness checks and obtained qualitatively and quantitatively similar results (Appendix A4). Results were also robust (Appendix A5) to weighting by average list size, to dropping 2005/6 as the there was a major change in the QOF from 2006/7, and to using a balanced sample of practices. We also estimated models with a balanced sample of practices and indicators, obtaining the same pattern of results but with reduced significance because of the loss of 75% of observations. Finally, we added interactions of indicator type (treatment, intermediate, and outcome) with the condition to which the indicator applied to further control for differences in the effects of the indicator on patient health and practice cost. Results were very similar to those in Table 2 with a slight increase in the effect of multi-reward.

4.2. Difference in differences models

Table 3 has summary statistics for the smoking indicators. The total points available for the 6 smoking record indicators in 2005/6 was 35 and for the single SMOKING1 record indicator in 2006/7–2007/8 it was 33. Total points available for smoking cessation advice was 33 in 2005/6 and 35 in 2006/7–2007/8. Thus points for recording and cessation

Table 5
Difference in differences estimates: removal of multiple rewards for smoking.

		Population achievement (N/P)		Exception rate (E/P)		Reported achievement (N/(P–E))	
		Recording smoking	Cessation advice	Recording smoking	Cessation advice	Recording smoking	Cessation advice
(a) DIDs	Control						
δ_1: 2006/7–2005/6	always multi	−0.0269***	−0.0316**	−0.0020	−0.0144	−0.0313***	−0.0487***
		(0.0051)	(0.0093)	(0.0030)	(0.0073)	(0.0031)	(0.0049)
δ_2: 2007/8–2005/6	always multi	−0.0258***	−0.0328**	−0.0019	−0.0116	−0.0302***	−0.0477***
		(0.0053)	(0.0099)	(0.0030)	(0.0073)	(0.0034)	(0.0054)
$\delta_1 - \gamma_1$: 2006/7–2005/6	always single	−0.0208**	−0.0205	−0.0094	−0.0173	−0.0315***	−0.0404**
		(0.0062)	(0.0182)	(0.0052)	(0.0090)	(0.0034)	(0.0127)
$\delta_2 - \gamma_2$: 2007/8–2005/6	always single	−0.0188*	−0.0235	−0.0124	−0.0062	−0.0324***	−0.0322*
		(0.0065)	(0.0219)	(0.0062)	(0.0091)	(0.0040)	(0.0153)
(b) Change in DIDs							
$\delta_2 - \delta_1$	always multi	0.0012	−0.0012	0.0001	0.0028	0.0012	0.0009
		(0.0010)	(0.0021)	(0.0009)	(0.0018)	(0.0006)	(0.0012)
$(\delta_2 - \gamma_2) - (\delta_1 - \gamma_1)$	always single	0.0020	−0.0030	−0.0030	0.0111*	−0.0009	0.0081*
		(0.0023)	(0.0049)	(0.0019)	(0.0047)	(0.0016)	(0.0033)
n obs		22936	35140	22936	35140	22936	35140

Notes. Specification (2). Robust SEs clustered on indicators. *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

advice changed little and there was no change in the total available points for both activities. (Appendix A1 has a simple theory model showing that even though the total points for the smoking indicators were similar before and after the change, the switch to single from multi-reward would have changed financial incentives to achieve the indicators.) In 2005/6 the 12 smoking record and cessation advice indicators have the same lower thresholds and nine had upper thresholds of 90% and three of 70%. In 2006/7–2007/8 the single smoking record and cessation indicators had a 40% lower threshold and 90% upper threshold. Average reported achievement was comfortably above the upper threshold for all indicators.

In Table 4 we compare the mean performance for the smoking and non-smoking control indicators before and after the change in the smoking indicators from multi to single reward. Population achievement, exception reporting, and reported achievement for the smoking indicators were lower in 2006/7 and 2007/8 than in 2005/6, except for population achievement in smoking status recording. Population achievement, exception reporting, and reported achievement increased for all the control groups of indicators. The unconditional DID are negative and statistically significant.

Table 5 has results from the DID specification (2). In panel (a) the 2006/7–2005/6 DIDs (δ_1) and the 2007/8–2005/6 DIDs (δ_2) for population achievement (N/P) and reported achievement ($N/(P-E)$) are negative and statistically significant for both recording and cessation advice when the control group is always multi-reward indicators. When the control group is always single indicators the DIDs ($\delta_t - \gamma_t$) for population achievement and reported achievement are generally less negative and are statistically insignificant in two of the eight DIDs. Exception rates for the smoking indicators were also reduced relative to the other types but the changes are always statistically insignificant. (Appendix A6 has similar results when 2006/7 to 2007/8 is used as a single post change period.)

Panel (b) compares the 2006/7–2005/6 and 2007/8–2005/6 DIDs. If the negative DIDs in panel (a) were due to an underlying downward trend, not the change in incentives, then the 2007/8–2005/6 DIDs would be more negative than the 2006/7–2005/6 DIDs: $\delta_2 - \delta_1 < 0$. With the always multi-reward control group the change in the DIDs ($\delta_2 - \delta_1$) is never statistically significant and is positive in five of the six cases. With the always single control group the change in the DIDs $(\delta_2 - \gamma_2) - (\delta_1 - \gamma_1)$ is statistically insignificant in four cases and in two cases it statistically significant but positive: the 2007/8–2005/6 DID is less negative than the 2006/7–2005/6 DID.

5. Conclusion

5.1. Study limitations

GPs are likely to be intrinsically as well as financially motivated. It is therefore possible that some of the difference between achievement on multi and single reward indicators found in the pooled panel analysis arises because GPs perceive multi-reward indicators as more beneficial effect for patients. However, it seems unlikely that all of the performance difference arises from intrinsic motivation. Our pooled models include potential controls for effects on patients: type and prevalence of the conditions to which indicators apply and whether the indicators are for recording, treatment, or intermediate outcome. Moreover, a recent study (Wilding et al., 2018) investigated the removal of some indicators from the QOF in 2015/16, so that the activities covered by the indicators were no longer financially incentivised. Although the health implications of the activities previously incentivised were unchanged, achievement fell for both multi and single reward indicators, implying that GPs respond to financial incentives even if this has negative health consequences for their patients.

This interpretation of the panel model results is supported by the results from the DID models which avoid the problem of distinguishing between intrinsic and financial motivation since they examine the effect of a change from multi to single reward for a given set of activities related to control of smoking. These models show that when smoking indicators changed from multi to single reward in 2006/7 achievement fell relative to control groups of indicator which were always single or always multi rewarded. The DID design requires that in the absence of the policy change the control and treatment groups would have had parallel trends. Since there was data for only one pre-policy change year 2005/6 we could not apply conventional tests for parallel trends over the pre-incentive change period. But we found that the 2007/8–2005/6 DIDs were either statistically indistinguishable from the 2006/7–2005/6 DIDs or were less negative. This suggests that the reductions in achievement for smoking indicators relative to the control indicators in 2006/7 and 2007/8 compared to 2005/6 were due to changes in incentives, not to pre-existing trends.

Both population and reported achievement for the two smoking indicators were closer to the 100% maximum than the control indicators. Thus their decline in achievement in 2006/7 and 2007/8 relative to the controls might be due to a ceiling effect: it was possible for achievement to increase for the controls but not the smoking indicators. However, reported achievement for recording smoking and population and reported achievement for smoking cessation advice fell in 2006/7–2007/8 relative to 2005/6 so that their reduction relative to the controls could not be due to the ceiling effect.

Our pooled and DID models are complementary: the pooled models utilise data on all multi and single rewarded indicators over eight years and the DID models examine the effect of changing from multi to single reward for a single group of indicators but thereby controls for any unobserved characteristics of indicators which might affect achievement.

We have practice, rather than patient, level data and so cannot directly compare the effects of single versus multiple rewards on individual patients, especially for multi-morbid patients. However, in examining practice level outcomes we at least partially control for practice level multi-morbidity by including prevalence rates for over 20 conditions, practice age and gender mix, and deprivation. Our administrative data has information on a large number of practices and we are able to allow for practice characteristics (such as list size and GP characteristics) and to attribute local small area socio-economic characteristics to the practice list. Although patient level clinical datasets are not panaceas, often being derived from a subset of possibly unrepresentative practices and having sparse information on practice characteristics and patient socio-economic characteristics, we plan in future work to use patient level data to examine patient outcomes under single and multi-reward indicators and the change in outcomes when indicators change from multi to single reward.

5.2. Discussion

Comparing practice performance on all multi-reward and single reward indicators between 2005/6 and 2012/13, we find, after allowing for other characteristics of indicators and for practice characteristics, that population achievement (N/P), which can be interpreted as a measure of quality, was 0.046 higher for multi-reward indicators compared to the overall mean of 0.830. Exception reporting (E/P) was lower for multi-reward indicators (by 0.026 compared to the overall mean exception rate of 0.091). Reported achievement ($N/(P-E)$), which determines practice revenue from the QOF, is higher for multi-reward indicators (by 0.025 compared to the mean of 0.911 for all indicators). The differences in population achievement, exception reporting, and reported achievement between multiple and single reward indicators were greater when the number of other indicators rewarded for the same activity was greater.

Holding the disease prevalence (P) constant, increases in the number of patients for whom an indicator is achieved (N) and increases in the number exception reported (E) both increase reported achievement ($N/(P-E)$) and, if reported achievement is below the upper

threshold, will increase practice revenue from the QOF. But we find that multi reward increase population achievement (*N/P*) and reduces the exception reporting rate (*E/P*). The explanation may be that GPs are intrinsically as well as extrinsically motivated (McGuire, 2000; Bénabou and Tirole, 2006). Increasing achievement on a quality indicator benefits patients whereas gaming exception reporting does not and offends against professional norms. If the incentives from multiple rewards lead the practice to choose a high enough *N* to take reported achievement above the upper threshold then the practice does not need to increase *E* and could reduce it without financial penalty. Thus our results are in line with Gravelle et al. (2010), who found that exception reporting for an indicator was lower if the practice was above the upper threshold in the previous year, and with Kontopantelis et al. (2012), who found that exception reporting was increased after an increase in the upper threshold if the practice was below the new threshold in the year before the threshold increased.

For two activities (recording smoking status and providing smoking cessation advice) population achievement and reported achievement fell in 2006/7 when multi-reward indicators were replaced by single reward indicators. Relative to indicators which were always multi-reward, population achievement for smoking recording fell by 0.0264 compared to the 2005/6 mean of 0.9573 and for smoking cessation advice it fell by 0.0322 compared to the 2005/6 mean of 0.9310. These DID estimated effects of changing from multi to single reward for smoking indicators are around half to one third smaller than the effect implied by the pooled model comparing multi with single reward indicators in general. They are, nevertheless, sizeable compared with those induced by other changes to the QOF, such as changing upper thresholds (Kontopantelis et al., 2012). The change from multi to single reward for smoking indicators also reduced exception reporting, though the effect was statistically insignificant and around half that suggested by the pooled panel models using all indicators. The contrast in exception reporting results for the pooled panel models for all indicators and the DID models for smoking may be because exception rates for smoking indicators were only around one third of those for other indicators.

The greater practice revenue from increased achievement for multi-reward indicators compared to single reward indicators suggests intuitively that achievement will be greater for multiple reward indicators. But in the absence of very detailed information about the extent of multi-morbidity, which will determine the effects on revenue, and about the cost of increasing achievement, we could not test specific hypotheses about individual indicators. Instead we have tested the general hypothesis that GPs respond to the differences in incentives for multiple reward indicators compared to single reward indicators. If we had found no overall effect of multi-reward versus single reward this might have been because there were offsetting positive and negative effects across the individual indicators classified as multi or single reward or it might be because GPs do not respond to the difference in incentives. Our finding of overall higher population achievement and reported achievement, and lower exception reporting, for multi reward indicators is intuitively plausible and is evidence that GPs do respond to the fine details of incentive schemes.

The QOF did not explicitly target multi-morbid patients but multiple rewards for screening, monitoring, and treating patients increase incentives to improve quality for multi-morbid patients relative to those with only one condition. The implicit incentives from multi-reward indicators may thus have led to better outcomes for multi-morbid patients. The replacement of the multi-reward smoking indicators by single reward indicators in 2006/7 may have disadvantaged multi-morbid patients since after the change the additional revenue the practice would earn for recording smoking status or offering smoking cessation advice was the same for multi-morbid patients as for patients with only one condition and achieving the indicator for multi-morbid patients may be more costly for the practice.

The financial incentives for treating multi-morbid patients when there are multi-reward indicators depend not just on the numbers of multi-morbid patients but also on the numbers of singly morbid patients with the same conditions. If multi-morbid patients have greater need for good quality primary care (Salisbury et al., 2011) it would be better to make the financial incentives for targeting them explicit and uniform across practices, rather than implicit and accidental.

Our results have wider relevance for the general design of P4P schemes. They show that decision-makers respond to the implicit incentives in P4P schemes, not just to the explicit links between measures of performance and financial rewards. Such implicit incentives may not be anticipated by P4P designers and can thus yield unintended consequences. For example, other studies of the QOF have shown that exception reporting, whose main purpose was to remove an incentive to over treat patients, led some GPs to exception report patients in order to boost their earnings from the QOF (Gravelle et al., 2010). Exception reporting is an explicit and prominent feature of the QOF clinical indicators. The implicit incentives created by multiple rewards for some QOF indicators are less salient but they nevertheless affected GP behaviour and hence patient outcomes. Given that P4P schemes often create complex multi-tasking incentives, the fact that decision makers respond to fine details reinforces the case for P4P schemes to be subject to simulations and trials to assess their effect on behaviour and outcomes.

Declaration of interest statement

While preparing the original submission of the manuscript, Yan Feng was an employee of the Office of Health Economics (a registered charity which received funding from a variety of sources including the Association of the British Pharmaceutical Industry). Hugh Gravelle declares no actual or potential conflict of interest.

Acknowledgements

We are grateful for helpful suggestions and comments from Rodrigo Moreno-Serra, Mark Monahan, and participants in the 2017 winter UK HESG. We also thank Grant Aitken (ISD General Practices) for help in locating archived QOF Business Rules, and James McNally (ISD Scotland) for help with the GP practice information.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.socscimed.2019.05.048.

References

Armour, B.S., Pitts, M.M., Maclean, R., Cangialose, C., Kishel, M., Imai, H., Etchason, J., 2001. The effect of explicit financial incentives on physician behavior. Arch. Intern. Med. 161 (10), 1261–1266.

Bénabou, R., Tirole, J., 2006. Incentives and prosocial behavior. Am. Econ. Rev. 96 (5), 1652–1678.

Christianson, J.B., Leatherman, S., Sutherland, K., 2007. Financial Incentives, Healthcare Providers and Quality Improvements. A Review of the Evidence. The Health Foundation, London, UK.

Coulson, J., 2011. Quality and outcomes framework review and assessment – guidance for GP practices. Accessed 25th November 2018. https://www.pcc-cic.org.uk/sites/default/files/articles/attachments/1_guidance_for_gp_practices_-_qof_process_2011-12.pdf.

Custers, T., Arah, O., Klazinga, N., 2006. Is there a business case for quality in The Netherlands? A critical analysis of the recent reforms of the health care system. Health Policy 82 (2), 226–239.

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. N. Engl. J. Med. 355, 375–384.

Duflo, E., 2017. The economist as plumber. Am. Econ. Rev. Am. Econ. Assoc. 107 (5), 1–26 (May).

Eijkenaar, F., Emmert, M., Scheppach, M., Schöffski, O., 2013. Effects of pay for performance in health care: a systematic review of systematic reviews. Health Policy 110 (2–3), 115–130.

Epstein, A., 2012. Will pay for performance improve quality of care? The answer is in the details. N. Engl. J. Med. 367, 1852–1853.

Gené-Badia, J., Escaramis-Babiano, G., Sans-Corrales, M., Sampietro-Colom, L., Aguado-Menguy, F., Cabezas-Peña, C., Gallo-Puelles, P., 2007. Impact of economic incentives on quality of professional life and on end-user satisfaction in primary care. Health Policy 80 (1), 2–10.

Gillam, S.J., Siriwardena, N., Steel, N., 2012. Pay-for-Performance in the United Kingdom: impact of the quality and outcomes framework - a systematic review. Ann. Fam. Med. 10 (5), 461–468.

Greb, S., Focke, A., Hessel, F., Wasem, J., 2006. Financial incentives for disease management programmes and integrated care in German social health insurance. Health Policy 78 (2–3), 295–305.

Gravelle, H., Sutton, M., Ma, A., 2010. Doctor behaviour under a pay for performance contract: treating, cheating and case Finding? Econ. J. 120 (542), F129–F156.

Greene, S.E., Nash, D.B., 2009. Pay for performance: an overview of the literature. Am. J. Med. Qual. 24 (2), 140–163.

Guthrie, B., McLean, G., Sutton, M., 2006. Workload and reward in the Quality and Outcomes Framework of the 2004 general practice contract. Br. J. Gen. Pract. 56 (532), 836–841.

Guthrie, B., Tang, J., 2016. What did we learn from 12 years of QOF? Scottish school of primary care. 29th November 2016. http://www.sspc.ac.uk/media/media_486342_en.pdf.

Holmstrom, B., Milgrom, P., 1991. Multitask principal-agent analyses: incentive contracts, asset ownership, and job design. J. Law Econ. Organ. 7, 24–52.

ISD Scotland, 2016a. NHS Scotland payments to general practice. Financial year 2015/16. Accessed 7th August 2017. http://www.isdscotland.org/Health-Topics/General-Practice/Publications/2016-11-01/2016-11-01-ScotlandGPPayments2015-16-Report.pdf.

ISD Scotland, 2016b. General Practice. GP Workforce and practice list sizes 2006 – 2016. Accessed 7th August 2017. https://www.isdscotland.org/Health-Topics/General-Practice/Publications/2016-12-13/2016-12-13-GPWorkforce2016-Report.pdf.

Kontopantelis, E., Doran, T., Gravelle, H., Goudie, R., Siciliani, L., Sutton, M., 2012. Family doctor responses to changes in incentives for influenza immunization under the U.K. Quality and Outcomes Framework pay-for-performance scheme. Health Serv. Res. 47 (3 Pt1), 1117–1136.

Kristensen, S., McDonald, R., Sutton, M., 2013. Should pay for performance schemes be locally designed? Evidence from the commissioning for quality and innovation (CQUIN) framework. J. Health Serv. Res. Policy 18 (2 Suppl. l), 38–49.

McGuire, T., 2000. Physician agency. In: In: Culyer, A., Newhouse, J. (Eds.), Handbook of Health Economics 1A North Holland.

Meacock, R., Kristensen, S., Sutton, M., 2014. Paying for improvements in quality: recent experience in the NHS in England. Nordic J. Health Econ. 2 (1), 239–255.

Mehrotra, A., Sorbero, M.E., Damberg, C.L., 2010. Using the lessons of behavioral economics to design more effective pay-for-performance programs. Am. J. Manag. Care 16 (7), 497–503.

Moulton, B.R., 1986. Random group effects and the precision of regression estimates. J. Econom. 32 (3), 385–397.

Perkins, R., Seddon, M., 2006. Quality improvement in New Zealand healthcare. Part 5: measurement for monitoring and controlling performance–the quest for external accountability. N. Z. Med. J. 119 (1241), U2149.

Petersen, L.A., Woodard, L.D., Urech, T., Daw, C., Sookanan, S., 2006. Does pay for-performance improve the quality of health care? Ann. Intern. Med. 145 (4), 265–272.

Pink, G.H., Brown, A.D., Studer, M.L., Reiter, K.L., Leatt, P., 2006. Pay-for-performance in publicly financed healthcare: some international experience and considerations. HealthcarePapers 6 (4), 8–26.

Rebitzer, J., Taylor, L., 2011. Extrinsic rewards and intrinsic motives: standard and behavioural approaches to agency and labor markets. In: In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. 4a Elsevier.

Rosenthal, M., Landon, B., Normand, S., Frank, R., Epstein, M., 2006. Pay for performance in commercial HMOs. N. Engl. J. Med. 355, 1895–1902.

Roth, A., 2002. The economist as engineer: game theory, experimentation, and computation as tools for design economics. Econometrica 70 (4), 1341–1378.

Ryan, A.M., 2009. Effects of the premier hospital quality incentive demonstration on Medicare patient mortality and cost. Health Serv. Res. 44 (3), 821–842.

Salisbury, C., Johnson, L., Purdy, S., Valderas, J.M., Montgomery, A.A., 2011. Epidemiology and impact of multimorbidity in primary care: a retrospective cohort study. Br. J. Gen. Pract. 61 (582), e12–21.

Scott, A., Sivey, P., Ait Ouakrim, D., Willenberg, L., Naccarella, L., Furler, J., Young, D., 2011. The effect of financial incentives on the quality of health care provided by primary care physicians. Cochrane Database Syst. Rev. 9, CD008451.

Scott, A., 2007. Pay for performance in health care: strategic issues for Australian experiments. Med. J. Aust. 187 (1), 31–35.

Scottish Government, 2018. The Scottish Index of multiple deprivation. Accessed 20th August 2018. https://www.gov.scot/Topics/Statistics/SIMD/.

Sutton, M., Nikolova, S., Boaden, R., Lester, H., McDonald, R., Roland, M., 2012. Reduced mortality with hospital pay for performance in England. N. Engl. J. Med. 367, 1821–1828.

Town, R., Kane, R., Johnson, P., Butler, M., 2005. Economic incentives and physicians' delivery of preventive care: a systematic review. Am. J. Prev. Med. 28 (2), 234–240.

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 Serv. Res. 10, 247.

Wilding, A., Kontopantelis, E., Munford, L., Sutton, M., 2018. Impact of Removing Indicators from the Quality and Outcomes Framework: Retrospective Study Using Individual Patient Data in England. Policy Research Unit in Commissioning and the Healthcare System Report to NHS England 28 June.