THE IMPACT OF PROVIDER INCENTIVES ON PROFESSIONALS AND PATIENTS

A THESIS SUBMITTED TO THE UNIVERSITY OF MANCHESTER FOR THE DEGREE OF PHD IN MEDICINE IN THE FACULTY OF MEDICAL AND HUMAN SCIENCES

2015

THOMAS ALLEN

SCHOOL OF MEDICINE

TABLE OF CONTENTS

Word count: 69,198

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACSC – Ambulatory Care Sensitive Conditions

DID – Difference in Differences

ELSA – English Longitudinal Study of Aging

FTE – Full-time Equivalent

GMS – General Medical Services

GP WLS – General Practitioner Work Life Survey

HSCIC – Health and Social Care Information Centre

LPM – Linear Probability Model

NHS – National Health Service

NICE – National Institute for Health and Care Excellence

OLS – Ordinary Least Squares

P4P – Pay for Performance

PBC – Practice Based Commissioning

PCT – Primary Care Trust

PMS – Primary Medical Services

QOF – Quality and Outcomes Framework

RCT – Randomised Controlled Trial

ABSTRACT

THE UNIVERSITY OF MANCHESTER
THOMAS ALLEN
PHD IN HEALTH ECONOMICS
THE IMPACT OF PROVIDER INCENTIVES ON PROFESSIONALS AND PATIENTS
2015

Healthcare providers are motivated by a combination of financial and non-financial incentives. This thesis focuses on two specific forms of these incentives: pay-for-performance (P4P) and reputation. Despite increased use, there is limited evidence on how financial and reputational incentives interact, or on how financial incentives affect patients and professionals. We further our understanding with respect to P4P and make recommendations about the design of future schemes. We achieve the thesis aims by producing four empirical studies. Each empirical study uses data collected as part of a national P4P in the English National Health System, the Quality and Outcomes Framework.

We begin with an investigation of the relevant importance of financial and reputational incentives in determining provider performance. We use administrative data for nine years of practice performance on a range of indicators totalling close to five million observations. This analysis covers a period during which the financial and reputational rewards were changing. We find that initially, financial incentives had a larger effect on performance. Over time, reputational incentives become more important.

Our second study uses changes in the organisational structure of healthcare providers to explore whether the observed similarity in the performance of nearby practices can be explained by peer effects. We measure the performance of 8,000 individual practices and their peers for five years. When peer groups are merged, there is a reduction in peer effects for old peers and an increase in peer effects for new peers. Practices seem to be pulled down by the presence of poor peers in their group.

In our third study we measure the impact of variations in the proportion of income at risk to P4P on the working lives of GPs. We combine administrative data with survey data from before and after the introduction of P4P. Our sample consists of approximately 2,000 GPs who provide detailed information about their working lives. We find that providers are unaffected by these variations despite income at risk being high.

Finally, to observe the relationship between quality of care reported at the patient level and at the practice level we link practice performance with a detailed survey of the English population aged over 50 years. Correlations are generally smaller than expected and negative for some areas. Practices may have lacked an incentive to communicate their care adequately to patients and may have diverted attention away from areas of care without financial incentives.

Non-financial incentives can be effective motivators when peer performance is observable. Professionals are also unlikely to associate negatively with income being related to their performance, along as incomes remain high. However, patients may suffer from a lack of communication about the type of care they are receiving.

DECLARATION

No portion of this work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

The analysis and writing for all chapters was done by Thomas Allen. Matt Sutton and William Whittaker provided supervision on all chapters.

Evangelos Kontopantelis (Centre for Health Informatics and NIHR School for Primary Care Research, University of Manchester) provided a cleaned dataset and advice for Chapter 4. James Banks (Economics Discipline Area, University of Manchester) advised on design and data relating to Chapter 7.

COPYRIGHT STATEMENT

ACKNOWLEDGMENTS

I would like to thank Matt Sutton and William Whittaker for their supervision and support throughout my time as a PhD student. There is a strong causal relationship between their guidance and advice, and my continued career and personal development.

My progress at MCHE has been enhanced by the quality of my peers and their feedback at my numerous “Monday Meeting” appearances. They have indulged my steadfast belief that a full page regression table has no place in a presentation. Colleagues have become friends, and for that I am pleased.

Generous funding has been provided by the Faculty of Human and Medical Sciences and the Alumni of The University of Manchester.

A special thank you is reserved for my fiancée Vicky. She has been by my side throughout my life as a student and I am grateful for her support.

DISSEMINATION OF WORK

Aspects of the literature review undertaken for Chapters 2 and 3 have been published in Allen et al. (2014b, 2014c).

An earlier version on Chapter 4 was presented and discussed at the HESG in York in 2011. Chapter 4 was presented (by Matt Sutton) as a seminar to the Health Economics Unit, University of Birmingham in 2015.

An earlier version of Chapter 5 was presented and discussed at the Workshop on spatial health econometrics, Cambridge in 2015

An earlier version of Chapter 6 was presented and discussed at the HESG in Glasgow in 2014 and the International Health Economics Association World Congress, Milan in 2015. This chapter was also presented to the Centre for Primary Care at The University of Manchester in 2014. A poster based on this chapter was presented at the Faculty of Medical and Human Sciences Student Showcase at The University of Manchester in 2014 and the European Health Economics Association (EuHEA) PhD Student-supervisor and Early Career Researcher (ECR) conference at The University of Manchester in 2014.

An earlier version of Chapter 7 was presented and discussed at the Health Economists' Study Group (HESG) in Warwick in 2013.

1. INTRODUCTION

Healthcare providers can be paid on the basis of block contracts, capitation, per diem rates, case-based payments, fee for service, pay for performance (P4P), or a mixture of these methods Appleby et al. (2012). P4P seeks to improve the quality of healthcare by incentivising providers though the use of financial rewards linked to quality measures (Van Herck et al., 2010). P4P is increasingly being adopted in many countries (Eijkenaar, 2012), despite the evaluations of such schemes often having flawed methods (Scott et al., 2011) and despite limited evidence that schemes are effective or cost-effective (Eijkenaar et al., 2013).

The impact of P4P extends beyond clinical effects and cost, and on to professionals and patients in domains such as job satisfaction and perceptions of care (Gillam et al., 2012; Van Herck et al., 2010). These areas have received relatively little attention in the literature as has the relationship between financial and non-financial incentives.

This thesis contains four empirical investigations of the existence and mechanisms behind reputational incentives, and the impacts of P4P on patients and professionals. We address the aforementioned gap in the literature by using the Quality and Outcomes Framework (QOF) as the basis for these empirical studies. The QOF is a large-scale P4P scheme for general practices in the UK and was introduced in 2004 as part of wider contractual changes to primary care (Roland, 2004). The scheme rewards practices based on their performance on a range of clinical and non-clinical areas and these financial rewards were significant in size (National Audit Office, 2008). The scheme was voluntary in principle but practice participation was near 100% (Lester and Campbell, 2010)

The thesis chapters are structured as followed: (1) introduces and summarises each of the four empirical chapters; (2) reviews of the literature on P4P; (3) introduces the QOF and its mechanics; (4) investigates how the financial and reputational incentives of a P4P scheme affect performance; (5) investigates the existence of peer effects under a P4P scheme in primary care organisations; (6) investigates the effect of P4P exposure on the working lives of GPs; (7) investigates the relationship between provider-reported performance under a P4P scheme and patient-reported quality of care; and (8) outlines the main findings, the strengths and weaknesses, and implications for future research and policy.

1.1 Revenue or reputation: quantitative analysis of general practitioner motivation

Purchasers have used both financial and reputational incentives to motivate providers of healthcare to increase effort and improve quality. The relative effectiveness and sustainability of these incentives is unknown. The QOF has generally been seen as a financial incentive scheme but the scores are published and therefore represent an externally-observed marker of reputation. We contribute to the literature on financial versus reputational incentives by using an unusual feature of the QOF, which was amended during our study period, whereby the revenue from activity is neither proportional to the required activity nor to the marker of reputation. We use 4.8million observations of the clinical performance of over 9,000 general practices on 60 indicators measured over nine years from 2004/5 to 2012/13. For each practice-indicator observation, we calculate separately the financial and reputational rewards for increasing performance. We find that the reputational incentive had a smaller effect on performance than the revenue incentive when the scheme was first introduced. However, over time the effect of the reputation incentive became more important, and the effect of the revenue incentive became less important. We attribute this change in the importance of each incentive as evidence that in the early years of the QOF, reputational incentives were weaker because a benchmark for performance had yet to be established.

1.2 Institutional peer effects: evidence from English Primary Care

Institutions providing healthcare do not operate within a vacuum. Organisations are grouped together under higher regional units for higher level management and accountability. Within such groups there is an observed relationship between the quality of neighbouring healthcare providers. This relationship could be due to spatial competition for patients, common influences or to peer effects. Peer effects may emerge because of reputational incentives. We exploit a partial reorganisation of regional bodies in the English NHS, whereby the peer groups of some practices were exogenously changed. We relate the performance of 8,000 practices to the performance of their peers during a five year period, from 2004/5 to 2008/9, within which their peer group was changed. Initially there were 303 peer groups with an average size of 33 practices. Many of these groups merged in 2006/7 to form 153

groups with an average size of 63 practices. This change in peer group allows us to estimate the change in peer effects when peer groups are expanded. We find that peer effects reduce for old peers and increase for new peers when the peer groups are enlarged. Performance also seems to be influenced more by the presence of poor performing peer than by the presence of well performing peers. We propose that institutional peer effects are a plausible explanation for the spatial correlation found in performance. This conclusion implies that P4P schemes should be designed to take advantage of this non-financial incentive in the form of peer effects.

1.3 Does P4P affect the job satisfaction of general practitioners?

In a systematic review, Scott et al. (2011) highlighted the lack of attention paid to the relative size of the financial incentives compared to other sources of income and queried whether this had implications for the effectiveness of a scheme or its unintended consequences. P4P schemes are associated with positive effects on professionals such as increased pay and productivity, and also negative elements such as increased monitoring and reduced autonomy. Provider satisfaction is important in healthcare as it is related to quality of care and workforce planning. The effects of P4P schemes on job satisfaction are important for retaining and replacing the workforce. However, little is known about these effects. We add to the literature on P4P by looking at a homogeneous workforce of general practitioners (GPs) in England who were all affected by the introduction of a large scale P4P scheme in 2004. We use data from the GP Worklife Survey (GP WLS), from before the QOF in 2004 and from after the introduction of the QOF in 2005 and 2008. The sample consists of 2,000 GPs. We calculate two measures of exposure to P4P. These exposure measures are used to test for associations between the proportion of income at risk and measures of job satisfaction and working lives. Continuous difference-in-differences (DID) estimation is used to control for endogeneity in the effect of P4P exposure on job satisfaction. We find that exposure to the P4P scheme does not significantly affect job satisfaction or satisfaction sub-domains, nor does it affect hours worked or intentions to quit. GPs seem to be unaffected, in terms of their working lives, by the amount of their income at risk due to P4P.

1.4 Is provider-reported performance under a P4P scheme correlated with patient-reported quality of care?

Despite the rise in the use of financial incentives for healthcare providers, the evidence that patients notice the benefits of such schemes is limited (Van Herck et al., 2010). Instead, the majority of analysis makes use of practice-level data on performance. Patient-level data however, measures performance independently from any practice bias and acts as an alternative method of quantifying quality of care. An absence of association between provider- and patient-reported measures of quality could suggest that patients are insensitive to provider performance, or that communication problems have affected the doctor-patient relationship. To test whether patients notice changes in the care they receive we link practice-level data on quality from the QOF to 3,500 respondents to the English Longitudinal Study of Ageing (ELSA) for 2004, 2006 and 2008. ELSA asks respondents to report whether they have received diabetes and hypertension care that was also incentivised under the QOF, thus enabling a linkage of provider-reported performance and patient-reported provider performance. We find significant correlations between the quality of care reported by practices and by patients. In first difference models, we find evidence that improvements made in practice quality were noticed by patients. We also find evidence that some of these improvements may have been made at the expense of other measures of quality that were not incentivised by the QOF.

2. LITERATURE REVIEW

As with all financial incentives, P4P is grounded in economic theory and has been often studied in the context of the profit maximising firm (Prendergast, 1999). Healthcare is not provided by typical firms and agents, instead having unique features setting it part from other sectors. For these reasons this literature review is split into two sections: (1) the economic theory of P4P; and (2) the use of P4P in healthcare. Chapter 3 follows with a specific focus on the background and evidence of the QOF, the P4P scheme that is the focus of the thesis.

2.1 The economic theory of P4P

P4P specifically links an agent's pay to a measure of their performance. This is in contrast to other financial incentives such as relating pay to hours or experience. P4P is used to address the principal-agent problem which arises when an agent acts on behalf of the principal, and these parties have different and competing objectives (Prendergast, 1999). The agent wishes to maximise his or her own utility by making decisions about how much effort to exert and where to apply this effort. On the other hand, the principal wishes to maximise the firm's profits, which requires maximising revenue while minimising costs. P4P aims to solve this problem by designing an incentive which will realign the objectives of the two parties. The principal may design a compensation scheme with an aim to incentivise the agent to allocate their effort to maximise the firm's profit. Asymmetric information, or hidden information, complicates the realigning of objectives as there are elements of the agents' utility function which remain unknown to the principal (Arrow, 1963). These complications will be introduced and discussed below.

An incentive must be designed such that the agent wishes to do the task and must satisfy the following constraints: *the incentive compatibility constraint*, and *the participation constraint* (Haubrich, 1994). The first states that the incentive must create a level of utility for the agent greater than the disutility from undertaking the task. The second states that the utility from the incentive is greater than the utility from not undertaking the task, that is, leisure time.

In the classic example of a profit maximising firm, the principal aims to increase the firm's value (Jensen and Meckling, 1976). Therefore, linking agents' pay to the firm's share value may be one method to align the objectives of the two parties. However,

individual agents will often lack the influence needed to change the firm's value, so this type of incentive is inappropriate for most agents. Executive pay is often linked to firm value as they possess sufficient influence over this measure (Mehran, 1995). For other agents with less direct influence, the principal should link pay to a measure of firm performance which the agent can control and which impacts on the firm's value. If the effort of the agent could be directly observed, then the principal might choose to relate pay to this and therefore incentivise the agent to maximise their effort. However, in most situations it is only output, loosely defined, which is observable and this output is not a perfect reflection of effort (Baker, 1992).

Relating pay to output has been used in the past with mixed results. A windshield replacement firm measured output as the number of replacements made by each worker (Lazear, 2000). Output was easily measurable with a clear link to effort and the firm was observed before and after introducing the new payment scheme. The firm attracted higher quality workers, increased productivity by 44% and earned higher profits. This is an example where P4P was appropriate and had positive results.

However, when Chicago teachers were rewarded with bonuses contingent on student test performance, there was not such a clear link between teacher effort and the performance of students (Jacob and Levitt, 2003). Test scores are dependent to some extent of the teaching quality but also on student effort and background. The size of the incentives was as much as $25,000 for some teachers. These high powered incentives created an environment whereby significant rewards would have been lost if students did not perform well. The authors detected systematic cheating by teachers in various forms, including changing the answers of students in order to gain higher test scores. Classrooms with test scores found to be suspiciously high were retested under controlled conditions and scores fell to the expected level. The scheme seems to have failed in its aim to increase the quality of teaching. Incentives were too highly powered and focused on a poor measure of teacher effort, and one which was too easily manipulated. These design problems created an environment where teachers became willing and able to cheat the system to gain rewards without the increased quality the scheme intended.

The selection of an appropriate measure of output to incentivise becomes more problematic when agents engage in multiple tasks and each task consists of unique inputs and outputs. This branch of the incentives literature is known multitasking, with

the key feature being that agents have several tasks with some being easier to measure than others (Holmstrom and Milgrom, 1991). When agents have multiple tasks, P4P will result in agents focusing on the incentivised aspects or tasks. This reaction may come at the expense of other, non-incentivised, activities and result in dysfunctional behaviour (Burgess and Ratto, 2003). In these cases it has been argued that a mix of incentives would be more suitable which combines the use of P4P to allocate risk and focus attention, and a fixed fee or salary to implicitly incentivise quality standards (Eggleston, 2005).

In addition to the incorrect selection of the incentivised measure, inefficiencies and adverse behaviour can be created by targeting the incentive at teams or individuals when one of these is suboptimal. Concerns about team behaviour under team incentives were discussed in the seminal work of Holmstrom (1982), and the important concept of free-riding was explained. When agents contribute to a team output and share a reward based on that output, individuals face an incentive to reduce their own inputs. The motivation for this behaviour is that the gains outweigh the losses. When an individual reduces their input they receive all the utility gained from the decreased workload. However, the losses that result from this action, a reduced reward based on team output, would be shared across all agents in the team. Holmstrom (1982) explains that a solution to this problem would be to penalise the team, to a greater degree than the loss in output alone would imply, if output falls below a certain level. Similarly, when the teams' output is higher than a certain level, they should receive a performance bonus. These types of rules are optimal when team output is perfectly measureable but when output is measured with uncertainty the rules only perform well with small teams (Ratto et al., 2001). Within a small team, an individual has a higher probability of influencing team output and hence a penalty on team output provides a sufficient incentive not to reduce effort. In large teams, individuals have less influence on team output so the incentive created by penalties will be reduced.

When individual output is subjected to a performance incentive, agents are not induced to free ride. Gains made by lowering effort are still felt by the individual alone, but so too is the reduced income from lower output. However, individual output is often unobservable or costly to measure. Team based incentives also have an advantage over individual incentives as they foster cooperation (Burgess and Ratto,

2003). For example, a clothing manufacturer saw increases in productivity following a switch from individual to group pay (Hamilton et al., 2003). The increased output was attributed to greater collaboration within teams and the authors found non-pecuniary benefits from working in teams as evidenced by high performing workers joining teams despite the potential to lose income as a result. In another example, from an airline, bonuses were awarded if the team performed well (Knez and Simester, 2001). The team size was very large, 35,000, with theory suggesting a high potential for free riding. However, performance increased, suggesting an absence of free riding which was explained by the formation of smaller groups with peer monitoring to dissuade free riding. With well-designed team incentives, the problems associated with free riding can be mitigated or removed.

The manner in which incentives interact with an individual's motivation has been studied and developed for many decades. Gagne and Deci (2005) discussed the history of this area, charting how models of motivation moved in and out of popularity and presenting the most recent theories. Early work dichotomised motivation into intrinsic or extrinsic sources (Porter and Lawler, 1968). The former came from the task itself, such as a sense of purpose, while the latter came from a source separate to the task, such as a reward. In this framework, motivation would be maximised through the combined use of both forms of motivators, that is, they were additive.

This theory prevailed for a time until research found a more nuanced relationship between intrinsic and extrinsic motivation (Deci, 1971). Certain extrinsic incentives such as tangible financial rewards would erode intrinsic motivation rather than support it, while non-tangible verbal rewards would tend to complement intrinsic motivation. These observations lead to the development of cognitive evaluation theory which could better explain the empirical findings (Deci and Ryan, 1980).

The theory explained that some external incentives, like greater choice and positive feedback, would improve intrinsic motivation. On the other hand, intrinsic motivation would be diminished by deadlines, surveillance and other autonomy reducing activities (Deci et al., 1999). The most recent developments in this area saw a broadening of these concepts into self-determination theory wherein a spectrum of motivation was described, moving from "amotivation", the absence of motivation, through to extrinsic and then intrinsic motivation (Deci and Ryan, 2000).

An awareness of the relationship between motivation and financial rewards is important to this thesis. Extrinsic incentives can lead to a reduction in motivation. Empirical evidence of these motivation interactions comes from lab and field experiments (Ariely et al., 2009; Frey and Jegen, 2001). The general findings are that in a range of tasks with various rewards, the higher the reward the worse the performance.

2.2 The use of P4P in healthcare

Healthcare differs from the private sector in a number of ways which are relevant when discussing how P4P should be applied. These differences mean that some of the issues explored above become more relevant, such as multitasking and intrinsic motivation, and some new issues are created. In this section these new issues are introduced followed by a review of the literature regarding P4P in healthcare. A critique of this literature finds a number of areas lacking sufficient attention. These areas are the focus of this thesis.

2.2.1 Modifications to standard P4P theory

In most healthcare systems, agents have two principals: the central payer of healthcare and the patient. This concept, often referred to as *dual agency*, implies the doctor is caught between their own objectives and those of both of their principals (Ellis and McGuire, 1986). Patients might have a preference for more treatment if they do not personally bear the financial cost, whereas the payer is mindful of cost and therefore has a preference for less treatment. This is just one scenario which is dependent on how the healthcare system is financed. In systems where patients face out of pocket costs, preferences may be reversed. From the doctor's point of view, they have an obligation to the patient and to the payer. Resource constraints necessitate that providing care to one patient will impact on the amount of care which can be given to another, at least at the margin.

The utility function of a doctor differs from a traditional employee due to the presence of altruism, or a concern for the welfare of their patients, in addition to the traditional income and leisure components (Ellis and McGuire, 1986; Ma, 1994; Siciliani, 2009). This difference may influence how doctors use their time or target their services. Altruism is an example of a source of intrinsic motivation which may be eroded by financial incentives (Sicsic et al., 2012).

There are many factors which affect the health of patients, not all of which are within the control of the doctor. For example, a doctor may apply effort in order to improve the health of a certain patient but, if the patient continues to live an unhealthy lifestyle, the patient may still remain unwell (Balia and Jones, 2008; Fichera and Sutton, 2011). Due to this, paying on observable outcomes, such as deaths or readmissions to hospital, would penalise doctors who treat more patients with unhealthy lifestyles (Axon and Williams, 2011).

2.2.2 Evidence from P4P in healthcare

Despite the added complications regarding the use of P4P, it is becoming increasingly common in healthcare. In 2012 a summary of non-US P4P schemes identified 13 programs from nine countries (Eijkenaar, 2012). These included UK examples such as the QOF in primary care (Roland, 2004) and Advancing Quality in hospitals (Maynard and Bloor, 2010). The review was too early to identify more recent UK schemes such as hospital Best Practice Tariffs (Allen et al., 2014a; Department of Health, 2011).

A systematic review of systematic reviews identified 22 studies between 2000 and 2011 focusing on the various effects of P4P (Eijkenaar et al., 2013). The timing of the review means that an additional systematic review of P4P in healthcare was not necessary for this thesis. Instead, reviews published after 2011 were identified from key word searches (P4P, performance pay, financial incentives, healthcare and primary care) and forward citation searches of Eijkenaar et al., (2013). This resulted in four additional systematic reviews of P4P in healthcare.

One review focused on low and middle income countries and is not particularly relevant this thesis. They concluded that the evidence lacked robustness and conclusions about P4P could not be drawn (Witter et al., 2012). Huang et al. (2013) reviewed the evidence on the management of diabetes, a disease area that is incentivised heavily in the QOF as well as other P4P schemes. In total, 33 studies were identified and the findings were generally in support of P4P for diabetes management. However, the low quality of the studies meant that effects were likely to be overestimated.

The third review published since 2011 synthesised evidence only from incentives targeted at individual practitioners (Houle et al., 2012). The motivation behind this restriction was not explained and excluding papers on this basis seems unnecessary. In

most settings, healthcare is provided by teams not individuals, or in cases where individuals do provide healthcare they are supported by others in doing so. Furthermore, Houle et al. (2012) include several studies which analysed the effect of the QOF, which is a practice based incentive scheme and does not reward the performance of individual GPs. Therefore, any QOF study should have been excluded as per the stated criteria. As with many reviews of P4P, the authors concluded that insufficient evidence was available and more robust study was needed.

The three reviews discussed above came to the same conclusions about P4P as Eijkenaar et al. (2013) suggesting there has not been any substantial changes to the literature since Eijkenaar et al. The final study focused on the substantial body of evidence on the QOF (Gillam et al., 2012). The literature relating to the QOF is directly relevant to this thesis and is covered in more detail in section 3.3. The remainder of this section will focus on the review by Eijkenaar et al. (2013) while drawing on individual systematic reviews for additional insight.

Of the 22 reviews identified by Eijkenaar et al., many were either of low quality in terms of the review methodology or were not conducted recently enough to capture the growth in P4P occurring towards the end of the 2000s. Additionally, some reviews were restricted by a narrow focus on a particular topic or setting, such as nursing homes (Briesacher et al., 2009) or inequalities (Alshamsan et al., 2010). The review by Van Herck et al. (2010), which was both up-to-date and of high quality, identified the most individual studies (128).

The review by Scott et al. (2011) focused on robust evaluations of financial incentives in primary care between 2000 and 2009. Robust evaluations were those using randomised controlled trials (RCTs), controlled before and after studies, and interrupted time series analyses. Seven studies met these restrictions. Of these seven studies, five were set in the US, one in Germany and one in the UK. The focus was not on P4P but on financial incentives more broadly. The UK study compared capitation, fee for service and salary as means of provider reimbursement for primary care (Gosden et al., 2000). Fee for service and salary would not be considered examples as P4P as there is no explicit link between payments and quality. However, the studies from other countries evaluated the effect of P4P making the findings of the review relevant, albeit perhaps not within a UK context. Three of the remaining six studies evaluated P4P schemes which incentivised smoking cessation advice and

related areas (An et al., 2008; Roski et al., 2003; Twardella and Brenner, 2007). The other studies focused on a more varied set of indicators in diabetes, cancer screening and immunisation (Mullen et al., 2009; Rosenthal et al., 2005; Young et al., 2007). Overall the review found that across these seven studies there was a lack of evidence as to the effects of financial incentives on primary care quality. However, the generalisability of the studies included in the review should be considered. The selection of the most robust studies omits the entire literature on the evaluation of the QOF. The size of this literature means that, although the evaluation methods are not comparable to RCTs, patterns have emerged which can inform future policy makers on the likely impacts of P4P. Individual studies should indeed be treated with caution as each is unlikely to remove all types of potential bias. However, as the body of literature grows and studies are repeated using different methods and time periods, conclusions can be drawn from a collection of studies (Craig et al., 2012).

Van Herck et al. (2010) chose a longer time period (1990 to 2009) and less stringent criteria on study methods. Van Herck et al. also included studies from any area of healthcare and only studies using P4P. The definition of P4P precluded financial incentives without a clear link to provider performance, for example, fee-for-service and public reporting. These differences meant that 128 studies were identified and reviewed, 57 of these coming from the UK. The majority of extra studies resulted from cross-sectional analysis which is less robust when compared to RCTs or other methods utilising control groups.

The effects of P4P found by Van Herck et al. were categorised into five groups: clinical effectiveness; access and equity of care; coordination and continuity of care; patient centeredness; and cost-effectiveness. The inclusion of sections on coordination, continuity and patient centeredness is beneficial as these are often overlooked. However, the evidence from these areas is limited and weak. The review did not consider how doctors and other related staff were affected by P4P. This omission was unfortunate as the impact of P4P on staff was studied within the search period of the review and these studies have shown positive and negative effects which deserved attention (Campbell et al., 2008; McDonald et al., 2007; McGregor et al., 2008). The study in Chapter 6 of this thesis adds to this area of the literature.

The effect on clinical effectiveness of P4P schemes varied between negative, not statistically significant and positive in the reviewed studies. Despite the variation,

generally positive effects prevailed and the strongest effects were found for chronic conditions as opposed to acute care. Across 28 studies, using a range of evaluation methods, access and equity of care were not made worse due to P4P. The review concludes that there was increasing support for the cost-effectiveness of P4P schemes. This conclusion is at odds with other reviews which claim that there is insufficient evidence to make firm conclusions about cost-effectiveness due to low numbers of evaluations which adequately capture either the costs or benefits (Emmert et al., 2012).

P4P has mixed effects on patient centeredness and continuity of care but research was relatively sparse in these areas. Van Herck et al called for more research in these alternative measures of quality of care. There was an absence of evidence exploring if P4P was affected by patient characteristics. Van Herck et al. found no evidence about how patient awareness of a P4P scheme impacted on performance. Patient awareness and patient-reported quality of care is covered in Chapter 7 of this thesis.

Van Herck et al. (2010) differentiate themselves from other reviewers by assimilating the evidence in order to comment on the design and context of P4P schemes. This information was used to provide insight into what approach might work best for future P4P schemes, thus generating a valuable resource for policy makers. The evidence points to the use of process rather than outcome measures and to target these measures in areas where there is greatest room for improvement. The involvement of stakeholders, such as doctors or managers, has been shown to improve the effects of P4P schemes. Van Herck et al. fail to comment on if this involvement led to stakeholders selecting measures based on an anticipation of high performance and therefore biasing results. A preference was observed for schemes to be financed using additional rather than existing funds. Effective communication improved the impact of P4P, this has been evidenced more recently when providers responded to a policy's announcement in advance of the activity being financially incentivised (Allen et al., 2014a).

The final point on design highlighted some evidence that P4P was more effective when combined with other non-financial incentives such as public reporting but more research was called for in this area. The study in Chapter 4 of this thesis explores these issues empirically. In terms of context, the US system of fragmented providers and insurers generally results in less effective P4P schemes, owing to the dilution of

incentives when providers interact with multiple payers with different incentives. The same was not found for the UK where P4P schemes tend to be introduced system wide.

The literature reviewed by Eijkenaar et al. (2013) was categorised into six areas: effectiveness; cost-effectiveness; unintended consequences; inequalities; applications with non-financial incentives; and design features. Unlike Van Herck et al. (2010), a specific section on patients was not included in the review. Instead, topics such as the holistic nature of care and the continuity of care were only briefly mentioned in the section on unintended consequences. There was evidence of selecting patients based on risk and also negative spillovers onto areas not covered by P4P programs. Evidence of gaming and effects on intrinsic motivation was not generally found. Gaming is when P4P schemes are manipulated to give the impression of improved performance without real changes (Bevan and Hood, 2006)

A useful distinction made by Eijkenaar et al. (2013) was to group the effectiveness findings into those from non-random studies, predominantly cross-sectional or longitudinal in nature, and findings from more rigorous randomised studies. This helps to highlight that the bias associated with non-random studies tends to lead to positive but modest effects. Randomised studies were more rigorous and able to control for such biases, but results were inconsistent and conclusions could not be drawn. Similarly, persuasive evidence of cost-effectiveness was not found, though some individual studies did suggest the potential to be cost-effective. Eijkenaar et al. disagree with the conclusions of Van Herck et al. (2010) regarding the evidence in support of the cost-effectiveness of P4P schemes. There are too few studies which have captured the full costs of schemes, including setup costs as well as running and payment costs. The costing methods used in Meacock et al. (2014) should be utilised.

With respect to the effect of combining P4P with non-financial incentives, Eijkenaar et al. (2013) conclude that clear patterns were hard to establish due to various types of control and treatment groups. For example, some studies compared P4P and non-financial incentives to no incentives, while others compared P4P and non-financial incentives to P4P alone. Some studies showed improvements when incentives were combined, but more research was called for. The relationship between financial and reputational incentives is investigated empirically in Chapter 4 of this thesis.

2.2.3 Discussion of evidence

The reviews of Eijkenaar et al. (2013) and Van Herck et al. (2010) provide an extensive view of the literature on P4P and how this literature might impact on the designs of future P4P schemes. Most of the reviewed studies were set in either the UK or US and both sets of authors acknowledge that the institutional differences limit the generalisability of findings from one setting to another. As more evidence on the effects of P4P emerge, future reviews should consider grouping the findings according to the manner in which the healthcare systems are financed and structured. Eijkenaar et al. (2013) conclusions were more conservative than Van Herck et al. (2010). This difference could be the result of Eijkenaar et al. assimilating such a broad range of evidence that decisive conclusions were not obvious. Despite Eijkenaar et al. suggesting more caution when interpreting the evidence on P4P, both reviews conclude that more research is needed. The areas mentioned have guided the empirical studies of this thesis. Hence, our focus is to add to the literature on how patients and professionals perceive P4P and the relationship between financial and non-financial incentives.

3. THE QUALITY AND OUTCOMES FRAMEWORK

Each of the empirical chapters of this thesis uses the QOF as a source of data and a setting in which to evaluate P4P. This scheme is introduced and explained in this chapter. We have first provided some background to the development of the QOF, followed by details about the design and data. An overview of the evidence emerging from the QOF is also included.

3.1 Background

The role of the GP is central to the NHS. People living in the UK and wishing to use NHS services must register with a specific practice, usually one close to their homes. Patients will visit their GP according to their primary care needs and it is via this GP that elective, non-emergency, hospital treatments must be scheduled.

In the years before the introduction of the QOF there were growing fears about a GP recruitment and retention crisis, in part due to more attractive career opportunities in other medical specialties (Young and Leese, 1999). The sentiment of the time is surprisingly similar to comments about the current state of the GP workforce (Abbt and Alderson, 2014). A new plan for the NHS, released in 2000, aimed to address a number of the emerging problems in primary care (Department of Health, 2000). The plan involved providing a wider range of services in a primary care setting, improvements to patient access and choice, greater flexibility for GPs and an increase in the use of P4P. The new General Medical Services (GMS) contract was introduced in 2004 as a means to implement these plans (National Audit Office, 2008).

The new contract added a large P4P element to the existing income streams for GPs (Department of Health, 2003a). The resulting system meant that GPs' income comprised five elements: the global sum (capitation payments); provision of enhanced services and out-of-hours care; seniority; and performance on the QOF. The majority of GPs were in support of the QOF when it was introduced having been consulted on the scheme during the design stages (Smith and York, 2004). GPs' incomes were expected to increase to bring them more in line with those of hospital consultants. In addition to the new P4P elements, the contract also changed from being held by individual GPs to being held by the practice (Smith and York, 2004). The previous system meant that GPs were individually contracted with the NHS and hence enjoyed a high level of professional autonomy and freedom. Practices would also be able to

employ salaried GPs whose income would be set by the GP partners who own the practice.

3.2 Design and data

The initial design of the QOF has been explained by academics (Roland, 2004; Smith and York, 2004) and in policy documents (Department of Health, 2003a, 2003b). The scheme rewarded practices with points based on their performance across four domains: clinical; organisation; additional services; and patient experience. The clinical domain consisted of 76 indicators in 11 chronic disease areas worth 550 points. The organisational domain consisted of 56 indicators in five areas relating to how well the practice organised its records, communications, training, and management. This domain was worth 184 points. Additional services included cervical screening and contraceptive services, and were worth 36 points across 10 indicators. Patient experience was worth 100 points and assessed on four indicators in two areas, a patient survey and the length of consultations. A further 180 points were available based on the depth of quality. These included holistic care (100 points), overall achievement (30 points), and patient access (50 points) (Health & Social Care Information Centre, 2005a). The maximum points a practice could earn was 1050.

The Health & Social Care Information Centre, (2004) provide examples of indicators and their definitions from each domain:

- Clinical (records): The practice can produce a register of patients with coronary heart disease. 6 points
- Clinical (diagnosis and initial management): The percentage of patients with hypertension whose notes record smoking status at least once. 10 points
- Clinical (ongoing management): The percentage of patients with diabetes who have a record of retinal screening in the previous 15 months. 5 points
- Organisational (Records and information about patients): There is a system to alert the out-of-hours service or duty doctor to patients dying at home. 2 points
- Organisational (Patient communication): If an answering system is used out of hours, the message is clear and the contact number is given at least twice. 0.5 points
- Organisational (Education and training): All new staff receive induction training. 3 points

- Organisational (practice management): Person specifications and job descriptions are produced for all advertised vacancies. 2 points
- Organisational (medicine management): There is a system for checking expiry dates of emergency drugs at least on an annual basis. 2 points
- Patient experience: The length of routine booked appointments with the doctors in the practice is not less than 10 minutes. 30 points
- Patient experience: The practice will have undertaken an approved patient survey each year. 40 points
- Additional services (cervical screening): The practice has a system to ensure inadequate/abnormal smears are followed up. 3 points

As the practice increases their performance on these indicators, they are rewarded with more points. Each point was worth £75 to the average practice in 2004/5 with the value of a point increasing in later years. The precise value of a point is determined by two adjustments: the adjusted disease prevalence factor (ADPF) and the contractor population index (CPI). These features increase the value of a point for practices with higher disease prevalence and larger lists of patients. They are discussed in more detail in Section 3.2.2.

The effect of the scheme on GP incomes was significant, due to practices scoring very well across all domains. On average 95.5% of total points were achieved in the first year (Doran et al., 2006). Between 2003/4 and 2005/6 GP incomes increase from £85,000 to £114,000 (National Audit Office, 2008).

In most years since its introduction, the QOF has been changed and details of these changes are documented online (NHS Employers, 2015). These changes range from minor revisions to the names of indicators, to more substantial changes like the increase in the lower threshold on performance from 25% to 40% for all but one indicator in 2006/7. The change in lower thresholds was motivated by the high levels of achievement by practices. Other common changes are the removal of indicators to free up points for new indicators. This helps the QOF to remain up to date (Reeves et al., 2010). Some have argued that the changes made to the QOF should have been informed by the past performance of practices. Instead, most changes were inconsistent and took no account of actual performance (Doran et al., 2014).

3.2.1 Raw data

QOF data are publically available to download from the Health & Social Care Information Centre, (2015). The data are separated into individual years, ten at the time of writing, and can be downloaded in various levels of aggregation: national; Strategic Health Authorities (replaced by NHS England in 2013); Primary Care Trusts (replaced by Clinical Commissioning Groups in 2013); and practice level. Practice level data are used in each of the studies presented.

The data and key variables which are used in multiple chapters are presented here. Table 3.1 [p.163] and Table 3.2 [p.164] summarise this information.

3.2.1.1 Raw practice performance

For each of the QOF domains, data are provided on the performance of all practices. For the organisational, patient experience and additional services domains performance is measured by points alone. For indicators in the clinical domains, more information on performance is provided. In addition to points, the numbers of treated patients and eligible patients is given for each indicator. Treated patients is defined as the number of patients which received the treatment in line with the specific indicator, for example having blood pressure taken. Eligible patients is defined as the number of patients with the appropriate disease(s) minus patients who were excluded or exception reported out. Exclusion of a patient is done on the grounds that, although they had the disease in question, they did not meet some other criteria, such as age. Exception reported patients are those that had the disease and met the other criteria but were removed from the group of eligible patients on other grounds, such as having comorbidity which made the indicator inappropriate or if the patient refused treatment. In 2005/6 the exception rate for patients was 5.6% (Health & Social Care Information Centre, 2005b). At the clinical indicator level, the number of exception reported patients has been available for download since 2005/6. Figures for exceptions in 2004/5 are not available. Exception figures can be added to the numbers of eligible patients to create a measure of practice population performance, irrespective of the level of exceptions. This adjustment may be necessary as exception reporting varies by practice (Doran et al., 2012).

3.2.1.2 List size and Disease prevalence

The number of patients registered with a practice is known as its list size. This variable is included in the QOF data for all practices. Registers of patients are

provided which count the number of patients with a disease covered by a QOF indicator. Disease prevalence measures are included which were calculated as the percentage of the practices list size on a particular disease register: $\left({}^{r_{idt}}/_{l_{it}}\right)$ where r_{idt} is the register and l_{it} is the list size for i indicators, d diseases and t time.

These variables were used to make adjustments to QOF payments for clinical indicators based on a practice's disease prevalence and list size.

3.2.2 Derived variables

Several important variables are not provided directly in the QOF data and have to be derived from the available data.

3.2.2.1 Derived practice performance

Practice performance is calculated as the percentage of eligible patients which received treatment for each indicator. This results in a comparable measure across indicators and practices.

Algebraically,

$$Y_{kidt} = {}^{\tau_{kidt}}/_{(r_{idt} - [E^1_{kidt} - E^2_{kidt}])} \qquad 3.1$$

Where Y denotes the measure of practice performance. τ denotes the number of patients treated, varying by k, i, d and t (indicators, practices, diseases and time respectively). r is the disease register. E^1 and E^2 denote the indicator specific excluded and excepted reported patients respectively. The performance variable is represented as a percentage for all analyses.

3.2.2.2 Performance thresholds

The upper and lower thresholds determine an interval of patients, above and below which no points are earned. The lower threshold ensures that a certain level of quality is provided by the practice before any payments are made and therefore promotes a minimum standard of care. The upper threshold is used to prevent practices over treating patients. The relationship between the number of points earned by a practice and practice performance is linear between the lower and upper threshold. This is depicted in Figure 3.1 [p.209].

The threshold values are not provided directly in the QOF data but can be sourced from the GMS contract and the various updates to this document (BMA and NHS Employers, 2012, 2011, 2009, 2008, 2006; Department of Health, 2003a).

3.2.2.3 Adjusted Disease Prevalence Factor (ADPF)

In the negotiation prior to the start of the QOF it was decided that adjustments should be made to payments in order to better reflect the workload involved with treating a certain percentage of a practice's patients with a disease (Guthrie et al., 2006). The argument in favour of the adjustment was that practices with higher than average disease prevalence would have to treat more patients, in absolute figures, to reach the lower threshold when payments started. They should therefore be compensated for the additional effort this involved. The resulting ADPF is applied to only clinical indicators, and changes were made to the definition of the ADPF in 2009/10 and in 2010/11.

In the initial iteration of the ADPF, very low prevalence rates were truncated by means of first calculating 5% of the range of prevalence rate for each disease. If a practice's prevalence rate was below this value, then the prevalence rate would be changed to this value. This increased the prevalence of practices with rates of zero or close to zero.

This truncated prevalence rate is then square rooted and divided by its disease-year mean to create an Adjusted Disease Prevalence Factor (ADPF) (NATPACT, 2004). The ADPF takes the value of one for practices with average prevalence rates. An ADPF greater than one means the practice has higher than average prevalence rates and is therefore compensated, in terms of revenue, for having more patients with certain diseases. Similarly, an ADPF less than one means the practice receives less revenue because they have fewer patients with these diseases.

The ADPF calculation changed twice during our period of analysis. In 2009/10 (the sixth year of the QOF) the square root transformation was removed. In 2010/11 (the seventh year) the truncation was removed. Both of these changes were announced in 2009 (NHS Employers and BMA, 2009).

Algebraically, the ADPF for 2004/5 to 2008/9 can be written as:

$$ADPF_{idt} = \sqrt{\text{trunc}(p_{idt})} \Big/ \overline{\sqrt{\text{trunc}(p_{dt})}} \quad 3.2$$

In which p_{idt} is the prevalence rate for practice i for disease d in time t. Disease prevalence is defined as the practice disease register divided by practice list size, that is, r_{idt}/l_{it}. Where r_{idt} is the register and l_{it} is the list size.

For 2009/10, the ADPF can be written as:

$$ADPF_{idt} = \text{trunc}(p_{idt}) \Big/ \overline{\text{trunc}(p_{dt})} \quad 3.3$$

Where the truncation of the prevalence rates requires that:

$$trunc(p_{idt}) = \max(p_{idt},\ 0.05\ (max_{dt}(p_{idt}) - \min(p_{idt})) \quad for\ t = 1,2,\dots,6 \quad 3.4$$

For 2010/11 to 2012/13 as, the ADPF can be written as:

$$ADPF_{idt} = p_{idt}/\bar{p}_{idt} \quad 3.5$$

3.2.2.4 Contractor Population Index

Payments are also adjusted according to the list size of the practice, with larger practices receiving greater payments. The list size adjustment involves creating the Contractor Population Index (CPI). The CPI is the practice list size divided by 5,891, the national average list size in 2003.

$$CPI_{it} = l_{it}/\mu \quad 3.6$$

Where l_{it} is the list size and μ is 5,891.

3.2.2.5 Maximum points and the value of a point

Indicators have a maximum point allocation and these values can be sourced from the contract documents. The maximum point allocation is the number of points the practice received once they reach the upper threshold of performance. These values are used to calculate the revenue a practice could earn if they maximised performance.

The monetary value placed on each QOF point is used to calculate practice revenue. These values are not provided directly but can be found in policy documents. In the first year (2004/5) a point was worth £75 which subsequently rose to £124.60 for

2005/6 to 2008/9 (National Audit Office, 2008). The value rose to £126.77 for 2009/10 and 2010/11 (NHS, 2010). In 2011/12 and 2012/13 the value was £130.51 and £133.76 respectively (BMA, 2011; NHS Employers and BMA, 2011).

3.2.2.6 Revenue

The revenue each practice receives from the QOF is not publically available. It is a function of the elements discussed above: points, ADPF, CPI and the value of a point. Using these elements it is possible to measure the actual income a practice receives and the maximum income a practice could receive. These two values differ as actual income is based on the observed performance of the practice and gives the income they received, whereas maximum income assumes the practices achieved 100% of the available points.

Revenue is calculated for each indicator individually and then summed over all indicators to arrive at practice income.

The expression for the revenue per indicator is

$$Revenue_{kidt} = \pi_{kidt} * ADPF_{idt} * CPI_{it} * \alpha_t \quad 3.7$$

Where π denotes the maximum points available and varies over indicator k, practice i, disease d and time t. α denotes the value of a QOF point which varies only over time. For non-clinical indicators the ADPF would take a value of one indicating that no adjustments were made for these indicators. List size adjustments are still made for non-clinical indicators.

We can illustrate this expression with a worked example. In 2004/5 the maximum number of points available for indicator BP4 was 20 points. This indicator rewarded practices for recording the blood pressure of patients with hypertension and had a lower threshold of 25% and an upper threshold of 90%. A practice with average hypertension prevalence and a list size of 5,891 who treated ≥90% of eligible patients would have received a revenue of £1,500 for this indicator (20 points * £75 * 1 * 1, since the CPI_{it} and α_t are both equal to one). Larger practices and those with higher prevalence would have received more income in line with these differences.

Total practice revenue is the sum of indicator income over all K indicators. The number of indicators varies each year.

$$Revenue_{it} = \sum_{k=1}^{K} (\pi_{kidt} * ADPF_{idt} * CPI_{it} * \alpha_t) \quad 3.8$$

3.3 Literature on the effect of the QOF

The QOF is the largest and longest running P4P scheme of its type, and as a result has been the subject of a substantial body of research which is reviewed in this section. Our approach is to rely on the three published systematic reviews of the QOF for the majority of the evidence (Gillam et al., 2012; Langdown and Peckham, 2014; Steel and Willems, 2010). With the most recent review being published in 2014, these reviews have covered most of the period in which the QOF has been in use. However, there have been a number of significant studies either not included in the reviews due to some inclusion or exclusion criteria, or published after these reviews. Therefore, in our review of the QOF, we supplement the evidence from systematic reviews with evidence from selected studies not covered in the reviews. We selected these additional studies based on our knowledge of the literature and forward citation searching of existing systematic reviews.

The first systematic review selected papers that used the QOF from previous reviews of financial incentives and added to these using the results from electronic literature searches (Steel and Willems, 2010). Their search period ended at January 2010 and 35 papers in total were identified. The second review include an additional 18 months (up to July 2011) in their search period and identified 94 studies in total (Gillam et al., 2012). The increase in the number of identified studies represents the selection criteria being more inclusive of cross-sectional studies, which are common in evaluations of the QOF, and the longer search period. The most recent review of the QOF included studies up to June 2012 (Langdown and Peckham, 2014). The review included only those studies which focused on QOF clinical indicators and met stringent inclusion criteria. This resulted in only 11 studies being included.

The approach of Gillam et al. (2012) is arguably the most appropriate as they include the most papers and hence are better able to reflect on the evidence of the scheme. The lack of a trial period or control group is a fault in the design and implementation of the QOF. To restrict a review of the QOF to only include studies which use before and after data is misrepresentative of the evidence base. Studies using cross-sectional data should be included, albeit with appropriate caveats on the evidence emerging from them. The three reviews have grouped the literature such that there are some

overlapping areas. The evidence from these areas will be compared and contrasted below, and where appropriate, new research will be mentioned.

Steel and Willems (2010) included a section specifically on the performance of practices using QOF data, as opposed to practice performance measures outside of the QOF. Performance was high from the first year and increased thereafter. Neither of the other reviews include such a section although Gillam et al. (2012) do mention that the increased use of computers has resulted in better recording of care.

Health outcomes are given a section in Steel and Willems (2010) and Langdown and Peckham (2014), while Gillam et al. (2012) incorporates these issues into a section on effectiveness. The QOF evidence relating to health outcomes largely comes from performance on intermediate health outcome targets such as controlling cholesterol or blood pressure within defined limits. All reviews come to the conclusion that performance increased initially only to plateau by 2007.

Since the reviews were published, the quality of care for diabetes and hypertension has been investigated further. For diabetes the same initial increase in performance followed by a plateau was found (Kontopantelis et al., 2013b). However, for hypertension, performance was increasing before the QOF and continued to do so afterwards. There was no change in the rate of this increasing quality attributable to the QOF (Serumaga et al., 2011). The effect of the QOF on population mortality has also been investigated recently (Kontopantelis et al., 2015). The study looked at all-cause mortality as well as mortality in areas incentivised by the QOF and found no relationship between practice performance and reduced mortality. The conclusion of the existing reviews is unlikely to change given these more recent studies. However, it has become clearer that although the QOF resulted in some initial improvements in some intermediate outcomes, it has yet to be shown to have reduced mortality.

The cost-effectiveness of the QOF is an important feature to establish. The high cost of the scheme has meant that expectations about resulting health improvements are also high. Langdown and Peckham (2014) do not review any studies which evaluate the cost-effectiveness of the QOF. Steel and Willems (2010) and Gillam et al. (2012) do review the evidence on the efficiency and cost consequences of the QOF but the evidence in this area was very limited. At the time of the review, Gillam et al. (2012)

identified one study showing the potential for QOF performance to reduce epilepsy hospital admissions.

Since the reviews several studies have explored the link between improved QOF performance and reduced hospital admissions for various conditions. Establishing this link is important as it represents a likely mechanism for the QOF being cost-effective. A small effect of reducing stroke hospital admissions was found for reaching the QOF targets for cholesterol (Soljak et al., 2011). Practices performing better on diabetes management indicators had fewer patients with diabetes related hospital admissions (Dusheiko et al., 2011a). A similar link between improved primary care management and lower emergency admissions and outpatient appointments was observed for stroke patients (Dusheiko et al., 2011b).

Some ambulatory care sensitive conditions (ACSC), those for which appropriate primary care treatment should prevent hospital admissions, are incentivised in the QOF and some are not. Conditions with incentives had lower hospital admissions than those without (Harrison et al., 2014). A positive relationship was found between patients receiving mental health checks and hospital admissions for related conditions (Jacobs et al., 2015). This is counter to the research noted above which suggests a negative relationship between primary care quality and hospital admissions. However, the authors explain that this is likely to do with the timing of hospital admissions and primary care visits; a hospital admission is likely to highlight the need for mental health checks, resulting in a primary care visit. This area of the QOF literature relating primary care quality to hospital admissions has been missed by systematic reviews due to the relatively recent nature of these studies. Taken together, they point out a potential mechanism by which the QOF can help reduce healthcare costs and therefore this literature is vital in any attempt to quantify the costs and benefits of the QOF.

The design of the QOF results in significant financial incentives for certain care processes and diseases, while leaving others without additional incentives. This design has prompted an interest in whether non-incentivised areas are the subject of negative or positive spillovers. The reviews by Steel and Willems (2010) and Langdown and Peckham (2014) include sections on the effects of the QOF on non-incentivised areas while Gillam et al. (2012) mention these issues on the context of effectiveness. Early findings show performance on non-incentivised areas was well below that of incentivised areas (Steel and Willems, 2010). This is most likely due to incentivised

areas receiving policy attention from other sources and not a direct result of the QOF incentives. However, it has been shown that by 2007, performance in non-incentivised areas was below what pre-QOF trends would have predicted (Campbell et al., 2009; Doran et al., 2011). This would suggest that non-incentivised areas were neglected due to GPs focusing on the incentives. Another study, not mentioned by any of the reviews, found evidence of positive spillovers onto other, non-incentivised, aspects of care for patients who had a disease with some QOF incentives (Sutton et al., 2010).

The QOF is not designed to reduce health inequalities, it is designed to standardise care. However, standardising care has the potential to impact on health inequalities. In theory the incentives to treat different groups of patients are the same, for example there is no difference between men and women in terms of payments. However, the workload needed to treat different groups of patients may differ, therefore creating an incentive to target those most easy to treat. Steel and Willems (2010) and Gillam et al. (2012) show that there is good evidence that socioeconomic inequalities in care have lessened during the QOF but inequalities in age, gender and ethnicity have increased. The reviews did not include the study by Norbury et al. (2011) which analysed inequalities in influenza immunisation. They found that age inequalities persisted following the QOF and so too did inequalities in socioeconomic status. Conclusions about the impact of the QOF on inequalities seem sensitive to the studied area, but it is likely to have reduced inequalities in some areas.

The QOF represented a major change to the GP contract and had significant impacts on the working lives of GPs. Steel and Willems (2010) and Gillam et al. (2012) both acknowledge the importance of the effect of such high powered incentives on professionals. The QOF led to an increased role for practice nurses, who were heavily involved in the treating of QOF patients, but who were often not the recipient of the financial incentives (Campbell et al., 2008). It has also been highlighted that the pressure to meet QOF targets has reduced continuity of care and the doctor-patient relationship (Gillam et al., 2012; Steel and Willems, 2010). More recent qualitative research into the effect of the QOF on health professionals has revealed a more positive outlook on the scheme (Lester et al., 2013). The feeling that the QOF has eroded clinical autonomy and professionalism remained, but there was greater acceptance of P4P and a willingness to work with the scheme to improve it.

Only Gillam et al. (2012) includes a section reviewing the literature concerning how patients have been affected by the QOF. Patients report a similar experience to GPs: they felt a reduction in the continuity of care following the QOF. This fits with the general idea that the QOF focuses too heavily on completing small tasks which detract from holistic patient care (Maisey et al., 2008). There is a lack of research in this area which explains why more cannot be said about the experience of patients.

3.3.1 Discussion of evidence

In recent years, probably prompted by the QOF turning 10 years old, a number of opinion articles have emerged discussing the history of the QOF and what the future might hold. Raleigh and Klazinga (2013) highlight some of the recent changes to the QOF which include the removal of the organisational domain, a new public health domain and new rules for increasing thresholds in line with practice performance. The changes are the largest since the scheme was introduced and have not been universally adopted by all health departments in the UK.

For 2014/15, 338 points have been removed from the QOF and the income attached to them reallocated into capitation funding (BMA et al., 2014) The movement of primary care funds is apparently welcomed by GPs (Roland and Campbell, 2014). The reallocation of funds seems to be motivated by an acceptance that the single disease approach of the QOF, which does not lend itself to multi-morbid or comorbid patients, is less appropriate for addressing complex cases or the determinants of health, such as obesity (Gillam and Steel, 2013). Roland and Campbell (2014) propose that a share of income coming from P4P exceeding 10% is likely to result in higher risks of unintended, negative consequences. Others have also suggested that the income share from the QOF is too high (Gillam and Steel, 2013).

Together these systematic reviews and opinion pieces highlight the sustained interest in P4P, particularly the QOF, which does not seem to be wavering as the scheme ages. The recent changes to the scheme mark the beginning of a second phase to the QOF. The early research focused on the effects of the introduction of the scheme, which was appropriate at the time. However, as the scheme is now an established feature of primary care and unlikely to be removed, research should shift away from the clinical effects in the early years and on to areas currently under researched. In concluding their systematic reviews, Steel and Willems (2010) and Gillam et al. (2012) suggest that these areas include the effects of the size of payment and how large these

payments should be, as well as topics addressing the patient experience of the QOF. These conclusions are similar to those of Van Herck et al. (2010) and Eijkenaar et al. (2013) who suggest a focus on patients and professionals and the interaction of financial and non-financial incentives. We add to these areas in the following four chapters.

4. REVENUE OR REPUTATION – A QUANTITATIVE ANALYSIS OF GENERAL PRACTITIONER MOTIVATION USING LONGITUDINAL DATA

4.1 Introduction

An often revisited question is how best to motivate providers of healthcare to increase effort and improve quality (Appleby et al., 2012). Some of the most common methods which purchasers have used are reputational and financial incentives (Lindenauer et al., 2007). Many financial incentive schemes in healthcare have an element of public reporting of provider performance. The public reporting of performance can create incentives to increase performance by inducing comparisons between providers (Fichera et al., 2014). The use of published waiting times for hospitals in the UK NHS is an example of reputational incentives which led to reduced hospital waits (Propper et al., 2008). Hospitals were expected to apply effort to avoid being named and shamed. Linking performance targets to financial bonuses was an example of a mainly financial incentive for US physician groups (Rosenthal et al., 2005).

A further example from the US paid hospitals a bonus for being placed in the top decile. This represents the clear combination of both reputational incentives, via public reporting of quality, and financial incentives (Ryan, 2009). This latter example highlights a common theme in many financial incentives; they are often coupled with some form of public reporting on the performance that they incentivise. Public reporting creates a potential reputational incentive in cases where providers of healthcare value their own reputation and act to improve it.

Reputational incentives can be problematic for researchers wishing to analyse the effects of the financial incentive as it becomes difficult to isolate the incentives created by public reporting and those of P4P. It is likely that any improvements in performance are motivated in part by financial incentives and in part by reputational ones. Understanding the mechanisms for motivation could shed light on the most efficient approach to design a financial incentive scheme. For example, should reputation incentives work to complement financial incentives, this could reduce the financial burden of a scheme in comparison to a scheme where financial incentives alone matter. It may also be the case that reputational and financial incentives conflict with one another, similar to how extrinsic incentives may weaken intrinsic motivation

(Ariely et al., 2009; Deci et al., 1999; Frey and Jegen, 2001). Under this scenario, financial incentives might impede the motivation from reputational incentives.

As discussed in Chapter 2, healthcare is often thought as different to other employment sectors in terms of the appropriateness of P4P schemes. This is due to task complexity (Eggleston, 2005; Kaarboe and Siciliani, 2011) and the doctor's utility function (Gravelle et al., 2002; Iversen and Ma, 2011). P4P schemes in healthcare tend to approach these issues through the use of mixed payment schemes, which financially incentivise performance where it can be measured, while still taking advantage of provider altruism and other sources of motivation such as reputation.

In practice, incentive schemes in healthcare tend to vary in the degree to which they adopt financial or reputational incentives. A spectrum can be considered with pure financial incentives at one end and pure reputational incentives at the other (Oliver, 2015). The evidence reviewed by Oliver (2015) suggests that the most effective schemes come from the combination of public reporting of performance and modest financial rewards. However, a consensus on what constitutes *performance* is rarely reached between providers and purchasers (Doran, 2015). The lack of a universally accepted definition of quality can result in doctors feeling torn between meeting the requirements of the incentive scheme and those of the patient. Oliver (2015) suggested allowing doctors a voice when designing incentive schemes as a possible solution, but in the past doctors have used such a voice to keep performance targets below the average provider achievement (Doran et al., 2014). A balanced must be struck between an effectively designed scheme that will drive improvement, and one which is accepted by doctors.

This chapter seeks to improve our understanding about the effects of combining financial and reputational incentives. Literature on this area is scarce and the need for more research has been highlighted in systematic reviews of P4P (Eijkenaar et al., 2013). We contribute to this area by using an unusual feature of the QOF, which was amended part way through our sample, whereby the revenue from activity is neither proportional to the required activity nor to the marker of reputation, the QOF points (Guthrie et al., 2006). The quality scores for individual practices are published and comparison between local practices can be made on the NHS Choices website (NHS, 2015). These scores therefore represent an externally-observed marker of quality with peers in other practices, with the PCT and with patients (Santos et al., 2015).

The revenue per point varies between diseases within practices, between practices within diseases, and over time across both dimensions. This distinct feature of the scheme allows for the reputational and the financial incentives to be measured separately. We estimate the effect each incentive had on practice performance to establish the relative importance of revenue versus reputation in determining performance. We are also able to track how these incentives interact with one another over a nine year period, further informing the literature on the effects of variations within these incentives over time.

We find that reputational and financial incentives can be separated. In the early years of the QOF, financial incentives have a larger effect on performance than reputational incentives. Overtime, perhaps due to the formation of benchmarks in performance, reputational incentives become more important in determining performance. These reputational incentives have the potential to increase performance and have low marginal cost.

This chapter is structured as follows. Section 4.2 discusses the data used. Sections 4.3 and 4.4 detail how incentives are measured and our econometric approach respectively. Our results are presented in Section 4.5. We conclude in Section 4.6 with a discussion and the implications for research and policy.

4.2 Data

Data are obtained via the Health and Social Care Information Centre (HSCIC) website (Health & Social Care Information Centre, 2015a). All data pertaining to QOF performance and revenue are either available directly from the HSCIC or can be generated from the available data. Practice level QOF data are downloadable for all clinical indicators for the financial years 2004/5 to 2012/13. These include the points awarded, the number of patients treated, the number of eligible patients, the lower and upper thresholds, disease prevalence and practice list size.

For all nine years the number of patients per GP was obtained from the HSCIC. The Index of Multiple Deprivation of the practice population was available for three years (2004, 2007 and 2010). This index was applied to the other years on the basis that 2004 can be used for 2004/5 and 2005/6; 2007 can be used for 2006/7, 2007/8 and 2008/9; and 2010 can be used for 2009/10, 2010/11, 2011/12 and 2012/13.

For 2004/5 to 2008/9 the following were obtained: average age of GPs in the practice; proportion of UK qualified GPs; proportion of GP partners; proportion of female GPs; and the age-gender composition of the practice population.

The units of analysis are practice-indicator combinations. As indicators were added and removed over the period, and practices opened and closed, the panel is not balanced. There are 7,686 practices appearing in all years, a total of 8,929 unique practices and an average of 8,345 practices per year. There are 42 indicators appearing in all years, a total of 97 unique indicators and an average of 65 indicators in each year. This results in 4,881,825 observations.

4.3 Measuring incentives

We focus on only the clinical indicators as these represented the majority of revenue available. Revenue from these indicators was also a function of points earned, practice size and practice disease prevalence. The majority of non-clinical indicators rewarded practices with a fixed amount of points for doing a certain activity, such as staff training, and did not utilise lower and upper thresholds or the prevalence adjustments. The non-clinical indicators did not have different financial and reputational incentives.

We treat the revenue as a financial incentive and points as a reputational incentive. They are separate rewards for the effort of treating patients. However, as these incentives are different functions of effort, they rewarded the effort of treating patients differentially. All points do not have the same *going rate* in terms of their financial reward because of the ADPF and the CPI. Once the revenue from a certain indicator has been separated from the points, what remains is an external marker of practice quality. Having stripped out the revenue incentive from points, any observed relationship between an indicator's points payoff and practice performance must be in response to an incentive other than the financial incentive. Points are published online and act as a standardised measure of performance; standardised in the sense that they account for differences in practice size and disease registers through the treated/eligible measure of performance. Given these qualities, and the fact that points come with an intrinsic association with quality, increasing points would be perceived as increasing reputation.

In Section 3.2.2.5 we describe how QOF revenue can be calculated using the available data and information about the design of the QOF: specifically the ADPF, CPI and

value of a point. We use the expression for indicator-level revenue shown in Equation 3.7 as a starting point to calculate the revenue incentive for each indicator.

To determine if practices responded differently to these separate incentives, we measured the payoff in terms of points and revenue as:

$$P = {}^{\pi_{kidt}}/_{[(u_{kdt} - l_{kdt}) * D_{kidt}]} \quad 4.1$$

And

$$R = \frac{[reveneue_{kidt}]/[(u_{kdt} - l_{kdt}) * D_{kidt}]}{\alpha_t} \quad 4.2$$

Where π is the maximum points for indicator k, practice i, disease d and time t. The terms u and l represent the upper and lower thresholds for each indicator. D measures the number of eligible patients in each practice, for each indicator. $(u_{kdt} - l_{kdt}) * D_{kidt}$ quantifies the number of patients needed to be treated for a practice to move from the lower to the upper threshold. This is illustrated in Figure 3.1 [p.209].

Revenue is equal to points * ADPF * CPI * α_t as in Equation 3.7. The revenue earnt for each patient between the lower and upper threshold is divided by the value of a point in that particular year, α_t (£75 in 2004/5 for example). This rescales the revenue incentive and puts it on the same scale as the reputational incentive. That is, the two incentives are now comparable. A value of 0.5 for both incentives means half a point per patient versus half *the value of a point* per patient in that year. The points and revenue payoffs can be thought of as the marginal reward per patient treated.

Differences in the ADPF and CPI result in practices having different revenue and points payoffs for the same indicator. Within a single practice, differences in the ADPF across indicators can result in a practice having different rewards for the same workload. The implication is that there exist different incentives for treating the same number of patients.

If a practice is more concerned with their reputation, a higher revenue incentive will not be associated with improved performance. Practices wishing to maximum their own revenues will be expected to identify and apply more effort to the indicators where these differential incentives create a larger revenue payoff.

That there are common elements in the measurement of revenue and reputational incentives may raise concerns that these variables are too closely related to include both in a single regression. The degree of multicollinearity between these two variables is tested using the variance inflation factor (O'Brien, 2007)

4.4 Econometrics methods

To test how practice performance responds to different revenue and reputational incentives we estimate the following regression.

$$Y_{kidt} = \beta_0 + \beta_1 R_{kidt} + \beta_2 P_{kidt} + \beta_3 U_{kdt} + \beta_4' \boldsymbol{X}_{it} + \delta_1 T_t + \epsilon_{kdi} + \varepsilon_{kidt} \qquad 4.3$$

Where k, i, d and t index indicators, practices, diseases and time respectively. Y is the measure of practice performance, the number treated patients as a percentage of eligible patients, and is explained further in Section 3.2.2.1. R and P are the payoff variables for revenue and points. U is the upper threshold for each indicator, $\boldsymbol{X}$ is a set of practice characteristics and T is a set of yearly dummies for 2004/5 to 2012/13. ϵ_{kdi} is an indicator-practice specific time-invariant heterogeneity term and ε_{kidt} is an idiosyncratic error term.

The upper threshold U is included as a benchmark of performance that practices may respond to (Kontopantelis et al., 2012). It is a strong determinant of practice performance. $\boldsymbol{X}$ is included to control for practice characteristics which affect performance. For example, larger practices with more administrative staff may perform better. T is included to control for yearly changes in performance which may be due to practices gaining experience of the scheme.

Due to data availability, the elements of $\boldsymbol{X}_{it}$ differ in the first five years compared to the last four years. For years one through five, the elements are: measures of social deprivation; list size; number of patients per GP; average age of GPs; proportion of UK qualified GPs; proportion of GP partners; proportion of female GPs; and the age-gender composition of the practice population. For years six through nine, the elements are restricted to list size and number of patients per GP.

The practice-indicator unobservable term (ϵ) allows for any time-invariant characteristic which affect how practices perform on different indicators. For example, certain practices may perform better on indicators involving the measurement of

processes or perhaps better on those involving patient communication for unobservable reasons. Equation 4.3 was estimated using fixed effects to remove the effects ϵ_{kdi} may have on the remaining parameters in the model. The fixed effect in the model is at practice-indicator level to allow for practices to differ in their unobservable quality/effort at an indicator level. This is the most flexible specification possible given the structure of the data.

To test whether the relationship between incentives and performance differed for stable indicators compared to transient indicators, separate regressions were estimated on samples of balanced and unbalanced indicators. Subsamples are also selected to separate the years before and after changes to the ADPF.

The ADPF and the CPI inject variation into the payoff variables such that they differ, and the models can be estimated. Practices in a given year will differ in their ADPF due to variation in their disease prevalence rates. As these disease prevalence rates change over time, more variation is added. This is in addition to changes in the calculation of the ADPF which affect practices to a varying degree depending on where in the prevalence distribution they lie. For example, the removal of truncation only affects those at the bottom of the distribution, and the removal of square rooting would not have affected practices with average disease prevalence.

Within practices, the ADPF varies over diseases, meaning that the revenue can fluctuate for indicators from different diseases, even if the workload is the same. To a lesser degree, the CPI varies over time and by practice but does not differ within a practice in a given year. The benefit of this variation is that a flexible specification for fixed effects can be used, which allows for unobserved differences at the practice-indicator level. With this specification of the fixed effects term it is not possible to estimate the models for a single year, due to a lack of variation. Similarly, for one year and one disease area, each practices' revenue payoff would be a perfect linear transformation of the points payoff.

It is plausible that during the nine years we observe performance, the response to the incentives may change. This may be driven by changes in the design of the QOF, such as the ADPF, or it may be driven by changes in practice behaviour. In order to test if there is a change in the response to incentives we estimate the following model

$$Y_{kidt} = \beta_0 + \beta_1[T_t * R_{kidt}] + \beta_2[T_t * P_{kidt}] + \beta_3 U_{kdt} + \beta_4' \boldsymbol{X}_{it} + \delta_1 T_t + \epsilon_{kdi} + \varepsilon_{kidt} \quad (4.4)$$

Where elements are defined as in Equation 4.3. Each incentive variable has been interacted with a set of time dummies which will estimate the response separately for each year.

4.4.1 Sensitivity analysis

To test the sensitivity of our incentives analysis shown in Equations 4.3 and 4.4 we perform several additional regressions.

Firstly, we perform a set of seven regressions which sequentially add additional control variables. Only the first five years of data are used for these regressions so it is possible to assess the importance of missing control variables in later years. We begin by including only the points payoff and then add the following in each subsequent regression: revenue pay off only; both payoff variables; upper threshold; controls available in all years; controls available in years one to five; and a balanced set of indicators. The purpose of these regressions is to quantify the effect of the inclusion of different variables on the estimated practice response.

Secondly, we perform three regressions using alternative specifications of the fixed effects term from Equation 4.3. Models are estimated with fixed effects at the practice level, at the practice-disease level and at the practice-indicator level. The practice-indicator level is our preferred specification as it is the most flexible, allowing practices to differ in their performance on indicators. These additional regressions determine if there are differences in our estimated effects due to the richness of the fixed effects.

Thirdly, we estimate the models using the different levels of fixed effects using a dummy variable regression (Wooldridge, 2009, p. 485). This is a specification which includes a dummy variable for each fixed effect: for example a dummy for each practice or for each practice-disease combination. This method allows for a comparison of the R^2 values from each regression. This comparison is not possible in the fixed effects regressions mentioned previously as the dependent variable is time-demeaned at different levels.

Finally, we perform a Hausman test to determine if fixed effects are preferred to random effects (Hausman, 1978).

4.5 Results

Our results are split into descriptive statistics, the incentives analysis and the sensitivity analysis.

4.5.1 Descriptive statistics

Table 4.1 [p.165] presents the descriptive statistics for all of the variables. Average performance is high in the first year at 83.8% and increases for the first five years. In 2004/5 the average revenue per patient was 0.405 which represented 40.5% of £75 rising to 0.836 or 83.6% of £133.76 by 2012/13. Similarly, points per patient started off lower and rose over time from 0.6 to 1.1 per patient. Payoff increases from 0.361 to 0.737 for revenue and from 0.531 to 1.182 for points between 2005/6 and 2006/7. These large changes are due to the lower threshold being increased from 25% to 40% for almost all indicators. This means that practices no longer received points for 15% of the patients they previously treated; hence the average marginal payoff increases for those patients that are treated between the lower and upper thresholds. It can be seen that this change in payoff per patient is driven by changes in the lower threshold as the revenue per indicator does not increase by the same degree.

The ADPF has a mean value of one by construction. The CPI has a mean slightly greater than one because this measure used the mean list size from 2003/4. The CPI increases as the actual mean list size increases compared to the fixed value used for the CPI calculations.

On average, the value of an indicator was between seven and eight points. The lower threshold is fairly stable over the nine years with significant changes happening only in 2006/7 and then 2012/13. Most indicators had an upper threshold of around 80%. Disease prevalence was around 3% when averaged across all included diseases. Average practice list size increased each year from a low of 6,226 and patients per GP decreased each year from a high of 1,802. The remaining variables are stable over time.

Table 4.2 [p.166] and Table 4.3 [p.166] provide more detailed descriptive statistics for the ADPF and the two payoff variables. There is significant variation in all measures which results in differing reputational and revenue incentives. The standard deviation

of the ADPF increases in 2009/10 and then again in 2010/11 when changes are made to the calculation. Overall, the standard deviation increases from 0.180 to 0.577 over the nine years, while the mean remains constant.

The extent of variation in the payoff variables also increases when changes are made to the lower threshold and to the value of a point. The standard deviation of points per patient increases from 1.714 to 3.028 while the standard deviation of revenue per patient increases to a similar degree. Overall the standard deviation in these measures is high which reflects the variation in incentives for difference practices and for different indicators.

Table 4.4 [p.166] provides further evidence that the two payoff variables are measuring separate aspects of the QOF. The correlation between these variables decreased over time and was low enough such that multicollinearity was not a concern. The variance inflation factor for these variables was 2.03 in 2004/5, 1.83 in 2008/9, 1.58 in 2009/10 and 1.38 in 2010/11. These are below the value at which researchers should be concerned (O'Brien, 2007). Figure 4.1 [p.210] illustrates the effects of the changes in ADPF calculations in nine annual scatter graphs of ADPF on list size. In the first five years, there was little variation in ADPF due to the truncation and square rooting. In the seventh year, the truncation is removed and the ADPF is allowed to take values of zero.

4.5.2 Incentive analysis

Table 4.5 [p.167] reports the fixed effects estimation results for Equation 4.3 on four subsamples of the data. In all models, both points per patient and revenue per patient were positive and statistically significant predictors of practice performance with t-ratios ranging between 6.04 and 18.14.

For the first five years, a change in the reputational incentive by one unit (one point per patient) leads to an expected change in practice performance of 0.205 percentage points. In comparison, a change in the revenue incentive of one unit (the value of a point in a given year per patient) leads to an expected change of 0.279 percentage points. However, an F-test testing if the difference between the coefficients was statistically significant failed to reject the null that the difference was zero ($p=0.3$).

Similar results concerning the incentives were found for years six to nine, again the difference was not statistically significant. The effect of the upper threshold was much

larger (0.45 compared with 0.0264) and had a smaller t-ratio (4.52 compared with 104.22), perhaps attributable to more changes being made to the threshold in this period.

Results for all years on the balanced panel of indicators find a larger difference between the responses to the two incentives. On this occasion the difference was statistically significant ($p<0.01$). The difference is smaller and not statistically significant in the final model in Table 4.5 [p.167], when analysing the larger dataset of unbalanced indicators.

Table 4.6 [p.168] shows how the effect of the reputational and revenue incentives changed over time by estimating Equation 4.4. The effect of the reputational incentive increases while the effect of the revenue incentive decreases over time. This pattern is similar in the panels of balanced and unbalanced indicators. Figure 4.2 [p.211] shows the estimated coefficients and 95% confidence intervals from the regression on the balanced panel. The statistically significant difference between the two incentives persists until 2008/9. After 2008/9 the two effects converge with overlapping confidence intervals.

4.5.3 Sensitivity analysis

In Table 4.7 [p.169], we investigate the impact of our choice of explanatory variables in models estimated using the first five years of data. The first two models include only one of the incentive variables and year dummies, each are statistically significant and larger than the model in which both are included, the third model. However, despite the effects falling when both are included, there is still a statistically significant effect for each. This suggests that although the variables are correlated, they each capture a distinct feature of the QOF. In the next models we include the upper threshold, then the practice characteristics for all years, then the practice characteristics for years one to five only. The final model restricts the sample to only indicators present in all five years: a balanced panel of indicators. These inclusions do not qualitatively change our results from Table 4.5 [p.167].

In Table 4.8 [p.169] all three possible specifications of the fixed effects were presented in one table: practice, practice-disease and practice-indicator. The choice of specification did not substantively change the estimated coefficients.

In Table 4.9 [p.170], a similar model is estimated where dummies for the appropriate fixed effects were included manually. The practice-indicator model has the highest adjusted R^2. This is further support for the specification used in Table 4.5 [p.167] and Table 4.6 [p.168].

Finally, in Table 4.10 [p.171] we compare fixed effects and random effects estimation, on the practice-indicator specification. A Hausman test strongly favours fixed effects (Hausman, 1978).

4.6 Conclusion

Since its introduction, the QOF has been the focus of a range of studies aimed at uncovering how effective a compensation scheme it represents and if there have been negative or unforeseen consequences. Previous literature assumes the QOF is a financial incentive and has focused on this feature. However, we argue that the QOF also has a significant reputational component which has been under studied. It was the aim of this chapter to investigate whether reputational incentives are present, and what effect they have on performance relative to financial incentives.

4.6.1 Summary of findings

When looking at the full nine years of QOF data and a consistent, balanced panel of indicators we find that the reputational incentive has a smaller effect on practice performance than the revenue incentive. This result suggests practices respond more to changes in revenue than changes in points.

Further analysis into these effects over time reveals a more complicated relationship between the two incentives. When the incentives are interacted with time dummies, a clear pattern is found whereby reputation becomes more important over time as revenue becomes less important. The effects of the incentives become more similar in later years. However, even in these later years, each effect is still statistically significant and measures distinct and separate incentives.

These findings suggest that in the early years of QOF, reputation is not a chief motivator for GPs. This was perhaps owing to an absence of a visible benchmark of practice performance with which to base their performance. In later years, performance norms are established and the reputational incentive becomes more relevant.

The falling magnitude of the revenue incentive may be attributable to GPs allocating a fixed amount of effort to maximise their utility from the QOF. When increasing reputation becomes more important, GPs may divert effort away from their revenue-maximising behaviour of previous years.

4.6.2 Strengths and weaknesses

Our analysis offers one of the few examples whereby the reputational and financial incentives of P4P can be estimated separately. We take advantage of a dataset covering almost the entire population of UK general practices for nine years of an ambitious and unique P4P scheme. The size and structure of the data permit a flexible allowance for unobserved characteristics; practice-indicator fixed effects allow the performance of practices to differ at a specific indicator level. We also take care to ensure the specification of our model was not affected by the multicollinearity of our main independent variables.

However, a number of potential weaknesses should be highlighted. Firstly, we have made the assumption that QOF points are a source of reputation for GPs. The fact that QOF points are published online supports this assumption, and GPs are also motivated in part by their reputation (Roland and Campbell, 2014). However, the extent to which GPs use points specifically as a source of reputational motivation is contestable. GPs response to the points incentives may be explained by reputation, it may also be due to an incentive to increase performance to attract more patients. By increasing the number of patients at a practice GPs will also increase their income. Therefore, the reaction to points may be motivated by an alternative source of financial incentives. It is plausible that reputation explains part of the response to points, but other incentives may also have a role.

Secondly, for our models to estimate the impacts of the relative incentives from reputation and revenue, it is a necessary for GPs themselves to observe these incentives and then react to them. An IT system came with the QOF which allowed practices to receive updates on how they were performing across different indicators (Kontopantelis et al., 2013a). The existence of this IT system should have meant that practices were aware of the relative incentives from points and revenue.

Finally, Gravelle et al. (2010) have shown that practices were able to influence prevalence rates which would suggest that the ADPF was not entirely exogenous to

practice behaviour. However, as the degree to which practices acted in this way was small, we believe that this lack of endogeneity would not have biased our results.

4.6.3 Future research

The QOF has been running for 10 years and has been subjected to small but frequent changes over that period. Changes planned for 2013/14 and 2014/15 were the most significant. First, a substantial number of indicators were retired and replaced with new indicators recommended by the National Institute for Health and Care Excellence (NICE) (NHS England et al., 2014). The following year the revenue associated with 338 points was reallocated into the capitation payments for practices and therefore this revenue was no longer linked to performance (BMA et al., 2014). A further 100 points was reallocated into rewarding practices if they avoided unplanned admissions. As these two changes were the most significant to date, they may have impacted on the incentives and motivation of GPs. Therefore, future work should analyse QOF data from these years.

To help address one of the main weaknesses of this chapter future research should focus on the extent to which GPs treat points as a reputational incentive. Qualitative interviews could attempt to capture the various incentives associated with the QOF and how these incentives motivate performance.

4.6.4 Policy implications

Policies such as the QOF tend to focus on the financial elements in order to motivate changes in behaviour and performance. Our research shows that the reputational incentives can be similar in magnitude to financial ones. If we assume that reputational incentives can be induced at lower costs, our results would imply that efforts to make the reputational elements stronger could be effective. Reputational incentives require an initial investment in the means to measure and report performance, but they do not require regular and frequent payments.

The benchmarking of an individual's performance against their peers seems to be an important component of the reputational incentive. In the early years, reputation did not seem to matter to GPs. We propose that this is due to the absence of a benchmark of performance. This would suggest that GPs take the performance of their peers as an indication of what their own performance should be. Future policies could do more to foster comparisons between practices.

The policy recommendations from this chapter are predicated on the assumption that the observed response to the points incentive is evidence that GPs are responsive to their reputations. The effect identified as a reputational response may be explained by alternative incentives but the result that GPs are motivated by financial as well as additional incentives remains.

5. INSTITUTIONAL PEER EFFECTS: EVIDENCE FROM ENGLISH PRIMARY CARE

5.1 Introduction

There is widespread and longstanding interest in mechanisms to improve the quality of services offered by healthcare providers. Payment reform is one mechanism and P4P is a popular example of such reforms (Appleby et al., 2012). The literature reviewed in Chapter 2 suggests that providers may respond to non-financial incentives, such as public reporting resulting in a reputational incentive. These non-financial incentives are often combined with financial incentives. This combining of incentives can be deliberate, as in cases where a specific comparison of the incentives in intended (Lindenauer et al., 2007). It can also be incidental, as in the QOF, where performance information is collected primarily for payments but is published as well. In this chapter, we explore whether GP practices demonstrate peer effects: when an individual is influenced by the behaviour or performance of those around them. We approach this issue by using the QOF as a way to measure performance and taking advantage of changes in practice peer groups induced by a partial re-organisation of regional healthcare organisations.

An important feature for our study comes from the QOF being an example of a non-contestable, non-rival mechanism. Practices do not compete in a traditional sense as there are neither winners nor losers. This means that practices are not in competition for QOF income because performance payments are made regardless of their relative performance compared to other practices.

However, because data on performance is widely shared, opportunities for comparisons exist and reputational incentives are created (Health & Social Care Information Centre, 2015b). Practices might use the published data to gain information about their peers and then respond to this information. These published data on performance might also be used by patients to inform their choice of practice. Previous studies of provider competition in healthcare frame the mechanism in terms of competition for demand from patients (Gravelle et al., 2014). Within primary care, there is evidence that patients choose practices based on the quality provided (Santos et al., 2015). However, the rates of patients switching practice are reported to be as low as 3% (Monitor, 2014). This low rate of switching could be due to the high

information costs involved in switching. Therefore it seems open to debate if practices would compete for patients using quality.

We propose a mechanism in which providers are motivated to respond to their peers' quality for reputational reasons. Unlike competition for patients, this is a form of non-financial intrinsic motivation. Peer effects can act in a positive way, when individual performance is pulled up by good peers, or a negative way, when performance is pulled down by poor peers. A priori it is not possible to determine whether PCT peer effects will be positive or negative. Indeed peer effects may not exist at all. However, if peer effects do exist, these incentives act alongside the financial incentives of the QOF. The difference being that peer effects specifically relate to the providers' response to the performance of their peers.

There are strong financial incentives to increase performance on the QOF, but there are no financial incentives to respond to peers performance. This would not be the case if the QOF was designed around a tournament or included relative performance bonuses. Under a tournament, providers at the top of the performance distribution might receive higher performance payments. These would be dependent on their relative performance compared to other providers, and not merely their absolute performance. The QOF does not take into account the relative performance of providers.

There is a substantial literature on peer effects for individuals but not for institutions, though there is a literature on relative performance evaluation and benchmarking that mainly focuses on hospitals (Fichera et al., 2014). Primary care providers are small institutions and, unlike hospitals, may behave more like individuals. We therefore focus on the peer effects of groups of practices brought together because of the institutional structure of primary care.

In the period we consider, practices held contracts with geographically defined organisations called PCTs. These organisations were responsible for monitoring and regulating quality but, crucially for our focus, they create an environment in which local GPs meet and compare performance (BMA, 2015). We use an exogenous change in the membership of PCTs to estimate the magnitude of institutional peer effects. This re-organisation was administrative, bringing PCTs more in line with the boundaries used for local government, and unrelated to variations in performance

(House of Commons Health Committee, 2005). It was also partial, with some parts of the country not exposed to changes in membership.

The identification of peer effects is complex as the relationship between individual and peer behaviour could be explained by three mechanisms: endogenous effects, exogenous effects or correlated effects (Manski, 1993). Endogenous effects are where the detected peer effect can be attributed to the peer groups' behaviour directly. Exogenous effects are where the detected peer effect is attributed to a common characteristic of the peer group that pre-dates the group's formation. Correlated effects are where the detected peer effect is attributed to common influences which affect the whole group. It is the endogenous effects which researchers wish to identify. Identification is also difficult due to self-selection into peer groups (Carrell et al., 2009; Rice and Sutton, 1998). A further problem, known as reflection, was introduced in Manski, (1993). This explains how identification of peer effects is difficult due to the simultaneous relationship between individuals and peers. That is, individuals affect peers and peers affect individuals.

Within the literature on peer effects there are several methods used to address the identification problems mentioned above. Self-selection is often addressed using situations where the peer group is randomly assigned. This sub-set of the literature has found mixed results. Within squadrons of the US Air Force Academy, strong and robust peer effects were found acting between individuals' academic performance and the performance of their peers with whom they were randomly grouped (Carrell et al., 2009). Peer effects were not observed for professional golfers when players were randomly assigned to their opponents (Guryan et al., 2007). The absence of peer effects was attributed to the professional nature of the individuals involved and it was suggested that high skilled professionals would not be influenced by peer effects to the same degree as low skilled workers or students. However, peer effects have been found to explain the prescribing habits of physicians in the US and Taiwan (Nair et al., 2006; Yang et al., 2014). So the extent to which professionals are influenced by peers is contested.

Even with random assignment into peer groups, the reflection problem of Manski (1993) still confounds the results. A possible solution is to combine random assignment with a measure of peer ability that pre-dates the formation of the peer group. This method was implemented when analysing peer effects of Dartmouth

College room-mates by using peers' previous academic ability (Sacerdote, 2001). Peer effects were found to impact on academic measures such as, the Grade Point Average, and social behaviour, such as joining a fraternity.

The use of current peers' past performance has been implemented elsewhere when not coupled with random assignment (Lavy et al., 2012). Consider school children tested at age 11 years and 14 years, where test scores were observed for the peer group at age 14 as well as the scores for the same students at age 11. This method breaks the simultaneous relationship as the performance of the individual at age 14 cannot affect the past test scores of their peers. However, if current peer group ability is approximated by the test scores of that group at age 11, then any effect detected must be a peer effect. Lavy et al. (2012) found that using this inter-temporal measure of peer effects coupled with multiple within-student observations and fixed effects addressed the methodological limitations of much of the previous research into peer effects. They found that school children were negatively affected by peers whose test scores were in the bottom 5% of the national distribution but were not positively affected by the proportion of peers in the top 5%. Girls were affected by top scoring students, but boys were not.

In our example, the unit of observation is individual practices whose behaviour will be dominated by the behaviour of the GPs who own and work in the practice. This approach contributes to the literature as it focuses on professionals working within small organisations. The majority of the peer effects literature has focused on school children and students, who are arguably more likely to be influenced by their peers than highly trained professionals. It is not clear whether peer effects will be as strong in professionals as they are in school children and students. The existence of peer effects in practices has implications for other small and locally grouped organisations.

In our setting we predict that practice performance will be correlated with the performance of practices from their PCT, as this is their peer group. The re-organisation of PCTs will impact on the nature of peer effects such that they should strengthen for new peers but become weaker for old peers. We find mixed evidence with respect to these two hypotheses depending on the identification strategy used.

Section 5.2 covers the data used. Section 5.3 provides more details on PCTs and how peer performance is measured. Section 5.4 explains our econometric methods. Section 5.5 presents our results and Section 5.6 provides a conclusion to this chapter.

5.2 Data

The QOF is the main source of data for this study and has been discussed in Chapter 3. In this section, the details of the QOF pertinent to this chapter are revisited.

5.2.1 Practice performance

To measure quality we use QOF scores for each practice which is publicly available online (Health & Social Care Information Centre, 2015a). We use data on the total number of points that practices have achieved. This is a simple and concise measure of practice performance that is visible to other practices online (Health & Social Care Information Centre, 2015b). These data are downloaded for a five year period from 2004/5 to 2008/9 which spans the re-organisation of PCTs.

In order for a consistent measure to be created, a practice must be present in all five years of the data. Therefore it is necessary to drop those practices that did not have five observations, either due to openings/closures or to missing PCT information. Before dropping observations, there were 44,532 practice-year observations from between 8,869 and 8,940 practices per year. The balanced panel is of 8,041 practices totalling 40,205 observations. Table 5.1 [p.172] reports some descriptive statistics for the sample of practices dropped due to attrition. Dropped practices are poorer performing, small and have an older GP workforce. They tend to be located in less rural areas with higher population deprivation.

5.2.2 Practice characteristics

In additional to data on practice performance, we use a range of practice and area characteristics to control for confounding factors. These were obtained from the HSCIC and include: the average age of GPs; the proportion of UK qualified GPs; the proportion of GP partners; the proportion of male GPs; number of FTE GPs; PMS practice; dispensing practice; Low Income Scheme Index; and age and gender of practice population. Descriptive statistics are shown in Table 5.2 [p.172].

We also measure the quality of the PCT according to the World Class Commissioning score system (Department of Health, 2007; Sobanja, 2009). This was a measure, taken in 2009, of the performance of each PCT on the competencies required in order to

meet the current and future health needs of their populations. The mean practice score was 54% with a standard deviation of 10%. The scoring system is limited in its capacity to measure the complexities of commissioning and the initiative is no longer active, having been cancelled in 2010 (McCafferty et al., 2012).

5.3 Peers and peer performance

In our context, the peer group of interest is the PCT in which practices are located and performance is measured using the QOF. Both of these elements are explained in detail below.

5.3.1 Primary Care Trusts

PCTs are responsible for the provision of primary care and commissioning of secondary care for a defined geographical area (Walshe et al., 2004). Initially, Primary Care Groups were formed and if they showed an ability to manage their own budgets they could gain 'trust' status and become a PCT (Bojke et al., 2001). By 2002, all PCTs were fully established (NHS Confederation, 2011).

The dual role of providing and commissioning meant that there was a conflict in terms of the appropriate size of PCTs. Small ones would have greater knowledge about the needs of the local population and have a stronger relationship with the GPs and practices in the area. Large ones would have greater bargaining power when it came to agreeing terms with large secondary care providers, particularly with respect to rare conditions affecting only a small population.

Beyond their role as commissioner and provider, PCTs would also ensure a level of bureaucratic and administrative oversight with respect to the practices in their area. PCT managers would relay performance information down to practices, therefore providing the means to observe and react to peer performance. For example similar organisations in Australia have been shown to influence the infrastructure of general practice but not the management of chronic conditions (Scott and Coote, 2010).

Within PCTs, smaller groups of practices formed Practice Based Commissioning (PBC) groups (Lewis et al., 2007). These groups were formed to engage GPs in commissioning services and improving patient health. The formation of these groups may act to strengthen the peer effects acting within a PCT as the PBC groups were expected to meet with their PCT in order to redesign services (Checkland et al., 2009).

Data do not exist on the membership of PBC groups so it is not possible to identify peer effects acting within these groups.

PCTs underwent a significant re-organisation in October 2006 which saw the pre-existing 303 PCTs merged into 152 large PCTs (House of Commons Health Committee, 2005). 31 PCTs in London, as well as 47 elsewhere, were not merged and remained the same size as before the re-organisation. Figure 5.1 [p.212] depicts the boundaries of the PCTs before and after the re-organisation. The PCTs that merged went from an average size of 26 to 76 practices. The range of practices in a PCT changed from 8-92 practices to 20-143 practices. It is not clear what effect the re-organisation had on PCT staffing levels as data on this are not available. It is possible that merged PCTs experienced a reduction in management staff and the staff that remained could have experienced greater pressures and responsibilities. This could have had a negative effect on the degree of bureaucratic and administrative oversight possible by the PCT.

The motivation for this reorganisation was to save money, improve commissioning and bring PCT boundaries in line with local authority boundaries (NHS Confederation, 2011). The merger decisions were not made by practices or PCTs. There was no evidence to suggest that variability in the performance of practices influenced the reorganisation in terms of merging PCTs containing poor performing practices with PCTs containing good performing practices. Therefore, it is reasonable to assume that the reorganisation was exogenous to our measure of practice performance. The stated intention was that larger PCTs would move away from their responsibility of providing services, instead being able to focus on the funding and contracting of primary/secondary care (Smith and Mays, 2005). More details of the reorganisation are given in NHS (2006).

5.3.2 Peer performance

The maximum available points in the first two years (2004/5 and 2005/6) was 1050, which then fell to 1000 in subsequent years. Due to this change in maximum points, performance was measured by the percentage of maximum points each practice was awarded in each year.

We observe the performance of each practice and the performance of all of their peers. The way in which the data is structured allows for the identification of practices forming the peer group before and after the reorganisation in all time periods. This

structure is shown in Figure 5.2 [p.212]. For a given practice, we observe the PCT before the reorganisation (A) as well as the PCT after the reorganisation (C). We also observe the PCT with which it merged (B). This structure means it is possible to create two measures of peer performance, one for the *old peers* (those in A) and one for the *new peers* (those in B).

5.3.2.1 Old peer performance

The average old peer performance is calculated as:

$$P_{ikt}^{O} = \left[\frac{\left(\sum_{j \neq i} p_{jkt} \Big/ (n_{kt} - 1) \right)}{Max\ points_t} \right] * 100 \qquad 5.1$$

That is, the sum of QOF points p over all practices in PCT k (except for practice i) in time t. This is divided by the number of practices in the PCT n_{kt} minus one to give the average total points of the old peer group. To transform this into a percentage measure we divided by maximum points available in each year (either 1050 or 1000) and multiplied by 100. This resulted in a measure of performance for the peers in the old group for all five years.

5.3.2.2 New peer performance

The average new peer performance was calculated as:

$$P^{N}{}_{ikt} = \left[\frac{\left(\left(\sum_{j \neq i} p_{jkt}^{C} \right) - \left(\sum_{j \neq i} p_{jkt}^{A} \right) \Big/ (n_{kt}^{C} - n_{kt}^{A}) \right)}{Max\ points_t} \right] * 100 \qquad 5.2$$

That is, the sum of QOF points for PCT C minus points for PCT A; this results in the sum of points for the new peers from PCT B. To take the average of these practices we divided by the number of practices in PCT B, which was the difference between the number of practices in C and A. As with the performance of old peers we then created a percentage measure for new peers.

5.4 Econometric methods

The econometric methods used to identify peer effects are split into two different approaches. The first takes advantage of the exogenous change in peer group while the second uses lagged performance to mitigate the reflection problem.

5.4.1 Exogenous change in peer group

As old peer performance is observable for practices from merged and non-merged PCTs, we first model the effect of the reorganisation on peer effects for these old peers. The following model is estimated separately for merged and non-merged practices. A comparison of the effect of the reorganisation can then be performed between the merged and non-merged practices.

$$Y_{ikt} = \beta_0 + \beta_1' \boldsymbol{X_{ikt}} + \beta_2 T + \beta_3 P_{kt}^{O} + \beta_4 P_{kt}^{O} * D + \alpha_i + \varepsilon_{ikt} \qquad 5.3$$

Where Y_{ikt} is the performance of practice i in PCT k at time t. $\boldsymbol{X}$ is a vector of practice characteristics used to control for correlated and exogenous effects. For example, an association between the performance of a practice and that of its peers could result from practices in the same PCT treating similar types of patients, or from practices being predominantly in rural locations. These are examples of exogenous effects as they pre-date the existence of the peer group, that is, practices were rural before they were in a certain PCT. Correlated effects are common elements of peer group that exist because of the peer group itself. For PCTs, a correlated effect could be the unobserved quality of the PCT management.

T is a set of year dummies. P_{kt}^{O} captures the performance of the peer group before the reorganisation (old peers) and is also interacted with D, which is a time dummy taking the value one for the years after the reorganisation (2006/7, 2007/8 and 2008/9) and zero otherwise. α_i is an unobserved practice fixed effect term while ε_{ikt} is an idiosyncratic error term $i.i.d.N(0, \sigma^2)$. This model was estimated using fixed effects to remove the confounding effect that α_i may have on individual performance (Wooldridge, 2010, p. 481). The elements contained in α_i may include practice characteristics for which data were not available such as practice population ethnicity. They might also include unobservable factors such as how outward-looking the practice is. The elements of α_i are assumed to be fixed over time. Estimation using

fixed effects at the practice level also accounts for practice performance being persistent over time.

From Equation 5.3 we make the following propositions regarding the hypothesised effects of the coefficient on peer performance.

Proposition 1: β_4 measures the effect of the PCT reorganisation on the effect of peer performance. We predict it will be negative for practices from merged PCTs, as it captures the effect of the dilution of the peer group, and zero for practices from non-merged PCTs. The negative coefficient would imply that the addition of new peers reduces the impact from the existing peer group. There should be no change in the non-merge groups.

Testing whether the difference between the β_4 terms is statistically significant reveals if there is a different effect of the reorganisation for merged and non-merged practices. To test this, the samples are combined and Equation 5.3 is estimated with each explanatory variable interacted with a merged/non-merged dummy. The practices from non-merged PCTs can be thought of as a placebo test. In the absence of a reorganisation we should not observe any significant change over the treated period for this group.

Estimating the effect of the reorganisation on both measures of peer performance (old and new) is only possible for practices in merge PCTs. Those in non-merged PCTs do not have new peers. The following model is estimated for practices in merged PCTs.

$$Y_{ikt} = \beta_0 + \beta_1' \boldsymbol{X_{ikt}} + \beta_2 T + \beta_3 P_{kt}^O + \beta_4 P_{kt}^N + \beta_5 P_{kt}^O * D + \beta_6 P_{kt}^N * D + \alpha_i + \varepsilon_{ikt} \quad 5.4$$

Where P_{kt}^N captures the performance of post-reorganisation peers (new peers). The other variables are as defined above. This model differs from Equation 5.3 as it estimates the change in peer effects for both the old and new peers, but can do so only for practices whose PCT merged.

From Equation 5.4 we make the following propositions regarding the hypothesised effects of the coefficients on peer performance.

Proposition 2: β_3 will be larger than β_4 and both will be positive. β_3 captures the effect of the old peers before the reorganisation and during this time these peers were the entire PCT. β_4 captures the effect of the new peers before the reorganisation and during this time these peers were in a neighbouring PCT. The assumption is that the correlations between practice performance will be strongest for practices within the same PCT.

Proposition 3: β_5 will be negative. This term captures the effect of a pre-existing peer group ceasing to be the entire membership of the peer group. The additional practices that are added to the PCT cause the correlation with the pre-existing peers to be reduced.

Proposition 4: β_6 will be positive. This term capture the effect of practices in neighbouring PCTs becoming part of the peer group. When these additional practices are added to the PCT they become part of a formal peer group and the correlation increases.

This method solves two of the main problems when estimating peer effects: the endogenous choice of peer groups and area effects. The method addresses the first problem due to the exogenous change in peer groups imposed by the PCT reorganisation. The second problem is addressed by the use of practice and PCT characteristics and practice fixed effects. It also controls for reflection under the assumption that the degree of peer reflection is fixed over time and not itself affected by the reorganisation. This model also assumes that it is current period peer performance which is most relevant. In an alternative specification we relax these assumptions.

5.4.2 Peer reflection

An alternative method uses a different approach to address the reflection problem which has been used in previous research (Lavy et al., 2012; Sacerdote, 2001). Past performance of current peers is used in place of current performance as the latter is reflective. The use of past performance also allows for a lag in the time it takes for information about peer performance to reach individual practices.

This method is straightforward in the education literature as the past performance of current peers is unlikely to pick up school specific trends in performance because the peers groups are made up of students from different schools. However, neighbouring

PCTs were merged in the PCT reorganisation, and we may be unable to fully control for practices with similar area characteristics. Therefore, using past performance of current peers could result in our "peer effect" simply picking up area specific trends in a measure of performance that is persistent over time. Put another way, neighbouring PCTs have similar performance due to area effects, and past performance is related to current performance due to persistence. In order to address the reflection problem we must control for the persistence of performance and area effects.

We therefore estimate the following specification:

$$Y_{ikt} = \beta_0 + \beta_{1n} {Y_{ikt-1}}^{n} + \beta_2 P^{O}_{ikt-1} + \beta_3 P^{N}_{ikt-1} + \beta_4 \boldsymbol{X_{ikt}} + \varepsilon_{ikt} \quad for\ n = 1, \dots, 4 \qquad 5.5$$

Where Y_{ikt} is the performance on an individual practice regressed on the performance of that practice in the previous period Y_{ikt-1}. As past performance is a strong predictor of current performance, a flexible specification is used to capture this relationship. This involves modelling past performance up to the fourth power. P^{O}_{ikt-1} and P^{N}_{ikt-1} measure the performance of old peers and new peers in the previous period. $\boldsymbol{X}$ is a vector of practice and area characteristics. ε_{ikt} is an error term $i.i.d.N(0, \sigma^2)$.

This specification controls for the reflection problem as well as unobserved area effects and persistence. The lag of individual performance controls for individual persistence. The lags of performance for old and new peers will capture peer effects but will also be contaminated by area effects not controlled for with the elements of $\boldsymbol{X}$. A change in peer effects would present as a change in the effects of lagged peer performance when the peer groups change. This model is estimated using OLS for each year separately resulting in four different sets of regression results.

This model specification introduces a lagged dependent variable in the form of practice performance. The inclusion allows the model to control for the persistence in practice performance from year to year. However, this specification also introduced endogeneity as the error term is correlated with one of the independent variables, namely the lagged dependent variable (Keele and Kelly, 2005). This endogeneity is likely to bias the estimates on the peer performance variables. However, if the estimated coefficients differ when comparing models using data from before and after

the PCT reorganisation, this difference can be argued to be in response to changes in peer effects.

5.4.3 Sensitivity analysis

In addition to the model specifications above, we perform a number of sensitivity checks to explore if peer effects are more prevalent in sub-samples of practices.

The association between the performance levels of practices in a PCT could be the result of managerial effort and not endogenous peer effects. We classify the sample on the basis of PCT quality measured by the World Class Commissioning score system. The detection of a change in peer effects could be explained by PCT managers focusing their attention on underperforming practices. Assuming that increased attention improves the performance of practices, then our models would detect this as an increase in the association between performance scores of practices in the same PCT. If a significant peer effect was detected only in the PCTs with high scores, this may suggest that the change in behaviour was due to managerial effort and not the performance of peers. This would be an example of an exogenous effect and not a peer effect using the terminology of Manski (1993). Low scoring PCTs are those in the bottom 50% of the score distribution, while high scoring PCTs are in the top 50%. Equations 5.4 and 5.5 are re-estimated on these sub-samples.

The weakening of the correlation between peer and own performance when PCTs become larger might also simply reflect size. We include size of PCT in the regression, but as a second sensitivity check we stratify based on the size of the old PCT compared to the size new PCT. For example, a practice in a small PCT with 20 peers before the reorganisation and a large PCT with 100 peers after the reorganisation would have a figure of 20%. This value distinguishes practices in terms of how diluted their peer initial group has become. Practices that have a significantly diluted peer group may react differently in terms of their behavioural response to the performance of new peers when compared to practices with a less significantly diluted peer group. This relative size variable is calculated for all practices. The sample is then split into the top and bottom 50% and Equations 5.4 and 5.5 are also re-estimated on these sub-samples.

For analysis using sub-samples by PCT management quality and PCT size, the difference in peer effects in each sample is tested. This test involves combining the two samples and interacting all variables with a dummy variable labelling each half of

the sample. The statistical significance of the interacted term tests if the difference in the coefficients in each sample is zero. A statistically significant difference would be counter to our earlier hypothesis that changes in the association between own and peer performance were due to peer effects.

In addition to our measures of peer performance based on the mean performance of each peer group, we also tested different measures based on low/high performance in a similar approach to Lavy et al. (2012). The justification for this further analysis comes from an appreciation that there are several ways to characterise peer performance. Practices may be more responsive to measures of peer performance which better capture the extremities of the distribution of peer performance. In total, three alternative measures are used. Firstly, the percentage of practices within the peer group whose performance falls below the 5th percentile of national performance for that year. Secondly, the percentage of practices within the peer group who attained 100% of available QOF points. Finally, the percentage of QOF points achieved by the 2nd lowest scoring practice in the peer group. The 2nd lowest was chosen to avoid extreme outliers in low performance. For practices whose score was equal to the 2nd lowest, the 3rd lowest was used.

5.5 Results

The presentation of the results is divided into subsections, first discussing the descriptive statistics of the sample, followed by the models exploiting the exogenous change in peer group and the peer reflection models. The final subsection presents the sensitivity analysis.

5.5.1 Descriptive statistics

Table 5.2 [p.172] contains the descriptive statistics for the sample of practices used in this study. The number of PCTs is reduced in 2006/7 due to the reorganisation which corresponds with an increase in the number of practices per PCT. The percentage of achieved QOF points remains above 90% in all five years with a high of 97% in 2007/8. List size and list size per GP remain consistent over time as do most of the practice characteristics. The proportion of GP partners decreases over time reflecting the increased use of salaried GPs over this period. Patient age and gender composition is also stable.

5.5.2 Exogenous change in peer group results

Table 5.3 [p.173] presents the results of Equation 6.3 which has been estimated separately on the merged and non-merged samples. The association between practice and PCT performance is positive and statistically significant for both groups of practices: a 1% point increase in mean PCT performance is associated with a 0.7% point increase in practice performance. The interaction of old peer performance with the post-reorganisation period measures how this association changes after the PCT reorganisation. There is a negative and statistically significant change in the association for merged practices (-0.1 percentage points), suggesting a decrease in the strength of the peer effects when the peer group is diluted. The interaction term is not statistically significant for practices in PCTs that did not merge, as expected.

Table 5.4 [p.175] shows the results of combining the samples and interacting with a dummy variable for practices from merged PCTs. These results show that the difference in the peer effects for the merged practices (-0.1 percentage points) was statistically significant at 10% with a t-ratio of -1.87.

Table 5.5 [p.174] contains the results of fixed effects regressions of practice performance on the performance of the old and new peers (Equation 5.4). The model includes the same set of explanatory variables as in Table 5.3 [p.173]. Peer effects are substantially larger in the old peer group when compared to the new peer group in the pre-reorganisation period (0.68 vs. 0.13). These peer effects are likely to be endogenous and the result of area or location effects that contribute to practice performance and which we have not controlled for. After the reorganisation, the decrease in peer effects for the old group is -0.1 with a t-ratio of -2.75. The hypothesised increase in peer effects for the new group is modest at 0.04 and is not statistically significant with a t-ratio of 0.99.

5.5.3 Peer reflection results

Table 5.6 [p.174] presents the results of four regressions of the specification shown in Equation 5.5. The lagged own performance terms are jointly significant in all models, highlighting the persistence in practice performance. The 2005/6 regression shows a negative association for lagged old performance but no statistically significant association for new peers. The 2006/7 regression, when the new peers had entered the PCT, showed no peer effect. The 2007/8 regression showed positive and statistically significant peer effects for new peers. In the 2008/9 regression, the coefficient for new peers is not statistically significant.

These results show that the lag of new peers' performance becomes significantly associated with own peer performance one year after these peers enter the PCT. This association lasts for only one year. Old peer performance does not demonstrate a reduced association during this period.

5.5.4 Sensitivity analysis results

Table 5.7 [p.175] contains the model of Equation 5.4 but stratified by PCT management quality. The coefficients from each subsample are of the same sign and of similar size. For example, the coefficient on old peers' performance is 0.53 in PCTs with poorer management compared to 0.74 in PCTs with better management. The coefficients of interest are those on the peer performance terms interacted with the post-reorganisation period. These show the effect of the reorganisation for practices from PCTs with low versus high management quality. These coefficients differ in each subsample, -0.067 compared to -0.09 for old peers for example. Testing the differences in the interacted coefficients from Table 5.7 [p.175] reveals the difference is not statistically significant with t-ratios of 0.34 and 0.81 respectively.

Table 5.8 [p.175] contain the models of Equation 5.4 but stratified by relative PCT size. The difference between the interacted terms is not statistically significant (t-ratios -0.04 and 0.89).

Regression results from Table 5.9 [p.176] are those of Equation 5.5 stratified by PCT management quality. For brevity only the coefficients on the lags of old and new peer performance are shown. The coefficients from the bottom 50% tend to differ to those from the top 50%. However, the only difference that is statistically significant comes from the lag of old peer performance in the 2008/9 regression. The coefficient from the bottom 50% (-0.02) is significantly smaller than the coefficient from the top 50% (0.1). The t-ratio for this difference is -2.11. However, the individual coefficients from these sub-samples are not statistically significant.

Table 5.10 [p.176] presents regressions using Equation 5.5 stratified by relative PCT size. As with the previous table, only the lagged peer performance variables are shown. For these regressions statistically significant differences are found for old peer performance in 2007/8 and for new peer performance in 2008/9. In 2007/8 the effect of old peer performance is significantly smaller for practices in the bottom 50% in terms of their initial peer group size. The coefficients are -0.02 compared with 0.11 with a t-ratio on the difference on -3.86. This suggests that lagged old peer

performance has a greater association with current performance for practices whose starting PCT was larger. In 2008/9 the effect of new peer performance is not statistically significant for either subsample. However, the coefficient for the bottom 50% is significantly smaller than the top 50%, -0.08 compared to 0.037 with a t-ratio of -2.67.

Table 5.11 [p.177] shows results where alternative measures of peer performance are tested. Practice performance is negatively associated with the percentage of peers who were poor performers. A unit increase in the percentage of poor performing peers decreases practice performance by 0.15 percentage points. Post-reorganisation, the size of this effect was reduced by 0.048 percentage points. These results suggest that performance is pulled down by the presence of poor performing peers and that this effect reduces when the peer groups are diluted. This sensitivity to poor peers is not mirrored when using a measure of well performing peers. The percentage of high performing practices in the peer group does not have the opposite effect. Practices also do not appear to be sensitive to peer performance when using the 2nd lowest score as the measure of peer performance.

5.6 Conclusion

The presence of peer effects has been often studied in the education sector and to a lesser extent in other areas of the workforce. However, the applicability of this body of research to professionals, such as GPs, is questionable. In this paper we test for the existence of peer effects within groups of GP practices. We propose that this relationship results from reputational incentives that P4P and public reporting create.

5.6.1 Summary of findings

A number of the propositions stated in the methods section have been confirmed by our analyses. Proposition 1 and 3 stated that peer effects would be reduced following a dilution of the peer group and was confirmed in Table 5.3-Table 5.5 [p.173-174]. Proposition 2 was confirmed as the peer effect was larger for the old peer group than the new peer group (Table 5.5 [p.174]). Proposition 4 stated that peer effects would increase when practices entered the new peer group but was not confirmed as the effect was not statistically significant (Table 5.5 [p.174]). The confirmation of these propositions leads to our main finding: that the peer effects acting within a PCT were diminished when that PCT was merged to create a new, larger peer group. More generally, when an existing peer group is merged, resulting in a larger peer group, the

peer effect within the initial group is reduced. The merging of PCTs created an exogenous change in peer group for some practices. Those practices that were not involved in PCT mergers did not experience the same decrease in peer effects.

The models using lagged performance would suggest that changes in peer effects may be sensitive to time, with increased peer effects for new peers one year after the reorganisation but not after two years.

The further analysis conducted using stratified samples found that peer effects did not substantially differ by PCT quality or size. Analysis using alterative performance measures provides evidence that practices were more sensitive to low performing practices than to high performing ones.

5.6.2 Strengths and weaknesses

We present one of the few examples of being able to observe changes in peer group, and hence use past performance of peers to solve the reflection problem. To the best of our knowledge, we present the only study that analyses how performance within groups is related from an institutional peer effects perspective. Our methods are similar to those of Lavy et al. (2012) and Sacerdote (2001) in that we have measures of past performance and observe a change in peer groups. We differentiate by having an institutional not an individual setting which involves professionals not students. We also benefit from an exogenous change in peer group, the result of the PCT reorganisation, which was not the case in Lavy et al. (2012).

Our paper makes a number of assumptions and assertions which should be highlighted. We assume that PCTs are a relevant peer group. This seems plausible given practices share a common source of management and oversight, and they provide a forum where performance is discussed and compared. However, it is also plausible that smaller and more relevant peer groups exist within PCTs and that peer effects may be stronger for these groups. An example of such peer groups would be PBC groups. The peer effects we detect and attribute to PCTs may in fact be, in part, PBC group peer effects, or some form of smaller peer group. In which case the PCT peer effect we present is potentially an aggregation of peer effects from other groups. This is analogous to attributing classroom peer effect which may be due to peer effects within smaller classroom social groups. It is not possible to identify these smaller peer groups within PCTs using currently available data. However, the PCT peer effects

reacted in a predictable way to the PCT reorganisation, suggesting that national policies have an impact on peer effects.

A further weakness concerns whether the response to the PCT reorganisation is motivated by peer effects or via PCT managerial quality or effort. This weakness provides the rationale for using the World Class Commissioning score as a means to stratify the sample according to PCT quality. However, this is a crude and simplistic measure and is unlikely to capture the full impact of variations in PCT quality. It is therefore possible that unmeasured PCT quality is still affecting our ability to estimate peer effects.

We also assume that practices are motivated by reputation and that this behaviour would cause practices to compare themselves based on QOF performance specifically. It is not clear in the literature if practices use QOF performance as a metric with which to benchmark themselves against their peers. Other measures may be more relevant but are not routinely collected.

Finally, although we have found no evidence to suggest otherwise, we have assumed that the PCT reorganisation was not motivated by practice performance in any way.

5.6.3 Future research

A natural extension to this work would be to analyse the more recent change in peer group composition when PCTs were dissolved and replaced by Clinical Commissioning Groups (CCGs) in 2013. These organisations perform a similar function to PCTs but they are sized somewhere in-between the old and new PCTs.

Future research should also consider if it is possible to identify more selective peer groups within PCTs. This is likely to involve qualitative research on practice social networks. Other measures of PCT quality could also be explored or developed depending on the availability of data. For example, PCT management may be approximated by PCT management budgets or staff expenditure.

5.6.4 Policy implications

In the past several decades commissioning bodies have undergone several key reorganisations: Primary Care Groups to PCTs; small PCTs to large PCTs; and PCTs to CCGs. These reorganisations have changed the composition and size of the commissioning body without knowing if there is an optimal size and if peer effects exist and/or are affected by these changes. Our findings imply that enlarging the peer

groups will diminish the peer effects within the initial group. This reduction in peer effects is an unintended but arguably positive consequence of the reorganisation. Positive in the sense that we find that practices seem to be influenced more by the presence of low performing peers than by high performing peers. This is evidenced by their response to the percentage of low performing practices in their peer group. The reduction of peer effects may have led to an increase in PCT performance.

An alternative discussion might argue that even these negative peer effects could be used to increase PCT performance. If a policy objective was to improve average performance within a PCT, it would be more effective to reduce the number of low performers than to increase the number of high performers. This is due to the spillover effects which reducing low performance would have.

We have found a correlation between practice performance and peer performance. The mechanism which explains this observation could stem from spatial competition or, as we assert, from peer effects. The policy implications of these competing mechanisms differ greatly. To promote spatial competition, *patients* need to be provided with more and/or better information on practices and the costs involved in switching practices should be reduced. On the other hand, to promote peer effects, *GPs* need to be provided with more and/or better information about their peers and further opportunities for comparison should be created. However, given the evidence that peer effects may result in group performance being pulled down, providing practices with more information may not be advisable.

6. DOES P4P AFFECT THE JOB SATISFACTION OF GENERAL PRACTITIONERS?

6.1 Introduction

GPs have always played a major role in the NHS since they are providers of primary care and gatekeepers to hospital treatment. They are often the first and only point of contact with the healthcare sector and as such play an important role in patient satisfaction and quality of care. Given their essential function in the NHS, understanding what factors impact on their job satisfaction and working lives is of importance for policy and services.

In this chapter we aim to uncover what impact P4P has on GP job satisfaction as well as a range of working life measures. This is currently a topical issue as the QOF has recently undergone the most significant redesign since its inception. A third of performance income has been shifted into capitation, substantially reducing GPs exposure to P4P (BMA et al., 2014). This change in direction is cause to highlight the lack of research on the appropriate ratio between income linked to performance and other sources. Despite Scott et al. (2011) drawing attention to the absence of research in this area, several recent articles have cautioned against having a large ratio of P4P to other income and have suggested the QOF incentives are too large (Gillam and Steel, 2013; Raleigh and Klazinga, 2013; Roland and Campbell, 2014).

P4P exposure is measured as the income from P4P as a percentage of total income. We use a measure of overall job satisfaction, nine sub-domains of job satisfaction, intentions to quit, hours worked per week and overall life satisfaction. We estimate this relationship using a range of methods including continuous DID which provides the causal effect of increased P4P exposure (Angrist and Pischke, 2008; Card, 1992).

This work adds to the existing literature in a number of ways. We link existing datasets in a unique way enabling us to measure job satisfaction and P4P exposure for the same sample of GPs. We take a wider range of measures of job satisfaction and working lives than previous studies. The continuous DID method estimates a causal relationship. GPs' exposure is not self-determined by the GP and GPs were all affected at the same time, therefore self-selection is not a confounding factor. We also

measure the immediate effects and the longer term effects of individual GP exposure to P4P and observe satisfaction from before and after the introduction of P4P.

As a preface to our finding, we conclude that P4P exposure did not affect overall job satisfaction, nor did it affect life satisfaction, intentions to quit or the sub-domains of job satisfaction. Changes to the QOF in 2014/15 will see a third of the income currently linked to performance removed, and channelled into capitation payments (BMA et al., 2014). GPs have reportedly welcomed this change (Roland and Campbell, 2014). Our results suggest that these exogenous changes to P4P exposure will not impact on the GP job satisfaction or working lives measures we have used.

This chapter is organised as follows. Section 6.2 reviews the literature on the importance of job satisfaction and how P4P impacts on job satisfaction. Section 6.3 discusses the two datasets we use and how these are linked. Section 6.4 explains how GP income and P4P exposure are measured. Section 6.5 presents our econometric approach. Section 6.6 contains the results. Section 6.7 discusses our findings and their implications for policy and research.

6.2 Literature

The extent to which P4P, and specifically the QOF, has impacted on professions is discussed in Chapters 2 and 3 respectively when reviewing the evidence from systematic reviews. In this chapter we introduce the literature uniquely related to the link between P4P and job satisfaction. The literature in Sections 6.2.1 and 6.2.2 was identified through key word searches for studies focusing on the effect of job satisfaction and the impact of P4P on job satisfaction. The literature in Section 6.2.3 was identified from existing literature reviews (Gillam et al., 2012) and from key word searches for empirical studies on the effects of P4P on job satisfaction.

6.2.1 The importance of job satisfaction

The job satisfaction of GPs is important for two main reasons: its effect on GPs leaving the workforce and its effects on the quality of care they provide. These effects can be measured in a number of ways, as can job satisfaction, and the effects are all the more important in public healthcare systems where local wages tend to be inflexible and cannot be used to compensate for dissatisfaction or stress (Scott et al., 2006).

Continuity of care is an important corner stone for primary care (Guthrie et al., 2008; Haggerty et al., 2003). GPs develop knowledge of the needs of their patients and wider community as well as the local services providers across all sectors of healthcare. This knowledge is accumulated over time and is necessary for GPs to fulfil their role as gate keeper and care provider. GPs leaving medicine interrupts continuity of care as new GPs have to develop this knowledge. Therefore, understanding and predicting when GPs quit is important for efficient healthcare planning and delivery (Teljeur et al., 2010).

Several studies have measured how job satisfaction is related to GPs making changes to their working arrangements or leaving medical practice all together. Satisfaction with their jobs was associated with decreases in GP intentions to switch from providing public to private healthcare in Finland (Kankaanranta et al., 2007). Structural models of the job satisfaction of GPs in England and Scotland show that personal and practice characteristics have an effect on sub-domains of job satisfaction which in turn affect overall job satisfaction (Scott et al., 2006). The same study models intentions to quit, finding that they are affected by job satisfaction measures and personal and practice characteristics. Lower levels of job satisfaction for US GPs were related to several measure of intentions to withdraw: quitting; working fewer hours; changing specialty; and leaving direct patient care (Williams et al., 2001). Similarly, dissatisfaction with pay was associated with plans to leave medical practice within two years (Pathman et al., 2002).

In analysis which followed GPs to determine if a stated intention to quit was actioned, dissatisfied GPs were more likely to report intentions to quit and those with intentions to quit where more likely to actually quit (Hann et al., 2011). However, GPs who reported high job satisfaction still quit in some cases, perhaps if they wanted to start a family. The authors suggest that job dissatisfaction is not the antithesis of job satisfaction; instead, they measure two distinct elements of working lives. Therefore, reducing the number of GPs who quit could be achieved by reducing dissatisfaction and not necessarily by increasing satisfaction.

A review from 2004 of the causes and consequences of doctor turnover found that many factors were linked to doctors leaving medical practice, including personal and financial factors (Misra-Hebert et al., 2004). A lack of job satisfaction was highlighted in several of the reviewed studies as a reason doctors left practice. Consequences on

doctor turnover included financial losses to the practice, lower morale for doctors not leaving and increase dissatisfaction in patients.

A further study of doctor turnover and attrition found that, for US doctors, age had the largest effect on actually leaving while dissatisfaction only predicted an intention to leave (Rittenhouse et al., 2004). With specific reference to those working in dementia care, low levels of job satisfaction was the largest reason for quitting (Vernooij-Dasssen et al., 2009). Doctor job satisfaction seems to be important with respect to stated intentions to quit, but the extent to which satisfaction affects actually quitting varies by setting.

The costs associated with replacing GPs when they leave medical practice was estimated over the period from 1987-1991 for GPs working in the US (Buchbinder et al., 1999). The replacement cost for a GP in 1991 was $236,000 ($412,000 in 2014 dollars) and covered transport costs, relocation costs and recruitment agency fees. The issues involved with replacing GPs extend beyond the financial cost but must consider the lengthy education and training period (10+ years), such that future workforce needs must be forecasted (Candace et al., 2009). This overview of the relationships between quitting and job satisfaction suggests that dissatisfaction can lead to quitting, which places a cost on the healthcare provider.

The importance of GP job satisfaction extends beyond turnover and retention. A survey of patients and GPs in the US found a positive correlation between job satisfaction reported by the GP and satisfaction with care reported by the patient (Haas et al., 2000). This relationship was also observed for US hospital doctors in addition to finding that more satisfied patients and doctors were associated with patients reporting better continuity of care and fewer patients missing appointments (Linn et al., 1985). Increased patient adherence is also correlated with higher job satisfaction in the US (DiMatteo et al., 1993). Again in the US, similar findings relate job dissatisfaction to GP perceptions of the quality of the care they provide (DeVoe et al., 2002). Low job satisfaction for GPs in England and Wales is associated with poor prescribing habits such as prescribing drugs with adverse effects or those which are no longer accepted by the medical community (Melville, 1980). The body of research discussed above highlights that GP job satisfaction can have an impact on the care they provide and how patients react to that care.

6.2.2 The theoretical effects of P4P on job satisfaction

There are several theories from economic and psychological literature which might predict the likely effects which P4P would have on job satisfaction.

6.2.2.1 Crowding-out

The concept of motivation crowding-out has been reviewed widely in psychology (Deci et al., 1999) and economics (Prendergast, 1999). Crowding-out occurs when an extrinsic incentive, such as an additional payment or punishment, overrides an individual's intrinsic motivation to complete a task, such as recognition or work ethic. This idea is particularly relevant to healthcare providers as they are usually considered to be motivated in-part by altruism or a desire to do good (Le Grand, 2003). Public perception or stigma may also interact with extrinsic motivation to further the erosion of intrinsic motivation (Siciliani, 2009). The extent to which GPs express motivation crowding-out in reality is not known but, if intrinsic motivation is eroded by P4P, the effect could be lower job satisfaction.

6.2.2.2 Loss of autonomy

Autonomy is a form of intrinsic motivation and professionals, above other types of workers, value their autonomy (Young et al., 2012). P4P schemes tend to reduce autonomy by relating pay to specific goals or targets, therefore removing a degree of discretion. Given that professionals value autonomy, we would expect the negative relationship between P4P and autonomy to be felt more strongly by professionals, such as GPs. A lack of control over their environment is associated with lower job satisfaction for GPs (Freeborn, 2001).

However, the interaction between professionals, P4P and autonomy may be more nuanced in practice. It has been argued that individuals can react differently to external motivation. An external incentive, such as a performance target, may be internalised by individuals causing them to be intrinsically motivated, perhaps by a personal desire to meet the target. Thus the individual perceives no loss in autonomy as they want to meet the target for personal reasons (Ryan and Deci, 2000). The implication being that P4P would not affect all GPs equally; instead, the effect would depend on the extent to which the GPs feels their autonomy is being limited (Young et al., 2012). If GPs perceive their professional autonomy is undermined by a P4P scheme, the likely effect is a reduction in job satisfaction.

6.2.2.3 Wage dispersion

Wage dispersion is the term often used to refer to the difference between high and lower earners in an organisation or industry. Greater wage dispersion can be the result of linking pay to performance as high performing individuals will benefit from higher pay. Negative consequences can arise when the indicator of performance is not accepted or supported by all individuals involved. Under these circumstances, increased wage dispersion resulted in lower satisfaction and reductions in collaboration for academic staff (Pfeffer and Langton, 1993). Closely related to this concept is relative income and income inequality. The Easterlin Paradox asserts that as countries improve their economic conditions, they do not improve their subjective wellbeing or happiness (Easterlin, 1974). However, there is also more recent evidence that contradicts this finding (Stevenson and Wolfers, 2008).

Within major league sports teams, the effect of wage dispersion varies from significantly positive, to insignificant, to significantly negative depending on the sport in question (Frick et al., 2003). This suggests that the relationship is variable and may differ by sector. The relationship also differs by unionisation, with non-union members experiencing job dissatisfaction when wage dispersion increased (Georgellis et al., 2008). An alternative explanation with empirical evidence asserts that jealousy can result from wage dispersion but the higher wages of co-workers also signals the individual's future earnings (Clark et al., 2009). The latter effect has a positive effect on job satisfaction which outweighs the jealousy effect. The majority of evidence would suggest that greater wage dispersion resulting from P4P would lower job satisfaction, particularly for those at the bottom of the income distribution.

6.2.2.4 Sorting

Compared to a fixed wage or salary, P4P schemes reward those who exert more effort and therefore perform better, but punish those who exert less effort. This results in employees with higher skill and ability preferring employment where they are paid via P4P (Eriksson and Villeval, 2008; Lazear, 2000). Not only are high ability workers attracted to P4P schemes but such schemes induce workers to perform better (Booth and Frank, 1999). Performance is higher even for workers who do not have a preference for P4P (Cadsby et al., 2007).

The introduction of a new P4P scheme may be disruptive as old workers who dislike the risk involved leave and are replaced with new workers. However, in the long term, this type of self-selection can result in a more productive workforce (Jensen, 2001).

High ability teachers have been observed leaving teaching when pay compression resulted in little differences in pay between high and low ability teachers (Hoxby and Leigh, 2004). A P4P scheme that could effectively reward high ability teachers based on better classroom performance could have reduced teacher turnover.

Having no P4P tends to create a working environment that is desirable to those who are more risk averse and less productive (Jensen, 2001). As GPs undergo time consuming and expensive training to become a GP, it is unlikely that they will immediately leave the profession due to an aversion to P4P. Instead, the effect may be on job satisfaction which, over time, could result in them leaving general practice. Medical students choosing a speciality may also be deterred from general practice if they dislike P4P. On the other hand, GPs with a preference for P4P could see their job satisfaction increase.

6.2.2.5 Risk aversion

In economics, individuals are described as risk averse when they display preferences for certain outcomes over uncertain ones. Even when the expected value of the uncertain outcome is higher than the value of the certain outcome, individuals will often prefer the lower value because it is less risky.

Most individuals demonstrate risk aversion (Holt and Laury, 2002). Those who prefer risk are known as risk lovers, while risk neutral individuals have no preference either way. This concept of risk aversion can also be related to preferences for compensation schemes. Risk aversion would show up as preferences for fixed salary as an alternative to P4P (Cadsby et al., 2007).

P4P is risky as performance is determined not only by an individual's effort but also other factors such as patient adherence. Employees who are risk averse were also found to dislike variable compensation schemes and preferred to work for organisations with fixed compensation schemes (Cable and Judge, 1994). The general finding that risk averse individuals prefer fixed compensation does not mean that firms should avoid P4P schemes where income is at risk. P4P is associated with productivity increases and attracting more productive workers, as discussed in Section 6.2.2.4. P4P is more appropriate as a compensation method the greater the amount of vested interest the individual has in the firm. For example, as CEOs own more firm shares, P4P becomes a more effective method of motivation (Jensen and Murphy,

1999). Increased work stress is another potential outcome from P4P. A literature review in 2011 finds a predominantly positive effect between P4P and increased work stress and that P4P induces work stress via uncertainly about future income (Ganster et al., 2011).

The concept of risk aversion and its relationship to P4P is related to that of sorting discussed above. However, they can have separate effects if GP job satisfaction is influenced more by their risk aversion than by their preferences for P4P because they are hard working.

6.2.2.6 Summary of theories

We have introduced various theoretical reasons why job satisfaction would be affected by P4P. These affects are summarised below with each demonstrating the individual-specific nature of the effects of P4P making firm hypotheses problematic.

- *Crowding-out*: increased exposure to P4P could reduce job satisfaction as intrinsic motivation is eroded
- *Loss of autonomy*: as exposure to P4P increases, the loss of autonomy should be felt more strongly. Hence job satisfaction should fall
- *Wage dispersion*: P4P creates more income inequalities which should result in lower job satisfaction for those who do not benefit as much as others. However, the higher incomes may signal future earning potential which could increase job satisfaction
- *Sorting*: high ability GPs will feel rewarded by P4P and have higher job satisfaction, whereas low ability workers will feel punished and have lower job satisfaction. Ability is loosely defined here, and could simply mean the GPs ability to perform as per the design of the P4P scheme
- *Risk aversion*: assuming GPs are risk averse on average, those who are more exposed to P4P should have reduced job satisfaction

6.2.3 The empirical effects of P4P on job satisfaction

A systematic review of the impact of the QOF highlighted six studies which researched the impact of the QOF on professionals and team working (Gillam et al., 2012). Four of these studies targeted themes such as professional wellbeing or job satisfaction and will be discussed in more detail below. Two studies focused on either the effects on performance in small practices (Doran et al., 2010) or the views of GPs

and practice staff on exception reporting (Campbell et al., 2011) and will not be discussed further.

An ethnographic study, observing 12 GPs, nine nurses, four healthcare assistants and four administrators, concluded that the internal motivation of GPs was not affected by the financial incentives of the QOF, but nurses showed some concern over changes to their role (McDonald et al., 2007). The study benefits from repeatedly observing several healthcare workers over a five month period but, as only two practices were sampled, the results may not be representative.

Semi-structured interviews of 21 GPs and 20 practice nurses in 2007 highlighted some of the issues with P4P that we address with this study (Campbell et al., 2008). Interviewees reported raised morale and improved work-life balance for GPs resulting from increased income for work which was already being undertaken but was not being specifically rewarded. However, GPs also expressed concerns about their extra income creating negative public opinion and concerns that a culture of monitoring and surveillance was being developed. Nurses felt that they contributed significantly to their practice's QOF performance but were not rewarded for this additional responsibility.

The study shows the conflicting effects that P4P may have on GPs and their job satisfaction, and how those not financially rewarded can be adversely affected. The study did not interview salaried GPs who, like nurses, do not receive a share of QOF incomes and therefore the study cannot comment on the effects on morale for these GPs. Further research involving semi-structured interviews of practice nurses reiterated their additional responsibilities following the QOF, which was largely welcomed, and the lack of reward attached to these responsibilities, which was not welcomed (McGregor et al., 2008).

Only one study made a specific mention of job satisfaction relating to the QOF (Maisey et al., 2008). As with previous research, semi-structured interviews were carried out with a nurse and a GP from 12 different practices. Job satisfaction was reported to be lower for GPs resulting from reduced continuity of care and from additional managerial roles. Nurses responded more positively on the development of their role and increased autonomy, but did report higher levels of workload stress and a lack of appropriate remuneration for this workload. With respect to GPs, the issues

which were linked to increased job satisfaction were the higher incomes and reduced hours. Those linked to decreased job satisfaction were additional managerial roles, the administrative nature of the QOF (box-ticking) and the feeling that the GPs' professional identity was being eroded.

The effect of financial incentives on GP job satisfaction has been studied in other countries. When studying GPs in the US and their response to managed care (an incentive to reduce healthcare costs), GPs were found to be indifferent to these types of incentives once GP and practice characteristics were controlled for (Grembowski et al., 2003). Only a GP being paid by salary was associated with dissatisfaction. It should be noted however, that salaried in this setting is compared to payment by fee-for-service. In the UK, the comparison would be between a salaried GP and a practice partner who takes a share of practice profits. GPs were also dissatisfied if they worked in larger practices associated with increased bureaucracy. The study uses cross-sectional data in a setting where GPs are likely to self-select into practices based on their own preferences. This means the findings could be biased. For example, the fact that productivity bonuses are not associated with job satisfaction could be influenced by GPs self-selecting to work in practices which match with their attitudes towards bonuses.

Financial incentives to meet certain quality targets, such as quality of clinical care or access to care, were applied to GPs in Catalonia in 2003 (Gené-Badia et al., 2007). The affected GPs and nurses were surveyed before and after the new policy. Questions related to intrinsic motivation, such as job satisfaction and team support, showed no change over the two years analysed for either nurses or GPs. However, GPs did report increases in workload, pressure, lack of time and stress. The study only focused on the year immediately after the new financial incentives were introduced and did not take a longer term view. Nor does the study employ rigorous statistical methods, instead only compared scores on questions before and after the policy change not controlling for confounding factors.

A negative relationship was observed between intrinsic and extrinsic motivation for GPs in France (Sicsic et al., 2012). This recent empirical finding suggests that extrinsic motivation may erode intrinsic motivation in GPs. Therefore caution should be applied when implementing P4P schemes as they may have a negative effect on overall motivation, via reduced job satisfaction for example.

The qualitative research on the effect of the QOF on professionals and the international research on the effects of financial incentives on job satisfaction and intrinsic motivation have a common theme. They analyse the effects of a financial incentive on some measure related to job satisfaction but cannot determine individual GPs' exposure to the financial incentive. Under such schemes, GPs will be differentially exposed to the incentives depending on the scheme's design. If a certain patient group is targeted by a financial incentive, then the power of that incentive will be related to the size of the patient group in a given practice.

Hadley et al., (1999) avoid this issue when surveying GPs in 1997 by asking them to self-report the extent of personal financial incentives in their practice. This approach was necessary in the US where GP contracts and local financial incentives vary more than they do in the UK. GPs reported if their incentives were to increase or reduce services. When GPs faced multiple incentives they were asked to give the net effect or to report neutral incentives. GPs who reported incentives to reduce services were 3.5 times more likely to be very dissatisfied when compared to those GPs who reported neutral incentives. Incentives to either increase or reduce services were associated with GPs reporting less professional autonomy. As with other studies discussed earlier, GPs could self-select into preferred contracts which may bias the findings and the data is from one cross-section of GPs.

The expected impact of the QOF on job satisfaction is uncertain. The policy significantly increased GPs' personal income which increases satisfaction, but the QOF also came with increased extrinsic motivators, lower autonomy, wage dispersion, rewards for effort and greater income risk. The overall effect is uncertain and likely to vary across groups of GPs.

When measured on a seven-point scale, job satisfaction increased when the QOF was introduced from an average of 4.58 to 5.17 (a value of four means neither satisfied or dissatisfied while a value of five means somewhat satisfied) (Whalley et al., 2008). However, the QOF came as part of a large change to GPs' contracts which also included the removal of the need to provide out of hours care, an unpopular job responsibility amongst GPs. Figures suggest that 91% and 87% of GPs surveyed in 2005 and 2008 worked at a practice which had opted out (Hann et al., 2009; Whalley et al., 2006b). As several factors which affect GP job satisfaction changed at the same time as the QOF was introduced, the role of the QOF is unclear.

6.3 Data

The data for this study come from two sources: the QOF and the GP WLS. The former provides a source of data that can be used to estimate performance-related pay at practice level. The latter provides a survey, conducted by The University of Manchester, of the GP population focusing on job satisfaction and working conditions. A bespoke linked dataset is created by linking QOF data with the GP WLS, thus enabling us to determine the relationship between P4P and job satisfaction.

6.3.1 GP WLS

The GP WLS is a postal survey and has been collecting data on the working lives of GPs in the UK since 1998 (Sibbald et al., 2000). Surveys have been conducted in 1998, 2001, 2004, 2005, 2008, 2010, 2012 and 2015. This time period captured several key changes in policy that affected the working lives of GPs; notably the new GMS contract in 2004. We utilise surveys from 2004, 2005 and 2008 for this study. Due to response bias the GP WLS is not fully representative of the total GP population, although research has shown that this response bias does not affect the determinants of job satisfaction (Gravelle et al., 2008).

The 2004 GP WLS was conducted in February 2004 and consisted of a random cross-section target sample of 1950 GPs and an additional longitudinal target sample of 2258 GPs who had responded to the previous GP WLS in 2001 (Whalley et al., 2005). Response rates for the 2004 samples were 53% for the cross-section and 54% for the longitudinal sample. Key findings were an increase in job satisfaction from 2001 levels but some trepidation about the effect of the upcoming change to their contract which would include the QOF.

The 2005 GP WLS was conducted in September 2005 and again consisted of a 2000 GP cross-sectional and 2122 GP longitudinal target sample (Whalley et al., 2006b). The response rate in 2005 was lower for the cross-sectional sample (45%) but higher for the longitudinal sample (64%) than in 2004. The key findings showed a reduction in hours worked by 5.5 hours per week and an increase in income of £14,000 per year. Job satisfaction increased from 2004 levels while job pressure and intentions to quit fell.

The 2008 GP WLS was conducted between September and November 2008 and had a target sample of 3,000 GPs and 1,986 GPs for the cross-sectional and longitudinal

samples respectively (Hann et al., 2009). Survey response was 44% for the cross-sectional sample and 70% for the longitudinal sample. The survey finds job stress increases, particularly concerning adverse publicity from the media. Job satisfaction fell when compared to 2004 levels but still remained above those seen in 2001.

The GP WLS provides a measure of overall job satisfaction for GPs as well as a number of GP characteristics. Descriptive statistics for the GP WLS are given in Table 6.1 [p.178]. The variables in the first section of Table 6.1 [p.178] are the dependent variables used in the various models estimated. The second section shows the control variables used.

Job satisfaction, the sub-domains of job satisfaction and life satisfaction were measured on a 7-point scale with 1='extremely dissatisfied' and 7='extremely satisfied'. Several of the job satisfaction sub-domains focus on elements of a GP's working life which are likely to have been affected by a large-scale P4P scheme: choice of working methods, remuneration and variety in job. Although these sub-domains have been singled out, P4P has the potential to impact on all sub-domains.

Intentions to quit are measured on a binary scale with 1 equalling considerable or high likelihood of leaving direct patient care and 0 equalling moderate, slight or no likelihood of leaving. The time scale for this question is within the next five years.

GPs are also asked to report the number of hours they typically work per week. This measure excludes out-of-hours care.

The GP WLS uses a banded measure of GP income. Surveys from 2004 and 2005 use the same bands while the survey from 2008 changed the bands to reflect the increases in GP income due to the QOF.

6.3.2 QOF

Details of the QOF are given in Chapter 3. For this chapter, QOF data are downloaded for the first QOF year (2004/5) and the fourth QOF year (2007/8) as these years correspond with the 2005 and 2008 surveys (Health & Social Care Information Centre, 2015a). Data from these years are used to first measure the maximum QOF income for each practice. This was then used to calculate the P4P exposure for each practice and each GP in the GP WLS.

6.3.3 GP WLS – QOF linkage

In order to determine the P4P exposure for GPs sampled in the GP WLS, the QOF and the GP WLS have to be linked together. The link between the two datasets is made using the following steps:

- A unique GP identifier (General Medical Council code) was available in the GP WLS
- GMC codes were linked to a practice identifier using the GMS data
- The practice identifier allowed QOF data to be linked

The first year of the QOF started in April 2004 and ended in March 2005. Therefore, the 2004 GP WLS was conducted just prior to the QOF, while the 2005 GP WLS was conducted after the first year payments had been made (Whalley et al., 2006a, 2005). We linked the 2005 survey data to the first year of the QOF payments and used the 2004 survey as our pre-QOF observation. The survey from 2008 provides an observation after the fourth year of the QOF, allowing for analysis of the effect of P4P exposure in the long run and also the effect of changes to exposure between the first and fourth years of the QOF (Hann et al., 2009).

6.3.4 Practice characteristics

Supplementary data from the Health and Social Care Information Centre are used to control for certain practice characteristics (Health & Social Care Information Centre, 2015c). These data are: the Low Income Scheme Index, GPs per practice, dispensing practice, PMS practice, population ethnicity and rural/urban classification of the practice. These variables are shown in Table 6.1 [p.178] along with those from the GP WLS.

6.4 Income and P4P exposure

First we discuss how QOF income can be calculated from the available data on practice performance, disease prevalence and list size. We then present two methods by which QOF income can be used to generate a measure of P4P exposure.

6.4.1 Measuring QOF income

For our analysis we calculated the maximum income a practice could receive if they achieved all available QOF points. Variations in this income measure do not depend on practice performance but on how the design of the QOF affects practices differently, which GPs could not influence. This distinction is important as we need to

avoid a measure of QOF income that was influenced by the GP directly. Achieved income would have been determined by the effort of the GP which itself may have been determined, in part, by their job satisfaction. We might expect a less satisfied GP to put in less effort and therefore earn themselves, and their practice, less QOF related income. By adapting QOF income to measure the potential income that could be realised, and not the income that resulted from achievement, we remove a possible source of endogeneity.

The formula for calculating practice income is explained in detail in Section 3.2.2.5, with the individual components of this formula being explained in earlier sections. The formula for income summed over all indicators (Equation 3.8) is used to calculate income, which is then used to allocate measures of P4P exposure to GPs.

6.4.2 Measurement of P4P exposure

In order to estimate the effect that P4P had on GP job satisfaction we need to calculate the percentage of income that came from the QOF. P4P exposure is measured using two methods: at the practice level and at the individual GP level.

6.4.2.1 Practice level P4P exposure

For each practice, QOF income per patient is calculated and this is divided by total income per patient. Algebraically, practice level P4P exposure is:

$$P4P_{jt} = \left(\frac{gQOF_{jt}/L_{jt}}{\gamma}\right) * 100 \qquad 6.1$$

Where $P4P$ denotes the practice P4P exposure. $gQOF_{jt}$ is gross QOF income which is divided by the practice list size. γ is the average income per patient from all income sources for all practices. Income from all sources is calculated using information from the GP Earning and Expenses Enquiry 2004/5 (Health & Social Care Information Centre, 2006). There were 27,334 GPs in 2004/5 with average gross income of £241,795. Total average gross income for this population was therefore £6,609,224,530 (£6.61billion). The English patient population for 2004/5 was 52,833,584 giving average gross income per patient of £125.10. This measure assumes that all practices earn the same per patient regardless of the demographics of their patient population.

6.4.2.2 GP level P4P exposure

P4P exposure at the GP level measures the percentage of individual GP partner income that came from the QOF. Salaried GPs are not included as they do not directly receive P4P income. In order to measure this exposure we first predicted GP income if the QOF had not been introduced. We estimate the determinants of income in the 2004 survey (before the QOF was introduced) and use these estimated coefficients to predict the income that would have been received post-QOF had the QOF not been introduced. These predictions are made for 2005 and 2008. The assumption in this approach is that, if the QOF had not been introduced, the determinants of income would have remained constant over the time period 2004-2008. We allow the values of the determinants of income to change over time, but not their effect on income.

This method provides the denominator in the measure of P4P exposure and the numerator is obtained by estimating the amount of individual GP income that came from the QOF. The maximum practice level QOF income is calculated (see Section 3.2.2.5) and used to create this numerator. The incomes calculated from QOF data are gross income and not net of expenses. The incomes predicted by our determinants models are based on self-reported GP incomes which are net of expenses. In order to have a comparable denominator and numerator, we adjusted the QOF income downwards to account for expenses. In 2004/5 GP partners in England had average gross earnings of £241,795 and average net earnings of £103,654, giving a gross/net ratio of 2.33 (Health & Social Care Information Centre, 2006). We divided gross QOF income by 2.33 to obtain a net figure. This rescales the income and the relative variation across GPs and practices is maintained.

We also account for the fact that not all GP partners in a practice will receive an equal share of QOF income. We assume that the share received is determined by the full-time equivalent (FTE) of the GP and how many FTEs where at the practice. For example, a 0.5 FTE GP partner in a practice with seven FTE partners would receive $1/14^{th}$ of the QOF income.

Algebraically, GP level P4P exposure is:

$$P4P_{ijt} = \left[\frac{\left(nQOF_{jt} * \left({FTE_{ijt}}/{\sum FTE_{jt}} \right) \right)}{\hat{Y}_{ijt}} \right] * 100 \qquad 6.2$$

Where $P4P$ is the P4P exposure for GP i in practice j at time t. $nQOF$ is net QOF income. ${FTE_{ijt}}/{\sum FTE_{jt}}$ is the FTE of the individual GP divided by the sum of FTE GPs in the whole practice. $\hat{Y}_{ijt}$ is the predicted "no-QOF" income for 2005 and 2008. The whole expression is multiplied by 100 to arrive at the percentage P4P exposure for each GP in our sample.

6.4.2.3 Discussion of exposure measures

The two measures were created as they quantify the extent to which practices and GP partners were exposed to P4P via the QOF. The first measure is based on practice income while the latter is based on individual GP income. Both measures should be thought of as measuring the percentage of income now at risk due to P4P. As the percentage of income at risk increases, the GP becomes more exposed to P4P. This increased exposure is what we expect to affect our measures of job satisfaction and working lives.

The exposure measures differ as practice exposure assumes that QOF income is divided equally amongst each GP. The measure of GP exposure assumes that QOF income is divided based on the FTE of the GP and practice. Each of these assumptions is likely to oversimplify a complicated set of income sharing rules. However, these sharing rules are unknown, and our approach makes use of the information that is available. The exposure measures also compare QOF income to different sources of income. For practice exposure, QOF income is compared to the average practice income per patient. For GP exposure, QOF income is compared to what we have estimated to be the GP's income had the QOF not been introduced.

6.5 Econometric methods

In this section we describe the econometric methods and models used to estimate predicted income for the measure of GP P4P exposure and the models used to estimate the effect P4P exposure had on job satisfaction and working lives.

6.5.1 Predicted income

To predict income for GPs in the GP WLS we use self-reported GP income as the dependent variable and regress this on a range of GP and practice characteristics. As GPs reported their income in bands, these models were estimated using interval regressions using the income bands as thresholds (Wooldridge, 2009, p. 601). The first and last income bands are open ended, for example less than £25,000 or more than £150,000. In these cases we assumed the low bounds were zero and the upper bounds were infinity.

$$y^*_{ijt} = \beta_0 + \beta'_1 \boldsymbol{X_{ijt}} + u_{ijt}, \qquad 6.3$$

Where y^*_{ijt} is an unobserved latent variable denoting to income of GP i in practice j at time t. $\boldsymbol{X_{ijt}}$ is a vector of GP and practice characteristics. u_{ijt} is the error term $i.i.d.N(0,\sigma^2)$. y_{ijt} denotes the observed banded income from the GP WLS:

$$\begin{aligned} y_{ijt} &= 1 \; if \; y^*_{ijt} \leq a_1 \\ y_{ijt} &= 2 \; if a_1 < \; y^*_{ijt} \leq a_2 \\ &\vdots \\ y_{ijt} &= J \; if a_{\mathrm{J}-1} \leq y^*_{ijt} \end{aligned} \qquad 6.4$$

Where $a_1 \ldots a_{J-1}$ represent the income band thresholds (Sutton and Godfrey, 1995).

The variables within $\boldsymbol{X_{ijt}}$ were: age, age^2, patients per GP, partnership size, dispensing practice, non-white GP, PMS practice, population ethnicity, population deprivation and rural practice. The estimation method and choice of independent variables are consistent with previous literature using the GP WLS (Morris et al., 2011).

6.5.2 Job satisfaction and working lives

Several model specifications are used to estimate the effect of QOF exposure on various measures of job satisfaction and working lives. Each model includes one of

the measures of P4P exposure, banded GP income and GP/practice characteristics. The characteristics are selected based on previous research on the determinants of GP job satisfaction (Scott et al., 2006). Holding other factors constant, increases in GP income would increase job satisfaction. Increased income is also likely to be associated with greater P4P exposure. By controlling for GP income in each model, we ensure that P4P exposure does not capture an income effect but only captures the effect of the source of income.

We do not include GP characteristics which might have been directly affected by the introduction of the QOF such as working hours. These characteristics are classed as "bad controls" and limit our ability to detect the effect of P4P exposure on job satisfaction (Angrist and Pischke, 2008, pp. 47–48). Exposure may lead to working more hours, which in turn may reduce job satisfaction. Including hours in the regressions would absorb the effect of exposure, which is what we wish to quantify.

We estimate three specifications: cross-sectional, changes in job satisfaction and continuous DID.

Equation 6.5 models the cross-sectional association between P4P exposure and measures of workings lives.

$$Y_{ijt} = \beta_0 + \beta_1 QOF_{ijt} + \beta_2 D_{ijt} + \beta_3' \boldsymbol{X_{ijt}} + u_{ijt} \qquad 6.5$$

Y is the dependent variable, a measure of job satisfaction or working lives, for GP i in practice j at time t. $P4P$ is a measure of exposure and D is a set of dummy variables for income bands. $\boldsymbol{X}$ is a matrix of practice and GP characteristics. u_{ijt} is the error term $i.i.d.N(0, \sigma^2)$. This model was estimated on the post-QOF data (2005 & 2008) using OLS regression.

Equation 6.6 models the association between P4P exposure and changes in job satisfaction.

$$\Delta Y_{ijt} = \beta_0 + \beta_1 P4P_{ijt} + \beta_2 \Delta I_{ijt} + \beta_3' \boldsymbol{X_{ijt}} + u_{ijt} \qquad 6.6$$

The dependent variable is the change in job satisfaction. $P4P$ is a measure of exposure from 2005 or 2008 and ΔI is the change in GP income. Other variables are defined as in Equation 6.5.

An interval regression is used for each year to predict a continuous and unconstrained value for income. The change in income is calculated from this prediction. An unconstrained prediction is used as we require out of sample predictions for income in 2005 and 2008. This additional step is needed as the values of the income bands changed between 2004/5 and 2008 meaning a simple difference between bands from each year would be misleading. We estimate this model first using the changes in the dependent variable and income between 2004 and 2005 and then the changes between 2004 and 2008.

Equation 6.7 models the effect of P4P exposure using a year dummy to denote when exposure occurs: $1[t = T]$, where T takes the value 2005 or 2008. The year dummy is also interacted with P4P exposure.

$$Y_{ijt} = \beta_0 + \beta_1 1[t = T] + \beta_2 P4P_{ij} + \beta_3 P4P_{ij} * 1[t = T] + \beta_4 D_{ijt} + \beta_5' \boldsymbol{X_{ijt}} + u_{ijt} \quad 6.7$$

This model is estimated on a sample combining observations from 2004 and 2005 and then a separate regression combining observations from 2004 and 2008. It is not a requirement that GPs appear in both years. Variables are defined as in Equation 6.5.

This estimation method is known as continuous DID and is used to estimate treatment effects when all subjects are treated but the treatment varies in intensity across subjects (Card, 1992; Gaynor et al., 2010). This is appropriate for the QOF as the P4P element was introduced in all practices at the same time so all GPs became treated. However, as P4P exposure measure varies across GPs and practices, we are able to determine the level of exposure received and use this to estimate the effect of P4P exposure on our dependent variables. This method has been used to estimate the effect of minimum wages (Card, 1992) and hospital competition (Gaynor et al., 2010). The treatment effect of minimum wages varied depending on the difference between the minimum wage and average wage by region. The treatment effect of hospital competition varied depending on proximity of hospitals.

P4P exposure is equal to the post-QOF exposure in both of the years used. For example, when 2004 and 2005 observations are used, the values for P4P exposure in 2004 measure the exposure GPs' would face in 2005. The coefficient on P4P exposure for 2004 is analogous to the treatment dummy from a standard DID model which measures the effect on the treated before they become treated. This variable also absorbs unobservable individual heterogeneity that is not explained by the model and is captured by the exposure variable. This unobserved heterogeneity would not confound the estimated effect of post-QOF exposure if we assume it to be time-invariant. Therefore, it is the interaction term that measures the additional effect of this exposure in the post-QOF period.

Regressions containing both 2004 and 2008 data have different income bands in each year. We include two sets of income band dummies and set the value of these dummies equal to zero for the year in which those dummies did not apply. For example, the 2004 dummies are set to zero for any observations in 2008. This allows income to have a different effect in 2004 and 2008. To allow the same differential income effects in 2004 and 2005, when the income bands were the same, we interact the income bands with a 2005 dummy. We also include a year dummy which is equal to one in either 2005 or 2008 depending on the model.

6.5.3 Sub-sample analysis

Heterogeneous treatments effects are when the effect of treatment varies within the study sample. In our case it is possible that GPs with different characteristics will react differently to being exposed to high levels of P4P. To explore this possibility we stratify the sample into various groups who may have been affected by P4P exposure differently. These stratifications are by gender, age, list size, contract type and partnership size.

Gender and ages have significant effects on job satisfaction and also are strong determinants of income, particularly in 2005 and 2008. List size and contract type (PMS or GMS contracts) also have a large impact on income and partnership size is a factor used to allocate P4P income across GPs. For these reasons it is possible that P4P exposure may have a different relationship on job satisfaction for these groups.

It is possible that GPs would react to the new P4P element of their contracts by moving practices. This behaviour might be intended to increase or decrease exposure depending on the preferences of the GP. Though unlikely to be a common occurrence,

particularly in 2005, we also stratify on characteristics from before the QOF in 2004. We only perform this additional analysis on the characteristics which GPs could change, not age and gender.

The impact of P4P exposure on job satisfaction for these sub-samples is tested using the continued DID models.

6.5.4 Salaried GPs

Only GP partners receive income from the QOF and hence only these GPs are exposed to P4P. Salaried GPs, who are employed by the GP partners, are not exposed directly to P4P. Salaried GPs are not included in the sample as these models test for the effect of varying exposure conditional on all GPs being exposed. Models using practice level P4P exposure are re-estimated on the sample including only salaried GPs. This approach is a placebo test as salaried GPs do not have income directly at risk due to performance on the QOF.

6.6 Results

Due to the number of models estimated and the different dependent and explanatory variables used, the results are presented in subsections.

6.6.1 Descriptive statistics

Descriptive statistics on the GP WLS and practice level characteristics are shown in Table 6.1 [p.178]. Job satisfaction increases substantially between 2004 and 2005, but falls between 2005 and 2008. A similar pattern is observed for all of the sub-domains of job satisfaction with the exception of working conditions and fellow workers. The changes in likelihood of quitting also reflect this improvement, followed by reduction, in working lives. Hours worked decreases in 2005 only to increase again in 2008.

Comparing the proportion of GPs in each income band reveals large increases in income between 2004 and 2005. Far fewer GPs are in lower income bands, while far more are in the higher bands. The comparison between 2005 and 2008 is complicated by the change in bands. However, it is clear that incomes also increased between these years.

GP workload, as measured by the number of patients per GP, decreased in 2008. This is likely to be the result of increases in the use of salaried GPs as the number of practice partners decreases. The other variables are largely static over the period.

6.6.2 QOF income and P4P exposure

Table 6.2 [p.179] contains the gross, unadjusted, maximum QOF income for each practice that was matched to GP WLS respondents. There is considerable variation in practice income from P4P despite the figures representing the maximum income. The variation observed originates from the adjustments to income made via the ADPF and CPI. The large increase in QOF income between 2005 and 2008 is due a change in the average payment per point from £75 to £125 (National Audit Office, 2008). This income variation between practices and over time materialises as variation in P4P exposure. The practices linked to GP WLS respondents are larger than average (Hann et al., 2009; Whalley et al., 2006b) and this resulted in higher maximum QOF income. The average practice should earn a maximum of £75*1050 points or £78,750 in 2005 and £125*1000 points or £125,000 in 2008.

The determinants of income models for all years are shown in Table 6.3 [p.179]. The results from the 2004 regression describe the determinants for GP income pre-QOF while the post-QOF determinants are described by the 2005 and 2008 regressions. The effect of the determinants of income is changed when the QOF is introduced. For example, the coefficient on patients per GP/1000 increases from £8,700 to £17,800 between 2004 and 2008. Having a larger practice becomes more lucrative in the post-QOF period. These results are similar to Morris et al. (2011) who present similar models also on the 2008 GP WLS.

The analysis of the determinants of incomes finds that incomes are greater for male GPs in all years and increase with age in 2005 and 2008. Single partner GPs tend to have higher incomes but partnership size is not statistically significant. Dispensing practices have GPs with higher incomes as do GP from practices on the PMS contract. Incomes are lower for non-white GPs and those serving a more deprived population. GPs from practices in rural areas have higher incomes as do GPs in areas where the non-white population is higher.

The predicted incomes from the 2004 income regression are shown in Table 6.4 [p.179]. In 2004, average predicted income was £73,800. The difference in the predicted incomes is caused by changes in the composition of the GP WLS sample in each year. The lower incomes predicted in 2008 could be due to the lower value for the number of patients per GP in that year.

Table 6.4 [p.179] also contains descriptive statistics for both measures of P4P exposure in both years. The mean values for 2005 are 14.6% for GP exposure and 10.8% for practice exposure. The practice income exposure has much less. The 90th percentile for GP income exposure is 20% and there are 11 GPs with income exposure in excess of 40%. Of these 11 outliers, 10 are from single partner practices where the individual GP cannot spread QOF income exposure between multiple partners. The other outlier comes from a practice with two partners but where the maximum QOF income is well above average, resulting in the GPs being particularly exposed. The increase in the payment per point in 2008 resulted in a large increase in P4P exposure for all GPs.

Figure 6.1 [p.213] plots a locally weighted regression (lowess) of the relationship between GP and practice P4P exposure in 2005 and 2008. There is little relationship between the two measures. The figures demonstrate the variation within the measures and the increase in exposure in 2008.

6.6.3 Job satisfaction

Figure 6.2 [p.214] and Figure 6.3 [p.214] plot locally weighted regressions (lowess) of the relationships between P4P exposure and changes in job satisfaction. Positive values on the y axis represent increases in job satisfaction between the two years. There is a slight "n" shape to the fitted line for GP exposure and a slight positive relationship for practice exposure in 2005. These relationships disappear in 2008. These regression plots do not control for confounding factors so multivariate analysis is required.

The regression results shown in Table 6.5-Table 6.7 [p.180-182] use job satisfaction as the dependent variable and measure exposure at the GP level. Table 6.5 [p.180] reports the results from the cross-sectional specification when 2005 and 2008 are regressed individually. This specification suggests that increased P4P exposure in 2005 was associated with lower job satisfaction; a one percentage point increase in exposure reduced job satisfaction by 0.023, statistically significant at 5% (t-ratio -2.00). Put differently, a ten percentage point increase in exposure had a similarly sized effect on job satisfaction as being male or non-white. This change in exposure would be equivalent to movement in the distribution from the 10th to the 90th percentile. This negative association was not found for 2008.

Table 6.6 [p.181] reports the results for the models using changes in job satisfaction as the dependent variable. This has the advantage of using the 2004 survey data as a baseline measure of satisfaction. These models can only be estimated for those GPs responding in both years and hence the sample size is reduced. The first results are for GPs in 2004 and 2005, the second results are for GPs in 2004 and 2008 and the third results are for GPs in all three years. GPs who experience higher levels of P4P exposure in 2005 have positive changes in their job satisfaction. The coefficient of 0.03 is statistically significant at 5% (t-ratio 2.49). The same relationship is not found in the 2004 to 2008 panel or the balanced panel.

Table 6.7 [p.182] reports the results from the continuous DID specification. The coefficient on P4P exposure in 2004 is negative and statistically significant at 0.1%. As mentioned in the econometric methods section, this variable captures unobserved individual heterogeneity. Therefore the association between exposure in 2004 and job satisfaction is the result of a correlation between the variables used to measure exposure and job satisfaction. It does not measure the effect of exposure since the observation is from before the exposure occurs. The interaction term suggests a positive, but not statistically significant, effect of QOF income exposure on job satisfaction in 2005 (t-ratio 1.81). In 2008 the coefficient is negative and not statistically significant (t-ratio -0.55).

Table 6.8-Table 6.10 [p.183-185] replicate the models from Table 6.5-Table 6.7 [p.180-182] but use practice level P4P exposure. In Table 6.8 [p.183] both 2005 and 2008 coefficients on exposure are negative and of similar size to each other, though neither are statistically significant (t-ratios -1.01 and -1.87). In contrast to GP exposure, all the coefficients on practice exposure in Table 6.9 [p.184] are not statistically significant. The coefficients are positive and of similar sizes to those found for GP exposure. Continuous DID models from Table 6.10 [p.185] do not show the same association between practice exposure and job satisfaction in the pre-QOF year. The interaction term is positive for 2005 and negative for 2008. As with GP exposure, these coefficients are not statistically significant.

Reviewing the results from this section we find only statistically significant results for GP level P4P exposure in the cross section and changes models but only for the samples using 2005 data. Cross-sectional models should be approached with caution as they are unable to control for unobserved heterogeneity. The continuous DID

models suggest an association between GP exposure and job satisfaction in 2004. This effect will be confounding the results from the cross-sectional models. Models using changes in job satisfaction and those using continuous DID can control for unobserved heterogeneity and reverse causation. The changes model for GP exposure in 2005 shows an increase in job satisfaction. However, to estimate using changes, the sample size is restricted, and the same effect is not found for these models in 2008. The continuous DID specification is the most preferred. It has the same benefits, in terms of heterogeneity and reverse causation, as the changes model, but does not restrict the sample size.

6.6.4 Measures of GP working lives

The regressions shown in Table 6.11 [p.186] used GP level P4P exposure and 2004 and 2005 data. For brevity the only control variables shown are the income bands, but the full set of controls were included. The pre-QOF term was statistically significant for hours worked per week, life satisfaction and satisfaction with working conditions, choice of method of working remuneration and hours of work. This suggests that P4P exposure in 2004 is capturing some aspect of these dependent variables that our control variables could not. As with job satisfaction models, the interaction term is not statistically significant in any model. The regressions shown in Table 6.12 - Table 6.14 [p.187-189] find the same result. There is no relationship between either measure of P4P exposure and the dependent variables, and this holds for both 2005 and 2008.

6.6.5 Sub-sample analysis

Table 6.15 [p.190] contains the gender and age results for the continuous DID models for both P4P exposure measures and both samples. GP P4P exposure in 2005 increases job satisfaction for female GPs and for GPs aged 40-49. The 2005 results for practice P4P exposure found no relationships. From the 2008 results, GP P4P exposure had a negative effect on job satisfaction for GPs over 60 years old and practice P4P exposure had a positive effect on job satisfaction for GPs under 40 years old.

Table 6.16 [p.191] contains the list size, contract type and partnership size result for the continuous DID models for both P4P exposure measures and both samples. The only statistically significant finding was for GP exposure in 2005 for practices with smaller list sizes.

Table 6.17 [p.192] presents the results of the sub-sample analysis on the characteristics which GPs could change, not age and gender. The significant effect previously observed for small practice in 2005 disappears.

6.6.6 Salaried GPs

Table 6.18 [p.192] contains the results from continuous DID models on only the sample of salaried GPs. In the 2004 and 2005 models there were 22 and 96 salaried GPs respectively. In the 2004 and 2008 models there were 12 and 196 respectively. We do not find a statistically significant effect of practice level exposure for salaried GPs in either sample.

6.7 Conclusion

The job satisfaction of GPs has been studied widely due to the relationship it has with a range of other important factors: patient satisfaction; perceptions of quality; adherence; hours worked by GPs; and GP retention. The connection between job satisfaction and P4P is strongly supported by theories from economics and psychology. There is also a body of empirical research from a range of disciplines using both qualitative and quantitative methods.

We aimed to uncover the effects on working lives and job satisfaction of a large P4P scheme introduced in the UK in 2004. Below we summarise our findings, acknowledge some important strengths and weaknesses, discuss further avenues of research and highlight the possible policy implications of this work.

6.7.1 Summary of findings

We estimated three different model specifications: cross-sectional, changes, and continuous DID. The continuous DID models control for unobserved heterogeneity and reverse causation making these the most reliable results.

With job satisfaction as the dependent variable, no effect of either GP or practice level P4P was found. This result was consistent whether an immediate effect using 2004 and 2005 data was tested, or a long term effect using 2004 and 2008 data.

Similarly, P4P exposure did not affect the 12 other dependent variables we tested: working hours; intentions to quit; life satisfaction; or the nine sub-domains of job satisfaction. Key sub-domains of satisfaction which are associated with the design of the QOF (choice of method of working, remuneration and variety in job) were not affected by P4P exposure.

Splitting the sample into those who might have been more or less affected by P4P exposures shows some evidence of differential effects by age and gender. However, these effects are not consistent over time and exposure measure. Female GPs were positively affected in 2005 only and age effects were present in 2005 and 2008 but for different ages. There are no consistent effects when the sample was split by list size, contact type or partnership size. Models including the un-exposed salaried GPs showed no effect of exposure of job satisfaction in either year.

From the various models and variables we have used to test for the relationship between P4P exposure and job satisfaction, we have found no consistent evidence that such a relationship existed.

6.7.2 Strengths and weaknesses

This study builds on an existing body of research using qualitative and quantitative methods. Previous qualitative studies suffered from small sample sizes but they benefitted from a greater focus on the specific subject of study. Quantitative studies had larger sample sizes but were biased by self-selection into P4P schemes. We feel that our study combines the benefits of both methods while avoiding these problems.

We have a large sample of GPs who could not self-select into or out of P4P. We have a survey which asks a range of questions about job satisfaction and working lives. This survey is linked to administrative data to create a unique linked dataset which has enabled us to measure the amount of P4P exposure for each GP in our sample. We employed an econometric method which estimates the causal effect of increased P4P exposure. In addition to job satisfaction we also model the effects of P4P exposure on intentions to quit, hours worked, life satisfaction and nine sub-domains of jobs satisfaction. We used two measures of P4P exposure and analysed the immediate effects and the effects four years after P4P was introduced. We take advantage of the natural control group of salaried GPs who were not exposed to P4P and compare them to GP partners who were exposed in varying amounts.

However, despite the strengths of this study mentioned above we should discuss some limitations. We had to estimate P4P exposure based on two assumptions. Firstly, that the sharing of QOF income within a practice is related only to a GP's FTE. The sharing rules for QOF incomes are unlikely to be as simple as being purely based on FTE. However, this is likely to play an important role in the allocation and we have taken this into account. Other factors which may influence the allocation include

seniority, tenure and additional responsibilities but it is unclear how these would all be incorporated into our measure accurately.

Our second assumption is that our predictions of GP income had the QOF not been introduced are accurate. We made our predictions of income based on the pre-QOF determinants of income. Provided the sample of GPs in 2004 was unbiased, when compared to 2005 and 2008 samples, these predictions would be accurate.

Another limitation of our study concerns the small numbers of salaried GPs in 2004. These were a relatively new type of GP and their numbers were low before the QOF was introduced. Results which include the salaried GPs are likely to be unreliable due to the small sample in 2004. Salaried GPs act as a natural control group as they are not exposed to P4P. The continuous DID method allows us to measure the effects of treatment intensity so our results without salaried GPs are still valid.

6.7.3 Future research

As the GP WLS has been conducted in 2010, 2012 and 2015 there are additional data points available. However, the QOF was introduced in 2004 and as more time elapses, GPs are more able to react to this P4P exposure and take steps to insulate or expose themselves depending on their preference for P4P. This means that later time points are more likely to be biased by self-selection. Further research would have to use methods which could correct for this but would benefit from a larger sample of salaried GPs.

6.7.4 Policy implications

Our findings suggest GPs are insensitive, in terms of job satisfaction, to the method of payments. Holding income constant, we find no effects of changes to the percentage of income which came from P4P. Policy makers should therefore be able to adjust the composition of GP income without affecting job satisfaction, as long as income levels do not change. For example, plans to remove a third of QOF points in 2014/15 should have no effect on job satisfaction as the income from these points will be transferred to capitation payments (Roland and Campbell, 2014).

7. IS PROVIDER-REPORTED PERFORMANCE UNDER A P4P SCHEME CORRELATED WITH PATIENT-REPORTED QUALITY OF CARE?

7.1 Introduction

A health service designed around the patient was one of the aims and visions of the 2000 NHS Plan (Department of Health, 2000). One of the tools suggested to achieve this was the increased use of P4P. At the time, P4P was in its infancy in healthcare but was growing in popularity. In 2004 the QOF was introduced, the system-wide nature of the scheme and the amount of practitioner pay that was tied to it made it the largest and most expensive P4P scheme of its time (Roland, 2004).

As discussed in Section 2.1, large financial incentives are known to adversely affect intrinsic motivation in the workplace (Ariely et al., 2009; Deci and Ryan, 2000, 1980; Deci, 1971; Deci et al., 1999; Frey and Jegen, 2001). Several years before the QOF was introduced, financial incentives for practitioners to control budgets where shown to reduce patients' satisfaction with their practitioners' opening hours and knowledge of the patient (Dusheiko et al., 2007).

In relation to the QOF, only a relatively small number of studies have analysed patient-level data to consider if the large incentives have impacted on patient experience. The systematic review of QOF evidence by Gillam et al. (2012) identified seven studies which investigated patient experience.

An ethnographic study of four practices in the early years of the QOF found that patients were exposed to a more biomedical style of care, with the disease being treated in place of the patient (Checkland et al., 2008). The implication of this style of treatment was a greater likelihood that the patient would be prescribed medication. This biomedical approach was more likely to be carried out by a nurse, with GPs claiming they turned their focus to more complex cases in response to the QOF.

These conclusions were drawn from interviews with GPs and nurses but not from direct patient experiences. The practices included in the study differed in size and structure but the results were consistent across practices. Despite this consistency, generalisability of these findings is still a concern. Particularly as the authors

acknowledge the sample of practices was influenced heavily by the practices' choice to allow interviewees to make observations.

Other studies have interviewed patients directly to assess their views of the incentivised use of a depression severity questionnaire in the QOF (Dowrick et al., 2009). Patients were generally more positive about the use of the questionnaire when compared with doctors who felt too much emphasis was placed on one mode of treatment. The results cannot comment on the experience of patients' outside of the depression questionnaire.

More generally, patients from practices scoring higher on the QOF self-reported better access to care (Kontopantelis et al., 2010). Surveys conducted in 2003, 2005 and 2007 asked patients about their experiences of primary care (Campbell et al., 2010). The findings suggest that quality of care remained unchanged as the QOF was introduced, and while access increased for patients with chronic diseases, continuity of care was reduced. Patients were able to get appointments faster, but not with their preferred GP.

Patient experience, in terms of satisfaction with access and communication, has been related to QOF overall performance using a linked dataset of the General Practice Patient Survey and the QOF (Llanwarne et al., 2013). A weak association was observed between patient experience domains and practice performance. This suggests that these two areas of healthcare quality are separate, each measuring a different element of performance.

Understanding how the QOF has impacted on the quality of care reported by the general population has not been investigated substantially. The use of patient satisfaction by Llanwarne et al., (2013) is the only study we found. Patient reported quality of care captures two important elements of value to patient centred healthcare: the quality of care received; and the communication about that quality. Low patient reported quality could suggest failings in either of these items. A welfarist perspective dictates that individuals are best placed to assess their own utility. While patients may receive high quality care without realising this, it is still important to measure the patients' perspective on their care.

The QOF directly targeted aspects of patient care, and therefore had the potential to increase the quality of care patients received. It may have also adversely affected patients through the negative consequences of financial incentives. These negative

consequences could present as reduced continuity of care and less holistic care, or present as effort diversion away from incentivised areas. A reduction in performance on non-incentivised areas has been documented in the literature (Campbell et al., 2009; Doran et al., 2011). However, positive spillovers from incentivised areas onto non-incentivised areas has also been observed (Sutton et al., 2010). Therefore, a priori, it would be difficult to hypothesise if the QOF had a negative or positive effect on non-incentivised areas.

The aim of this chapter is to evaluate the QOF by comparing practice-reported quality of care with care reported by the patients. This is achieved by linking individual level data on the quality of primary care from ELSA with practice performance on QOF indicators. This generates a bespoke dataset with which it is possible to test whether investments in healthcare in the form of P4P are visible to the general population. This association between practice and patient reports could vary depending on if the patient reported measure is also incentivised by the QOF. To account for this, quality of care measures from ELSA which are incentivised are used, as well as measures not incentivised. We also use both disease level measures of practice performance and performance on specific indicators. The analyses are performed in levels and first differences. We utilise longitudinal data which Llanwarne et al. (2013) did not have.

We predict that the correlation between practice and patient reported care will be imperfect. This is due to a random measurement error component and aspects of practice performance not being sufficiently communicated to the patient. For example, a patient may misremember having a process measure of quality if this process is not fully explained by the doctor or nurse.

To preface our findings, we found a positive, but modest, correlation between quality of care reported at patient and at practice level for incentivised indicators. This correlation was slightly smaller when using a more aggregated measure of practice performance at disease level. When examining different indicators, the correlation existed only for taking ACE inhibitors and for checking feet, both of which were quality indicators for diabetes. First difference models suggested improvements at practice level have mixed effects at patient level; positive for one out of 15 indicators but negative for two out of 15.

This chapter is structured as follows: a description of the data in Section 7.2; how the two datasets are matched in Section 7.3; the representativeness of the matched sample in Section 7.4; our econometric methods in Section 7.5; results in Section 7.6; and conclusions and discussion in Section 7.7.

7.2 Data

The data used in this study came from two main sources: the QOF; and ELSA. They will be discussed separately in this section, followed by a discussion of the methods used to link the two datasets and finally a discussion of the representativeness of the linked data when compared to the datasets they originated from.

These data were selected due to a number of features which made them appropriate for answering the research question. Firstly, the QOF was the largest and most comprehensive P4P scheme in primary care. A wide range of indicators covered many common disease areas. This feature meant it was likely to find overlap within a survey dataset. Secondly, ELSA focuses on those aged over 50 years and it has been well documented that chronic conditions, such as those incentivised under the QOF, are more prevalent in older individuals (Strong et al., 2005). ELSA also has a large health component to its questionnaire which is more detailed than other surveys such as the British Household Panel Survey or the Health Survey for England. ELSA also benefited from features, discussed below, that would make linking it to the QOF possible.

7.2.1 Data from the QOF

The QOF was discussed in detail in Chapter 3. Here we discuss the aspects most pertinent to the study in this chapter.

Practice performance on QOF indicators is calculated as the percentage of patients treated. Relating back to Chapter 3, the specific measure used is treated patients divided by the sum of eligible patients and exception reported patients, see Equation 3.1.

Exception reported patients are added back into the denominator to create a population measure of performance which reflects who the practice treated and which is more closely related to corresponding measure in ELSA. A respondent in ELSA would not know if they had been exception reported from an indicator for which they were eligible.

Data for QOF exceptions are available for all but the first year (2004/5). In the absence of data on exceptions, the maximum number of eligible patients across all indicators for a disease can be calculated and used in the place of eligible patients plus exceptions (McLean et al., 2006).

This method is applied to all indicators that are used with the exception for the indicator measuring the use of ACE inhibitors (Diabetes 15). For this indicator patients are first required to have a diagnosis of proteinuria or micro-albuminuria meaning that only those with such a diagnosis would form the group of eligible patients for Diabetes 15. As information on proteinuria or micro-albuminuria is not recorded in ELSA, the maximum number of eligible patients across all diabetes indicators was used for all years for diabetes 15. This creates a measure more in line with the corresponding question from ELSA.

A combined measure is created for the two QOF indicators that measure aspects of feet checking (Diabetes 9 and 10) by taking the average of performance across both indicators. This is necessary as these indicators measure two types of feet checking activity: checking peripheral pulses and neuropathy. The feet checking question in ELSA does not specify the type of check performed.

Indicators used in this study are selected if the related disease area is also included in ELSA and if the indicator is present in each of the relevant years of the QOF. The first of these criteria leads to choosing diabetes and hypertension indicators, of which there are 22 and five respectively to select from. Only 12 diabetes indicators and three hypertension indicators are collected in the relevant years. More details on the chosen and excluded indicators are given in Table 7.1 [p.193] and Table 7.2 [p.194].

In addition to indicator specific measures of performance, an overall disease performance measure is created for diabetes and hypertension. This involves calculating the total number of times a practice was successful according to the indicators for a given disease. This figure is divided by the total number of times a practice could have been successful. Success is determined by the number of patients receiving the indicator of quality. The disease performance measures include the 12 diabetes and two hypertension indicators that are consistently collected for the QOF over the period analysed. The combined measures capture a more general measure of quality of care for each disease.

To summarise, QOF data are used to create indicators of practice-reported performance for diabetes and hypertension indicators and for overall performance in these same two disease areas.

7.2.2 Data from ELSA

ELSA is a longitudinal household survey of a representative sample of the English population over 50 years old (Banks et al., 2006). First collected in 2002 and followed up every 2 years, ELSA documents the social, economic and health effects of aging (Marmot et al., 2003). Several modules are included in the survey, each with a main theme. The health module of the core questionnaire contains numerous quality of care indicators covering measures within the control of primary care. Example questions include whether a doctor or nurse ever suggested taking medication to lower blood pressure or whether an A1c test (blood sugar test) had been performed (Steel et al., 2008). A list of these questions is included in Table 7.3 [p.195].

Selected questions on patient reported quality of care are used in this study if they relate to QOF disease areas and are present in all three waves. Three hypertension questions and 12 diabetes questions are used; these are bold in Table 7.3 [p.195]. Each question measures an aspect of care that the practice would have some control over, as opposed to other measures such as BMI which are affected to a larger extent by the patient's lifestyle. Four of the measures used are directly related to a QOF indicator:

- having blood pressure checked (#1 in Table 7.3 [p.195])
- taking an ACE inhibitor (#12 in Table 7.3)
- having an A1c test in the last 12 months (#17 in Table 7.3)
- having feet examined (#18 in Table 7.3)

The other indicators measure aspects of care that are not directly incentivised by the QOF, though there may have been an implicit incentive to target these measures (Sutton et al., 2010). The implicit incentives come from positive spillovers onto activity not incentivised by the QOF but relating to a disease area that is incentivised. For example, doctors do not have a direct financial incentive to provide training to the patient to facilitate self-management of their diabetes. However, there is arguably an implicit incentive to provide such training if it leads to improved self-management and

therefore better practice performance on other indicators. The measures in this category were:

- suggested medication to lower blood pressure (#4 in Table 7.3 [p.195])
- taking medication for high blood pressure (#5 in Table 7.3)
- currently injects insulin (#9 Table 7.3)
- taking medication (#10 in Table 7.3)
- ACE inhibitor or A2 receptor blocker discussed by doctor (#11 in)
- checked for protein in urine in past year by doctor (#13 in Table 7.3)
- been told has protein in urine by doctor (#14 in Table 7.3)
- been told has kidney trouble by doctor (#15 in Table 7.3)
- received training to manage diabetes (#19 in Table 7.3)
- how much knows about managing diabetes (#20 in Table 7.3)

Of the 15 questions used, 14 have "yes", "no" or "do not know" answers. The few respondents answering "do not know" are recoded as "no". These changes affected 559 responses in total. This is approximately 2% of the responses from these questions.

When asked about the amount of knowledge they had about managing their diabetes, respondents answered in five categories ("all", "most", "some", "little" and "none"). This variable is recoded to group "all" and "most" into one group and the remaining categories into a second group.

The selection and recoding resulted in a total of 15 binary variables about the quality of care that patients reported receiving from their practice. Four of these measures were directly related to incentivised activities on the QOF. The remaining measures covered areas not incentivised directly by the QOF.

A summary of the ELSA variables that we use and the corresponding QOF measure of performance is shown in Table 7.4 [p.196]. Four ELSA variables are matched with QOF indicators as well as QOF performance for the entire disease area. The 11 remaining ELSA variables are matched with only overall disease performance as they do not have a direct equivalent as a QOF indicator.

ELSA also provides a range of respondent characteristics to control for factors affecting quality of care or the individuals recall about their quality of care. Age, gender and marital status are used along with self-assessed health and existence of long-standing illness. Self-assessed health is measured on a five point scale (excellent, very good, good, fair and poor). Long-standing illness is measured as not present, non-limiting or limiting. Memory and executive function are measured on a five and six point scale respectively. Dummy variables capture if the observation is from the respondents first, second or third interview as part of ELSA and another dummy captured if the respondent changed GP since the last interview.

7.2.3 Practice characteristics

Additional data on practice characteristics are sourced from the Health and Social Care Information Centre (Health & Social Care Information Centre, 2015c). These additional variables are: practice population size; age proportions of the practice population; number of GPs; rural practice indicator; deprivation measured by the Low Income Scheme Index (LISI); practice contract type; and disease registers. The LISI measures deprivation by how many of the practices' population receive financial support for prescriptions. Disease registers are provided as part of the QOF data (Health & Social Care Information Centre, 2015a). Previous research has shown these variables to be associated with practice performance (Doran et al., 2006). They are included to help account for potential confounding in the relationship between patient- and practice-reported quality of care.

7.2.4 Data structure

ELSA data are at individual level and in a "wide" format: one observation per respondent per interview year with many variables for each respondent. In order to analyse multiple quality of care variables in a single model, these data are reshaped into a "long" format. The resulting dataset is at indicator level with multiple observations per respondent per interview year. For example, respondents providing answers to all 15 quality of care questions relating to diabetes and hypertension would have 15 observations per year. The same structure is used for data from the QOF.

7.3 Matching of ELSA respondents to practices

To link ELSA to QOF we need to first observe the respondent's practice. A nurse visit formed part of the interview in 2004 and 2008. This visit was used to collect more detailed information about the respondent's DNA, blood pressure, height, weight and

other aspects of physical health. Respondents were asked to provide details of the general practice with which they were registered during the nurse visit.

This information was not provided as standard with either the ELSA core questionnaire or with the nurse data. A special request was submitted in order to obtain data for the purpose of using the address information to assign each respondent an official practice identifier. Using these details, it was possible to link respondents to their general practice and therefore data collected about that practice. Linking the two datasets involved matching the practice address information collected as part of the ELSA nurse visit with practice address information available from the Health and Social Care Information Centre (Health & Social Care Information Centre, 2015c). The latter also contained a unique practice identifier that was used to subsequently link to practice QOF data.

Not all respondents to ELSA where given a nurse visit and not all of those with a visit provided address information. Of the 7,666 visited in 2004, 7,332 (95.7%) provided information and of the 8,643 visited in 2008, 8,138 (94.3%) provided information. The quality of the address information varied substantially between respondents. Some gave full addresses and postcodes while others failed to provide more than a partial postcode or street address. This lack of precision may have been the result of out of date records or recall error.

7.3.1 Matching method

The Stata command "reclink" is used to achieve the highest number of possible matches between address information provided by the respondent and official records (Blasnik, 2010). The command allows matches to be made where the values from both datasets are not unique, due to alternative spellings or the address format. The command assigns a score between 0 and 1 to potential matches based on the probability that the match was genuine. Matches with low scores can then be assessed individually. This manual checking is necessary to distinguish between genuine matches and mistakes.

The matching of respondents to practices using practice address information is performed systematically using the 2004 wave of ELSA and data on practice addresses in 2004 and the 2008 wave of ELSA and data on practice addresses in 2008. Approximately 15% of practices in these samples are either based at the same site or on the same street as another practice, which results in them not having a unique

postcode or street address. These practices are identified and separated from practices with unique postcodes so that different matching methods could be applied. This separation yields two sets of official address data for each year, one containing practices with unique postcodes and one containing practices with non-unique postcodes.

7.3.2 Matching results

The first attempt to match practices is made using only the postcode information from ELSA and uses only the practices with unique postcodes. Postcodes are standardised to lower case and have spaces removed. This results in 2,214 matches in 2004 and 3,739 in 2008. After these direct matches had been made, no other deterministic matching was attempted due to the common errors in spelling. Given the number of different address fields (practice name, address line 1-4, postcode), several combinations of these fields are used in successive passes of the reclink command. The details of these combinations and how many successful matches were made are given in Table 7.5 [p.197] and Table 7.6 [p.198].

In all cases, some manual checking of the matches is needed to ensure the accuracy of the match. For example, the reclink command might assign a high score when the name of the practice in both datasets is "the surgery". However, given this is a common practice name, the high score is no guarantee of a genuine match. The next step uses the reclink command on the practices with non-unique postcodes. The final passes of reclink are made after renaming some of the address fields. This is needed due to information being saved in the wrong field. For example, if the street name and number are saved in the field meant for the practice name.

Once all combinations of reclink had been tried, further matches where made manually in Excel. This involved searching for matching address information in any of the address fields.

7.3.3 Additional matches

After matching individuals to practices which could be uniquely identified by their address and postcode, there still remained some unmatched individuals. A different strategy is adopted for these cases. We collapse the dataset containing the non-unique postcode practices to postcode level. This process results in data at the postcode not practice level. We then match respondents who have indicated that they are registered to a practice at one of these postcodes, but could not identify which practice it was.

This is common as some practice names only differ by the presence of a number or letter (practice 1, 2, 3 or a, b, c). With these matches it is necessary to average any practice characteristics across all practices at that postcode. This is a weighted average with the weight proportional to the practice list size.

Given that ELSA is a panel dataset, further matches are made by three forms of imputation. First we impute matches across couples in the same household where one member of the couple has been matched but the other has not. This assumes that couples will register with the same practice. Second we impute across waves in cases where a respondent provides sufficient information to be matched in one wave but not another. This assumes that individuals do not change practices often which is reasonable given rates of practice changes are low (Monitor, 2014). The final imputation is made for the ELSA wave from 2006, where no address information was collected. If a match is made for 2004 or 2008, this is applied to 2006. If respondents are matched to different practices in 2004 and 2008, or reported to have moved home recently, no imputation across waves is made. Table 7.7 [p.199] provides details of the number of individuals surveyed in ELSA waves 2-4 and how many individuals were matched to practices from these waves.

7.4 Representativeness of matched sample

The full ELSA sample is representative of the English population aged over 50 years (Banks et al., 2006), but the sample of ELSA that we match to a practice may not be. A lack of representativeness may be the result of certain individual characteristics increasing or decreasing the likelihood of being matched to a practice. Not having a representative sample may bias our results. For example, if older age respondents are less likely to be matched then the estimated effect of practice performance on individual reported performance would be downwardly (upwardly) biased if older people are also more (less) likely to report in line with practice performance. If we believed that age did not affect how a respondent reported practice performance then the lack of older respondents would not have affected our results.

Similarly, it is not possible to match ELSA respondents to all practices in England, because the ELSA sample size is not large enough. Therefore, it is possible that the sample of included practices is systematically different from the average English practice.

7.4.1 Patients

To check that the sample of ELSA respondents matched to practices is representative of the full ELSA sample we test if any individual characteristics are correlated with being matched. To do this, a dummy variable is created taking the value one for matched individuals and zero for not matched. This new dummy is used as the dependent variable in a Linear Probability Model (LPM) regression. A probit regression was also used and the results were qualitatively the same.

The independent variables in the regression are various individual characteristics that may have been associated with being matched: gender; age; age^2; living in same house as last interview; self-assessed general health; presence of limiting and/or longstanding illness; marital status; memory function; and executive function.

The coefficients in Table 7.8 [p.200] show the change to the probability of an individual being matched to a practice associated with changes in the independent variables. The results of these regressions tell us that matched respondents are:

- More likely to be older up to an certain point, at which further increases in age were associated with a lower probability of being matched
- More likely to live in the same house as during the previous interview
- Less likely to be in fair or poor health rather than good health
- More likely to suffer from a long-standing illness or a limiting long-standing illness
- More likely to be part of a married or co-habiting couple
- More likely to score higher memory functions and less likely to score lower memory functions
- Less likely to have the lowest executive function score
- More likely to be matched in 2004 and less likely in 2006 when compared to 2008

7.4.2 Practices

Another LPM regression is used to test whether included practices differ from practices not included. In this case the dependent variable takes the value one if the practice was included and zero otherwise. The independent variables are: rural/urban location of the practice; practice performance on the QOF; local deprivation; practice

contract type; patient population size; number of GPs at the practice; and the age/gender composition of the practice population.

The coefficients in Table 7.9 [p.201] show the changes in the probability of a practice being included associated with the changes in the independent variables. The results of these regressions tell us that included practices were:

- Less likely to be from an rural area
- More likely to score higher on the QOF
- Less likely to be from a deprived area
- Less likely to be on the PMS contract
- Smaller in terms of patient population size and GPs at the practice but had more patients of ELSA age when defined as closely as the data would allow (aged 45 and over)
- More likely to be included in 2008

7.4.3 Discussion of representativeness

Representativeness of our sample is important for our analysis. Most of the patient characteristics included in the LPM results are significant. A similar lack of representativeness exists for the included practices. Our final samples have some statistically significant differences from the populations they are drawn. We are able to control for these observable differences and the effect they may have on patient-reported quality of care. We are unable to control for if differences affect the strength of association between patient- and practice-reported quality.

However, if unobservable differences exist, we are unable to control for these unless they are time-invariant. The first difference estimation discussed below controls for time-invariant unobservable differences. These limitations and their impacts on our conclusions are discussed further in the conclusion to this chapter.

7.5 Econometric methods

We are primarily interested in the association between what the patient reports and what their practice reports concerning quality of care. A negative association would be worrying, as it would suggest that improvements made at practice level are at the expense of care reported by patients. Finding no significant association would also be worrying, suggesting that practices reporting better quality do not have patients who

share that view. A positive but not perfect correlation would suggest that some proportion of the quality reported at practice level is also reported at patient level.

If we assume that the quality reported at either level is a perfect account of the quality provided/received, then an imperfect correlation may indicate that the marginal patient is not one from our population. Practices may target certain groups of patients first in order to maximise the QOF points they can earn. If the marginal patients are those in less need, then they would be less likely to be in our population of older people. On the other hand, the practice may target patients who are easier to treat, suggesting that the marginal patient may be from our population.

Another factor which would explain a less than prefect association between patient and practice reported quality of care is random measurement error. Once observable and time constant unobservable characteristics have been controlled for, any measurement error in patient reported quality of care is likely to be random with respect to practice performance. This would bias the association downwards.

Various models are estimated with different combinations of dependent and independent variables, all aiming to answer the same central research question. These models are grouped into three categories based on the use of (1) four ELSA indicators and QOF indicator performance, (2) four ELSA indicators and QOF disease performance, and (3) 15 ELSA indicators and QOF disease performance.

Within these three categories alternative specifications are estimated for three difference sets of independent variables for models in levels and one model using first differences. In total, 12 models are estimated.

7.5.1 Four ELSA indicators on four QOF indicator scores

The first category of models used patient-reported quality of care on the set of four ELSA measures as the dependent variable. This is modelled against the corresponding four QOF indicators. In total four variations are presented here, all of which regress a patient-reported measure of quality of care on a practice-reported measure for the same indicator of quality.

$$ELSA_{mijt} = \beta_0 + \beta_1 I + \beta_2 D + \beta_3 QOF_{mjt} + u_{mijt} \qquad 7.1$$

In Equation 7.1 ELSA measures of quality are modelled on: a set of interview dummies I, measuring if the observation is from a 1st, 2nd or 3rd interview; a set of indicator dummies D, measuring which of the four quality of care indicators the observation is from; and a single variable QOF, measuring practice performance in the corresponding QOF area. The data structure allows for observations at quality measure m, individual i, practice j and time t. The error term u is $i.i.d.N(0, \sigma^2)$.

The set of interview dummies captures the effect on patient-reported quality which comes from being interviewed on more occasions. The motivation for their inclusion is that recall might be affected by being asked the same questions in successive interviews. The quality of care indicator dummies allow for patient-reported quality of care to differ by indicator. In these models the four indicators are for: blood pressure test, ACE inhibitor prescription, A1c blood glucose test and feet checking. The dummies allow differences in the proportion of ELSA respondents receiving these tests. The QOF performance variable captures the association between practice-reported and patient reported quality of care. Having a single variable for practice performance means this model has restricted this association not to differ across the four indicators.

$$ELSA_{mijt} = \beta_0 + \beta_1 I + \beta_2 D + \beta_3 D * QOF_{mjt} + u_{mijt} \qquad 7.2$$

Equation 7.2 addresses the restriction imposed on the QOF performance variables by interacting this with quality of care indicator dummies: $D * QOF$. The equation does not differ in other respects. The use of this interaction term will result in four separate coefficients measuring the association between practice-reported and patient-reported quality of care, one for each indicator. This speciation allows for a different association for feet checking when compared to blood pressure tests, for example.

$$ELSA_{mijt} = \beta_0 + \beta_1 I + \beta_2 D + \beta_3 D * QOF_{mjt} + \beta_4' \boldsymbol{X_{ijt}} + u_{mijt} \qquad 7.3$$

Equation 7.3 is a modification of the Equation 7.2 to include characteristics of the individual and of the practice ($\boldsymbol{X}$) to help control for omitted variable bias. For example, we might expect age to the positively related with ELSA quality of care measures. If practices with a higher number of older patients tend to perform better on

QOF indicators, then the omission of age would upwardly bias the coefficient on practice quality β_3.

$$\Delta ELSA_{mijt} = \Delta\beta_0 + \beta_1\Delta I + \beta_2\Delta D + \beta_3\Delta\left(D * QOF_{mjt}\right) + \beta_4'\Delta \boldsymbol{X_{ijt}} + \Delta u_{mijt} \quad 7.4$$

Equation 7.4 estimates the specification of Equation 7.3 but using first differences. This specification achieved two desirable outcomes: it estimates the effect that changes in practices-reported quality of care have on patient-reported quality of care and it controls for unobservable heterogeneity (Wooldridge, 2009, p. 462). The latter is achieved, as any time-invariant characteristics of the individual or practice are removed from the model. This is the case for those characteristics which are observable and those which are unobservable.

All the above models are estimated using the LPM as estimation and interpretation of probit models is confounded by the use of interaction terms (Norton et al., 1998). Standard errors are clustered by practice to account for correlations in patient-reported quality of care by patients in the same practice (Angrist and Pischke, 2008, p. 231).

In addition to these four variations, we regress models including only the practice-level measures of quality of care and: (1) no constant; (2) a constant but no other independent variables; and (3) a constant and interview dummies. This approach shows the effect of including different independent variables on the patient-practice relationship.

7.5.2 Four ELSA indicator on two QOF disease scores

The second category of models uses the same four ELSA indicators as previously discussed but uses the disease level measure of QOF performance. ELSA indicators are measured at the same level and vary over measure m, individual i, practice j and time t. The QOF disease scores do not vary by measure m, instead they vary over the two diseases: diabetes and hypertension. This model is used to test if the relationship between patient- and practice-reported quality of care is weaker or stronger when using a more aggregated measure of practice quality. For this different independent variable, each of the four models discussed previously are estimated.

7.5.3 15 ELSA indicators on two QOF disease scores

The third set of models differs by regressing 15 ELSA quality measures on practice disease level performance. These models include the four indicators with a direct incentive in the QOF and 11 indicators which relate to either diabetes or hypertension quality of care but are not directly incentivised. These models tested whether there was a positive or negative association between QOF disease scores and non-incentivised ELSA indicators i.e. if there were positive spillovers or effort diversion.

7.6 Results

The results from this study are presented in four sections. In Section 7.6.1 are descriptive statistics on patient-reported and practice-reported quality of care. In Section 7.6.2 the results from models using four indicators at patient- and practice-level are presented. In Section 7.6.3 the results from models using four indicators at patient-level and disease performance at practice-level are presented. In Section 7.6.4 the results from models using 15 indicators at patient-level and disease performance at practice-level are presented.

7.6.1 Patient-reported and practice-reported quality of care

Descriptive statistics for the matched ELSA and QOF samples are shown in Table 7.10 [p.201]. ELSA respondents generally report high quality of care on most of the measures, though there was some variation. The average quality of care was lowest for patients being told that they had kidney trouble by a doctor and highest for patients having an A1c test in the last 12 months. For patients with diabetes, Alc testing is recommended (American Diabetes Association, 2003) whereas kidney trouble is not typical for all diabetes patients. These differences highlight that although these questions have been categorised as quality of care measures, some questions are aimed at accessing the severity of the disease. Although these disease severity questions ask the respondent about aspects of their disease which has been revealed to them by their doctor, there exists an implicit quality of care dimension.

Improvements in quality of care over time on these questions are modest or absent. Many questions do not follow a trend over the three years; instead, they fluctuate around similar values. The greatest improvements come from the three questions relating to taking medication for diabetes and hypertension. Diabetes medication improves from 0.62 to 0.69, ACE inhibitor use improves from 0.46 to 0.53, and hypertension medication improves from 0.8 to 0.89.

Approximately 25% of respondents report receiving training for their diabetes, but 80% report they knew all or most of what they needed to know to manage the disease. Sample sizes indicate that more respondents are asked the quality of care questions in later waves of the survey. The prevalence of hypertension is higher than diabetes.

QOF performance is also high on the measures used, and remains high in all years. One exception to this high level of performance comes from the use of ACE inhibitors. This indicator was redefined to remove the requirement for practices to first diagnose patient with proteinuria or micro-albuminuria. Performance is highest on the A1c indicator, as it is for the A1c question in ELSA.

The disease level measures of performance are lower than the individual indicators presented. This is due to aggregating over all indicators for those diseases and not only the ones presented in Table 7.10 [p.202]. Performance increases over time for diabetes but not for hypertension.

7.6.2 Four ELSA indicators on four QOF indicator scores

Table 7.11 [p.203] presents the results for ELSA indicator performance regressed on the corresponding QOF indicator performance. The first model includes QOF performance and no constant term, thus restricting the relationship to pass through at the origin. The coefficient of 1.04 indicates a strong correlation between practice- and patient-reported quality. An increase in practice performance from 0 to 100% results in the probability that a patient reports the same quality of close to 100%.

The second model includes a constant term, allowing for patient-reported quality to average 46.5% when practice-reported quality is zero. The association between practice and patient quality is reduced by half. Including interview dummies in the third model, to explain a possible effect of recall error, did not change this result.

The forth model in Table 7.11 [p.203] is that on Equation 7.1. The addition of indicator dummies reveals that patient-reported quality varies over indicators. Average ACE inhibitor use is indicated by the constant of 48.1%. Other indicators have higher levels with increases of 20.9%, 14.9% and 25.9% for A1c testing, feet checking and blood pressure testing respectively. Having controlled for differences in average quality across indicators, the association between patient- and practice-reported quality is reduced to 0.253. This suggests that a quarter of the practice level performance is reflected at patient level. The fifth model (Equation 7.2) includes QOF

performance interacted with indicator dummies, therefore allowing the quality relationship to differ by indicator. ACE inhibitor remains statistically significant and increases to 0.451. Feet checking is also statistically significant with a coefficient of 0.234. Both A1c and blood pressure tests are not statistically significant.

Coefficients from this fifth model are used to plot the fitted values for ELSA performance against QOF performance in Figure 7.1 [p.215]. The figure helps to show the relationship in the regression results. Taken together with the results from Table 7.11 [p.203], these results suggest that an association exists between practice- and patient-reported quality of care but only for certain indicators.

Table 7.12 [p.204] presents results controlling for various groups of respondent and practice characteristics. The first model repeats the fifth model from Table 7.11 [p.203] for comparison. Subsequent columns have models that include variables measuring the characteristics of: the individual; their household; their health; how well they function; and the practice with which they are linked. The final model in Table 7.12 [p.204] is that of Equation 7.3. The important coefficients in this table are those on the interactions between QOF performance and indicator dummies. As successive groups of control variables are added, the coefficients on ACE inhibitor and feet checking do not change substantially. The overall change for ACE inhibitor is from 0.451 to 0.505 and for feet checking from 0.234 to 0.209.

Results from first difference estimation are shown in Table 7.13 [p.205]. The first model uses one variable to measure practice performance across all indicators. An increase in practice performance by 100% results in an increase of 24.4% in the probability of a patient reporting they received the quality indicator. The second model interacts practice performance to allow for different relationships across different indicators. ACE inhibitor use is the only statistically significant relationship. The coefficient on ACE inhibitor suggests that changes made in the quality of care at practice level are fully observed at patient level. The third model estimates Equation 7.4 and adds the same full set of controls seen in Table 7.12 [p.204]. The inclusion of these variables does not significantly change the estimated effect.

7.6.3 Four ELSA indicator on two QOF disease scores

Table 7.14 [p.206] contains the results from the second category of dependent and independent variables which compared the quality on four ELSA indicators to the disease quality of diabetes and hypertension.

The first model presents the model with no constant, again finding a coefficient close to one. This suggests that disease quality at the practice level is observed at patient level. The inclusion of the constant in second model or the interview dummies in third model does not change this finding. There is evidence that this relationship goes through the origin.

The fourth model estimates the specification of Equation 7.1. The indicator dummies show the difference in indicator quality at patient level, as it does in Table 7.11 [p.203]. The use of interaction terms in the fifth model (Equation 7.2) allows the relationship between ELSA indicator quality and QOF disease quality to differ by indicator. ACE inhibitor use and feet checking are the only statistically significant coefficients. An increase in practice disease performance of 100% is associated with increased probabilities that patients will receive ACE inhibitors and feet checking of 57.2% and 46.1% respectively. These values are both higher than those from Table 7.11 [p.203] particularly for feet checking, but they have lower statistical significance. These associations can be seen in Figure 7.2 [p.216]. Again, an association was found but this varied by indicator.

Table 7.15 [p.206] adds patient and practice characteristics to the model (Equation 7.3). The association between practice disease performance and ACE inhibitor use increases from 0.572 to 0.629 as controls are added. The association for feet checking decreases from 0.461 to 0.338 and is not statistically significant once practice controls are added.

Table 7.16 [p.207] presents first difference models (Equation 7.4) using disease performance. These results did not contain significant coefficients for any of the indicators. Practice level changes in disease performance did not result in changes being observed at the patient level.

7.6.4 15 ELSA indicators on two QOF disease score

Table 7.17 [p.208] contains the results for models using all 15 indicators from ELSA regressed against disease performance for diabetes and hypertension. For brevity not all model specifications which have been presented previously are shown. Instead the first model is the specification using interactions between indicator dummies and disease performance at practice level (Equation 7.2). The second model includes all the control variables previously shown (Equation 7.3). Model three is estimated in first differences (Equation 7.4).

Of the 15 indicators used, only ACE inhibitor and feet checking are statistically significant. The sizes of the coefficients are similar to the previous models. Again, the fitted values in Figure 7.3 [p.216] visualise this result. This model suggests that disease performance was not associated with patient-reported quality of care for any of the additional areas.

The second model includes all control variables (Equation 7.3). This results in feet checking to lose statistical significance, while ACE inhibitor remains the only statistically significant coefficient.

The third model is estimated in first differences with all controls (Equation 7.4). Three of the new indicators relating to diabetes have statistically significant coefficients: taking medication; checking for protein; and receiving training. Increases in disease performance for diabetes results in an increased likelihood of a patient reporting they are taking medication for diabetes. However, the opposite is true for checking for protein and receiving training. This result suggests that increases in disease performance are, in some cases, made at the expense of other indicators which are not incentivised.

7.7 Conclusion

Substantial investments have been made in primary care since the introduction of the QOF in the form of performance payments. This investment was new money and was not diverted from other sources of primary care funding. Given the size of the investment, approximately £1 billion per year, there is an expectation of a measurable response. A considerable body of research has provided evidence of a response by primary care providers but has yet to show if this response was noticed by patients. The aim of this paper was to add to the literature by establishing if a relationship existed between quality of care reported by practices and that which was reported by patients.

7.7.1 Summary of findings

Different measures of quality of care are used in this research. These include individual questions asked to ELSA respondents that overlap with individual QOF indicators and questions that do not overlap. Both of these measures are compared with indicator and disease level practice performance. This method enables the association between patient- and practice-reported quality of care to be analysed at

various levels. Models with and without interaction terms are used so it is possible to separate out the effect of individual indicators. Models using first differences are used to estimate the relationship between changes reported by providers and improvements reported by patients. First differences also control for unobserved heterogeneity.

We find correlations between practice- and patient-reported quality for two indicators: ACE inhibitor use and feet checking. This correlation persists when using practice indicators or measures of disease performance. For ACE inhibitor use, the correlation is robust to controlling for various patient and practice characteristics. In models using first differences only ACE inhibitor was significant when indicator performance was used and neither ACE inhibitor nor feet checking was significant when disease performance was used.

When using the 15 available ELSA indicators, we find evidence that improvements in practice performance were associated with increases in patients reporting they took medication for diabetes. However, the same increases in diabetes performance was associated with decreases in patients reporting their urine had been checked for protein and they had received training to manage their condition. These latter results suggest the incentives of the QOF may have led to poorer care on indicators not incentivised.

To summarise, our models find an association between patient-reported and practice-reported quality of care that is sensitive to the quality indicator in question. Increases in disease performance at practice level are also linked to increases in medication use but decreases in the quality of care relating to protein in urine and the training of patients to manage their condition.

7.7.2 Strengths and weaknesses

The data that were linked together for this research created a unique dataset that has not been available before. This is a main strength of the research. The approach we use differentiates this study from the existing literature in a number of key ways. Llanwarne et al. (2013) also link practice and survey data so their approach is similar to ours. We differentiate through our use of panel data, allowing for estimation in first differences which helps control for unobserved heterogeneity across individuals. A rich set of observable characteristics is also available which is more extensive than other surveys. ELSA also asks questions specifically relating to incentivised measures in the QOF, as well as more general measures not incentivised.

A weakness of this research was that many ELSA and QOF indicators could not be used due to them being absent from some years. This restricted the amount of diseases and indicators we could analyse. The timings of the ELSA interview and QOF data collection are not perfectly contemporaneous. For example, ELSA field work is typically conducted in the autumn, whereas QOF data refers to a whole financial year. However, ELSA questions and QOF indicators typically measure the quality of care in the preceding 12-15 months which means the timings are closely aligned.

The focus of ELSA on older ages is a potential restriction in the external validity of our findings. It may also be the case that recall error is more common in this age group. The inclusion of memory function variables should mitigate the bias caused by recall error that is due to poor respondent memory. However, as chronic conditions are more prevalent in these ages, this population is arguably of more interest than the general population.

7.7.3 Policy implications

The aim of this study was to investigate the potential relationship between the levels of quality reported by practices and the level of quality reported by patients. In doing so we were able to comment on whether patients felt the benefit of the higher quality of care their practice may have provided. Our results suggest that, overall, there was a correlation between practice and patient reported quality. However, this correlation was only significant for two out of 15 indicators when indicators were analysed individually. The implication for policy is that practices are reporting higher levels of performance than patients are reporting, suggesting that some quality is being overstated by the practice or not translated to the patient. The QOF could have more patient experience elements incorporated to help incentivise doctors to focus on this area.

8. CONCLUSION

The aim was to contribute to three areas that had been underrepresented in the P4P literature: the role of non-financial incentives; how providers are affected; and the impact on patients. These gaps in the literature have emerged notwithstanding an extensive body of research into the various effects of P4P.

The first of these gaps is addressed in both Chapters 4 and 5. Chapter 4 uses a large dataset on practice performance to study the relative importance of financial and reputational incentives in determining practice performance. In Chapter 5, a natural experiment is exploited whereby some practices had their PCT reorganised while other practices were unaffected. This created the opportunity to investigate the existence of peer effects, which is an example of a non-financial incentive, acting between practices in the same PCT. Addressing the second gap in the literature was the aim of Chapter 6, and required creating a bespoke linked dataset by combining administrative data on practice P4P income with rich data on the working lives of GPs. With this dataset it became possible to explore the relationship between GPs' exposure to P4P and aspects of their working lives. The third and final gap in the literature is addressed in Chapter 7 by linking together data on practice- and patient-reported quality of care. This results in a unique dataset enabling questions to be answered about the relationship between practice performance on financially incentivised aspects of care, and how patients perceive their own care in these same areas.

Collectively these chapters further our knowledge and understanding on the impact of P4P and provide feedback that can inform the design and implementation of future P4P schemes. Below we provide a discussion of the main findings from each chapter. In the sections that follow on from these findings we identify the strengths and weaknesses of our research, avenues for future research and the policy implications.

8.1 Summary of findings

8.1.1 Chapter 4

Chapter 4 aims to determine the relative impact financial and non-financial incentives have on performance, and if these impacts change over time. This is achieved by exploiting a design feature of the QOF which results in variation in the payoffs to performance. If performance is found to be motivated by non-financial reputational

incentives, the QOF could plausibly extract additional performance from providers by developing this low cost performance incentive. Reputational incentives may be cheaper as they require an initial set up cost, but do not require continued payments.

Chapter 4 utilises nine full years of QOF data for approximately 9000 practices and 60 indicators. This results in a dataset at the indicator level of close to 5 million observations. Analysis at the indicator level allows the within practice variation in the financial and reputational incentives to be modelled.

Our initial analysis restricts the revenue and reputational incentives to be constant in each year of the QOF. When using a sample of balanced indicators which were present in all years of the QOF, the performance response to the revenue incentive is found to be larger than that of the reputational incentive. This would imply that practices are more motivated by financial incentives.

However, allowing each incentive to have a year specific effect reveals that the impact on performance changes over time. This change is best illustrated when the estimates are plotted over time. The revenue incentive is positive and strongest in the first year of the QOF and falls gradually over time. The reputational incentive is negative and not statistically significant in the first year of the QOF and gradually increases over time. The two incentives are significantly different for the first four years of the QOF. They converge in year five and remain statistically indistinguishable through to the final year. This observed convergence in incentives could be explained if reputational incentives were weak in the early years of the QOF because a benchmark in performance had not been established. During these early years, income is more of a performance incentive. Reputation becomes a significant incentive once the performance of peers can be observed.

Our main finding that GPs are not motivated solely by the income associated with P4P is supported by qualitative research in this same area. GPs have expressed a competitive nature with respect to the incentives and a desire to achieve points, both for sake of the points themselves and the income implied (Campbell et al., 2008). A possible effect of high powered financial incentives is that they erode or crowd-out over sources of motivation (Siciliani, 2009). However we do not find evidence for this type of behaviour, instead GP seem to be motivated by both financial and non-financial incentives with the response to the latter increasing over time. This lack of a

detrimental impact on internal motivation is also found in qualitative research (McDonald et al., 2007).

The extent to which GPs perceive an increase in QOF points as an increase in their reputation should be explored in future research. We assert that an increasing response to points incentives over time is evidence that GPs became more concerned about their reputations as the QOF developed. However, it may be an over simplification to treat points and reputation as the same. Instead, a response to points could capture a range of non-financial sources of motivation such as a desire to meet targets or an acknowledgment that achieving points is a by-product of improved clinical care.

8.1.2 Chapter 5

The aim of Chapter 5 was to investigate the presence of peer effects acting between practices in the same PCT. Peer effects are a potentially influential non-financial incentive and their existence would imply that practices are motivated, in part, by their reputation. Reorganisations of practice peer groups have been common in the NHS, but it is unclear if these reorganisations have an effect on the relationships between practices within these groups.

Peer effects are difficult to identify due to a combination of endogeneity, resulting from self-selection into peer groups, and reverse causation, resulting from the reflective nature of peer interactions. Our econometric approach is designed to address these identification problems. One approach takes advantage of the quasi-experimental nature of the PCT reorganisation, namely that not all practices were in a PCTs that were merged. The structure of the data also allows this approach to observe the performance of old and new peers in all time periods. The second approach addresses the reflection problem through the use of lagged peer performance which could not have been influenced by a practice's current performance. Both approaches recognise the need to control for persistence in practice performance. This is done through the use of practice fixed effects or lagged practice performance. The two approaches also differ in the assumption they make regarding whether it is contemporaneous peer performance or lagged peer performance which influences own performance most.

The findings from our first approach show that the impact of old peers is reduced when PCTs are merged. The same reduction is not observed for practices in non-merged PCTs. However, statistical tests show that the difference in effects is

marginally rejected at the 5% level. This reduction in peer effects is an expected result following a dilution of a peer group. We also expect peer effects to increase for new practices entering the peer group but these models do not support this expectation.

Our second approach assumes a stronger influence of peer reflection and addresses this using the lag of peer performance. We find that the effect of lagged performance for the new peers becomes statistically significant one year after the reorganisation. This effect is not found in the second year after the reorganisation.

The sensitivity analysis finds that peer effects do not differ substantially according to PCT management or the relative size of the merged PCTs. This provides further evidence that the relationships we find are due to peer effects but a more reliable and accurate measure of PCT management quality could provide more convincing evidence. The limitations of the available measure of PCT quality means that it is not possible to completely rule out that our results are caused by changes in managerial effort.

Peer effects are also found to be statistically significant when using a measure of peer performance which quantifies the number of poor performers in a peer group. This is not found when quantifying the number of well performing peers.

Our analysis finds convincing evidence that the PCT reorganisation had an impact on the relationship between the performance of an individual practice and the performance of their peers. The qualitative evidence that GPs are competitive (Campbell et al., 2008), would suggest that the performance of their peers might affect their own performance. It is less clear whether the whole PCT is an appropriate peer group and if QOF points are the appropriate measure of peer performance.

Additional research could be conducted to classify peer groups based on geographical proximity or by smaller social networks such as groups of practices engaged in PBC. These groups would arguably have stronger peer effects but they would not be changed by the PCT reorganisation meaning a source of exogenous variation would be lost. These smaller groups may be driving the apparent peer effects which we have attributed to PCTs.

8.1.3 Chapter 6

In Chapter 6 we seek to discover if variation in the level of income exposed to P4P has an effect on the job satisfaction and working lives of GPs. The proportion of GP

income that became dependent on performance was substantial and unprecedented, prompting an interest into its impact on GPs. We add to a largely qualitative literature base with a quantitative study, using linked data, of the impact of P4P exposure on various working life measures.

The estimation methods used to approach this research question consisted of a series of cross-sectional models, models using changes in job satisfaction and a continuous DID specification. The last of these methods is the most robust as it controls for unobserved GP heterogeneity and reverse causation. These features are achieved as the models use the post-QOF level of P4P exposure in the pre-QOF period to absorb any unobserved relationship between exposure and our dependent variables. The interaction of P4P exposure with a post-QOF dummy variable captures any additional relationship that emerges when the P4P exposure takes effect. It is not possible for this interacted coefficient to be confounded by reverse causation because any impact on exposure from our dependent variable is captured in the pre-QOF observation. Prior to estimation, we had already mitigated the effect of the potentially endogenous relationship between satisfaction and P4P exposure through the use of maximum QOF income at practice level and predicted non-QOF income at GP level. Endogeneity would have caused identification problems as more satisfied GPs, or those working longer hours, could have performed better on the QOF and hence resulted in increased exposure.

Using continuous DID, we tested for an effect of P4P exposure on a range of working lives variables: overall job satisfaction; sub-domains of job satisfaction; life satisfaction; intentions to quit; and working hours. We also measure P4P exposure at the GP and at the practice level, in addition to testing the effect of increases in exposure occurring between 2005 and 2008. Overall, and across each of these different approaches, we find no effect of P4P exposure on working lives.

Further analysis was performed to investigate if an effect existed for sub-samples of the GP population. An overall finding that was not statistically significant could have been caused by an aggregation of positive and negative findings for sub-samples: for example male and female GPs. The sample was split by gender, age, list size, contract type and partnership size. In the majority of cases there was no difference in the estimated effect of P4P exposure between sub-samples. The differences that existed for age and gender were inconsistent with respect to measures of P4P exposure and

year. Overall, there is no convincing evidence that the impact of P4P exposure differs according to these GP characteristics.

The continuous DID models were also estimated on samples including only salaried GPs, who do not receive P4P income and are considered to have zero direct P4P exposure. We also found no effect of P4P exposure in this placebo group. There were only a very small number of salaried GPs included in these samples.

The absence of an effect resulting from considerable P4P exposure was not expected given the theory and literature reviewed in Chapter 6. Qualitative research has shown that GPs were affected by the QOF in terms of increased morale and better work-life balance but disliked increased monitoring (Campbell et al., 2008). Similarly higher incomes and fewer working hours were positive effects associated with P4P while increased administrative and managerial work were negative effects (Maisey et al., 2008). Despite this evidence suggesting both positive and negative effects of the QOF, we find no quantitative impacts on the measures of job satisfaction and working lives used in this study. An explanation for this absence could be due to the QOF impacting on these measures of working lives but having a constant effect which does not vary by exposure to P4P. The implication would be that our sample of GPs experienced equal changes in their working lives, and these changes were unrelated to differential P4P exposure.

8.1.4 Chapter 7

The aim of Chapter 7 was to estimate the relationship between the quality of care reported by practices and by patients using a linked dataset. The value of this question is rooted in understanding how patients perceive their care, and if P4P has negatively impacted on these perceptions. To the extent that a patient's perception of care is a true reflection of the care they receive, we are also able to comment not only on perceptions, but the care delivered.

Several combinations of dependent and independent variables were analysed using two main estimation specifications: cross-sectional and first difference regression models. The results of Chapter 7 will be summarised according to these two specifications, as they differ in their limitations and overall findings.

The findings from the cross-sectional models vary depending on the choice of dependent and independent variables. When patient reports are regressed on

individual QOF indicators corresponding to the same measure of quality of care, it is only for ACE inhibitor use and feet checking where a statistically significant and positive correlation is found. When the models switch to use a disease level measure of practice quality of care, it is only ACE inhibitor use that maintains a statistically significant and positive correlation. Two statements can be made following these results. Firstly, practices that report higher levels of quality of care relating to feet checking and ACE inhibitor use also have patients who report higher levels of quality of care in these areas. Secondly, practices that report higher performance across multiple quality of care indicators for diabetes have patients who reported higher use of ACE inhibitors. These statements must be caveated by highlighting that these models only identify correlations between practices and patients. They cannot claim that higher patient-reported quality of care is the result of the practice performance.

The models using first differences also exhibit different results depending on the measure of practice quality of care. Results from corresponding indicators at patient and practice level reveal that when practices report improvements in ACE inhibitor use, patients also report improvements. However, when disease level measures of quality of care are used, the relationship for ACE inhibitor use does not persist. Improvements in disease level measures of quality of care are associated with improvements in patients reporting taking medication for diabetes but also associated with declines in patients reporting protein checks and receiving training. These results have different implications to the cross-sectional results as they remove the effect of time-invariant unobserved heterogeneity at the patient level. The cross-sectional models are only able to relate high levels of one measure with high levels of another. First difference models related *changes* (increases or decreases) in quality and therefore can determine if patients respond to increases in quality and not only pre-existing high quality.

In conclusion, our main results relating to Chapter 7 are that improvements in diabetes disease management at practice level lead to patients reporting greater use of medication and fewer patients reporting that their protein levels are checked and that they have received training. The first of these three patient measures is incentivised by the QOF, the latter two measures are not.

It should be noted that our finding that better diabetes management, as measured by the QOF, leads to patients reporting fewer checks for protein may not be an indication

of poor quality of care or effort diversion. It is possible that these checks are carried out and not communicated to the patient merely because it is not necessary to communicate this information. Patients with diabetes will routinely provide urine samples and the reasons for doing so vary and may not be fully explained in every case.

The findings for diabetes training and medication use are perhaps more plausible and can be linked to qualitative research in the area. Specifically, interviews with GPs have revealed an increase focus on a biomedical approach to primary care (Checkland et al., 2008). This approach would imply a greater reliance on medication for treatment, possibly at the expensive of a more holistic approach. This qualitative research lends support to our findings that when practices improve their QOF scores, their patients tend to report more medication and less training.

8.1.5 Overall thesis

When viewed as a single and unified body of research, this thesis can be condensed into a number of key statements which are supported by the exposition, analysis and results presented.

Practices are motivated by both financial and non-financial incentives, though the relative size of these incentives is not fixed during our study period. These non-financial incentives are, in part, the result of peer effects acting on practices within the same PCT. When financial incentives are dominant over reputational incentives, GPs' working lives are unaffected by exposure to P4P. During this same period, practices focus on incentivised measures of quality more than they focus on non-incentivised measures. We cannot comment on whether the latter two statements would remain true once the financial incentive converges with the reputational incentive.

8.2 Strengths and weaknesses

A strength which is common across Chapters 6 and 7 is the use of linked data to answer research questions which could not be answered using existing individual datasets. The relationship between practice- and patient-reported quality of care can only be investigated when data are available at both levels. Similarly, data on practice P4P income and GPs' working lives are not provided in any single dataset. Linking data to create the means to answer these research questions is how we are able to add to the literature in unique ways.

In Chapters 4 and 5, linked datasets are not used. Originality is achieved through the use of existing routinely collected data, combined with original study design. In the case of Chapter 5, this centres on the use of the natural experiment that resulted from the partial reorganisation of PCTs in 2006. This provides a scenario in which peer effects can be identified. With respect to Chapter 4, a long time period is used which captures several exogenous changes in revenue and reputational incentives. These changes provide the variation across time, practices and diseases which are necessary to identify responses to these incentives and if these responses change.

Where possible confounding factors which could introduce bias into our results are removed or mitigated. In all chapters, methods are used which remove the effects of time-invariant unobserved heterogeneity. The use of these methods is possible as we have panel data for all chapters.

Despite these strengths, limitations remain which restrict what this thesis is able to add to the literature on P4P. The setting for all analysis is the UK NHS which means that generalising our results for different healthcare systems is not straightforward. In the NHS, GPs have the dual role of provider and gatekeeper; although this is not unique to the NHS, nor is it a ubiquitous feature of healthcare systems worldwide. This dual role may impact on GP's internal motivation, with particular respect to their rationing behaviour, and hence effect how GPs react to financial incentives. Along these lines, the majority of GPs are self-employed contractors working in for-profit practices. In healthcare systems which do not share these features, our results and conclusions may not be as relevant.

Besides the institutional setting, another limitation which impacts on all empirical chapters is the non-experimental introduction of the QOF. There was no trial period or control group which necessitates the use of non-experimental econometric methods. With these methods it becomes more difficult to generate causal inference.

The QOF was also almost universally adopted by practices in the UK. This is an indication that practices have a preference for this type of P4P which is likely to influence the results. Alternative schemes, with lower levels of uptake, may not demonstrate the same impacts on patients and providers.

Although we feel that the most appropriate sources of data are used for this thesis, there are limitations to these data which impact on our conclusions. ELSA asks

respondents about a range of quality of care measures, but only a relatively small number of these overlap with QOF measures. This means that only two disease areas can be analysed in Chapter 7. An interesting feature of UK primary care in recent years has been the increased use of salaried GPs. These GPs are not generally in receipt of P4P income. The GP WLS in 2004, 2005 and 2008 only have small samples of salaried GPs which restricted the analysis which could be performed with respect to comparing salaried GPs to GP partners.

8.3 Future research

Each of the datasets that are used in this thesis continue to be collected. ELSA data have recently become available for 2010 and 2012. There are GP WLS for 2010, 2012 and 2015. QOF data are published yearly. This means that future research could involve updating the analysis of this thesis using the more recent data.

Samples from the 2010, 2012 and 2015 GP WLS have a higher proportion of salaried GPs so it is possible that the use of these later surveys can be used to investigate the impact of P4P on an unexposed group of GPs. However, later surveys are likely to have been affected to a greater extent by self-selection into salaried GP roles. Therefore, simply repeating the analysis of Chapter 6 on later surveys will result in endogenous effects.

In 2012, PCTs were dismantled and replaced with CCGs. In terms of the number of practices, these newer commissioning bodies were in-between the pre- and post-2006 PCTs. Repeating the analysis of Chapter 5 and factoring in this additional reorganisation could be a future study. The reorganisation into CCGs allowed practices to self-select their new peer group so the change was not exogenous in the same way to the PCT reorganisation that we analyse.

The QOF has been subjected to further changes in design since the analysis in Chapter 4. In 2014/15 a large amount of QOF income was shifted from P4P to capitation. It would be interesting to analyse the effect of this change to the design of the QOF and assess to what extent it impacted on the relationship identified in Chapter 4.

8.4 Policy implications

The research questions of this thesis have been approached using a combination of linked data, natural experiments and the novel use of existing datasets. These various methods have been necessitated due to the non-experimental way in which the QOF was introduced. Although we have been able to add to the literature, the research in this thesis, and much of the literature discussed in Chapter 3, could have been more robust had the QOF been designed to aid its own evaluation.

If the QOF had been randomly introduced in certain practices, the analyses from Chapters 6 and 7 could have incorporated a control group of practices which were unaffected. This would have facilitated a direct comparison of P4P practices with non-P4P practices. Although commonplace in medicine, this type of experimental design is much rarer in healthcare policy (Finkelstein and Taubman, 2015). There are examples such as the Oregon Health Insurance Experiment in the US whereby health insurance was allocated via a lottery to a group of previously uninsured individuals (Finkelstein et al., 2012). Another example is the Diabetes Care Project in Australia whereby financial incentives for diabetes care were allocated to certain practices in a cluster RCT (McKinsey and Company, 2015). These examples of random policy introduction are evidence of a changing approach in healthcare.

When the QOF was introduced, the use of RCTs in policy was far less common. However, the changes made to the QOF since its introduction could have been implemented in a controlled manner. Certain practices could have been exposed to new indicators before all practices were. This would have posed few practical difficulties as the existing financial infrastructure and data collection systems could have been utilised. For future changes to QOF, and new schemes of a similar nature, a randomised and controlled approach should be adopted where possible.

Our conclusions regarding the existence of reputational incentives from Chapter 4 imply that an increased focus on these non-financial incentives could result in observable changes in behaviour. We recommend that P4P schemes should promote reputational incentives explicitly through the use of published performance data and/or league tables. This information should be easily accessible and understandable to maximise its use. We also conclude that it may be possible to reduce the size of payments linked to performance over time. The falling financial incentives may be offset by the increasing reputational incentives.

The analysis of peer effects in Chapter 5 highlights a potential negative consequence of encouraging practices to compare themselves with their peers. Practices were observed to be pulled down by poor performing peers to a greater extent than they were pulled up by well performing peers. This finding suggests that caution should be applied when promoting practice comparisons. We advise an approach whereby the high performing practices are highlighted, as opposed to naming and shaming the practices with low performance. We also suggest that caution should be applied when forming new organisations, such as PCTs. Peer effects are influenced by such changes, and policy makers should be mindful of this influence.

We find individual GPs are unaffected by the proportion of income at risk. Of particular relevance, we do not find an effect on job satisfaction or working hours. Therefore, policy makers should not expect to be able to use exposure to P4P in the QOF as a policy lever to affect these working life variables. The concerns expressed about the QOF being too powerful and having payments that were too high are unfounded.

For some measures of quality of care there is no observable correlation between the quality reported by the patients and that which is reported by the practices. This leads us to advise that the design of P4P schemes should account for the potential for financial incentives to have negative effects of doctor-patient communication. The QOF has elements which measure patient experience, but it is arguable that this is under represented. Care should also be taken in the design of schemes to mitigate any deterioration in performance for non-incentivised measures of quality of care.

REFERENCES

Abbt, N., Alderson, S., 2014. Why doesn't anyone want to be a GP: and what can we do about it? Br. J. Gen. Pract. 64, 579–579. doi:10.3399/bjgp14X682345

Allen, T., Fichera, E., Sutton, M., 2014a. Can payers use prices to improve quality? Evidence from english hospitals. Health Econ. doi:10.1002/hec.3121

Allen, T., Mason, T., Whittaker, W., 2014b. Impacts of pay for performance on the quality of primary care. Risk Manag. Healthc. Policy 7, 113–20. doi:10.2147/RMHP.S46423

Allen, T., Sutton, M., Cookson, R., 2014c. How Do Economic Incentive Schemes Influence Rationing Decisions by Primary Care Physicians?, in: Danis, M., Hurst, S.A., Fleck, L., Forde, R., Slowther, A. (Eds.), Fair Resourse Allocation and Rationing at the Bedside. Oxford University Press, p. 480.

Alshamsan, R., Majeed, A., Ashworth, M., Car, J., Millett, C., 2010. Impact of pay for performance on inequalities in health care: systematic review. J. Health Serv. Res. Policy 15, 178–184. doi:10.1258/jhsrp.2010.009113

American Diabetes Association, 2003. Tests of Glycemia in Diabetes. Diabetes Care 26, S106–S108. doi:10.2337/diacare.26.2007.S106

An, L.C., Bluhm, J.H., Foldes, S.S., Alesci, N.L., Klatt, C.M., Center, B.A., Nersesian, W.S., Larson, M.E., Ahluwalia, J.S., Manley, M.W., 2008. A randomized trial of a pay-for-performance program targeting clinician referral to a state tobacco quitline. Arch. Intern. Med. 168, 1993–9. doi:10.1001/archinte.168.18.1993

Angrist, J.D., Pischke, J.-S.S., 2008. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press, Princeton, NJ.

Appleby, J., Harrison, T., Hawkins, L., Dixon, A., 2012. Payment by Results. How can payment systems help to deliver better care? [WWW Document]. URL http://www.kingsfund.org.uk/sites/files/kf/field/field_publication_file/payment-by-results-the-kings-fund-nov-2012.pdf (accessed 6.6.15).

Ariely, D., Gneezy, U., Loewenstein, G., Mazar, N., 2009. Large Stakes and Big Mistakes. Rev. Econ. Stud. 76, 451–469. doi:10.1111/j.1467-937X.2009.00534.x

Arrow, K.J., 1963. Uncertainty and the Welfare Economics of Medical Care. Am. Econ. Rev. 53, 941–973.

Axon, R.N., Williams, M. V, 2011. Hospital readmission as an accountability measure. JAMA 305, 504–5. doi:10.1001/jama.2011.72

Baker, G.P., 1992. Incentive Contracts and Performance Measurement. J. Polit. Econ. 100, 598–614.

Balia, S., Jones, A.M., 2008. Mortality, lifestyle and socio-economic status. J. Health Econ. 27, 1–26. doi:10.1016/j.jhealeco.2007.03.001

Banks, J., Marmot, M., Oldfield, Z., Smith, J.P., 2006. Disease and Disadvantage in the United States and in England. JAMA 295, 2037–2045. doi:10.1001/jama.295.17.2037

Bevan, G., Hood, C., 2006. What's measured is what matters: targets and gaming in the English public health care system. Public Adm. 84, 517–538. doi:10.1111/j.1467-9299.2006.00600.x

Blasnik, M., 2010. RECLINK: Stata module to probabilistically match records. Stat. Softw. Components.

BMA, 2015. Local Medical Committees [WWW Document]. URL http://bma.org.uk/about-the-bma/how-we-work/local-representation/local-medical-committees (accessed 3.3.15).

BMA, 2011. Focus on QOF payments. April 2011 [WWW Document]. URL file:///C:/Users/msrajta2/Downloads/Focus_on_QOF_payments_-_April_2011.pdf (accessed 10.2.15).

BMA, NHS Employers, 2012. Quality and Outcomes Framework for 2012/13 [WWW Document]. URL http://www.nhsemployers.org/~/media/Employers/Documents/Primary care contracts/QOF/2012-13/Quality and Outcomes Framework for 2012-13 - Guidance for PCOs and Practices.pdf (accessed 5.1.15).

BMA, NHS Employers, 2011. Quality and Outcomes Framework guidance for GMS contract 2011/12 [WWW Document]. URL http://www.nhsemployers.org/~/media/Employers/Documents/Primary care contracts/QOF/2011-12/Quality and Outcomes Framework guidance for GMS contract 2011-12 - Delivering investment in general practice.pdf (accessed 5.1.15).

BMA, NHS Employers, 2009. Quality and Outcomes Framework guidance for GMS contract 2009/10 [WWW Document]. URL http://www.nhsemployers.org/~/media/Employers/Documents/Primary care contracts/QOF/2009-10/Quality and Outcomes Framework guidance for GMS contract 2009-10 - Delivering investment in general practice.pdf (accessed 5.1.15).

BMA, NHS Employers, 2008. Quality and Outcomes Framework guidance for GMS contract 2008/09 [WWW Document]. URL http://www.nhsemployers.org/~/media/Employers/Documents/Primary care contracts/QOF/2008-09/Quality and Outcomes Framework guidance for GMS contract 200809 - Delivering investments in general practice.pdf (accessed 5.1.15).

BMA, NHS Employers, 2006. Revisions to the GMS contract 2006/07 [WWW Document]. URL

http://www.nhsemployers.org/~/media/Employers/Documents/Primary care contracts/QOF/2006-07/Revisions to the GMS contract 200607 - Delivering investments in general practice.pdf (accessed 5.1.15).

BMA, NHS Employers, NHS England, 2014. 2014/15 General Medical Services (GMS) Contract Quality and Outcomes Framework (QOF) [WWW Document]. URL http://www.nhsemployers.org/~/media/Employers/Documents/Primary care contracts/QOF/2014-15/14-15 General Medical Services contract - Quality and Outcomes Framework.pdf (accessed 5.1.15).

Bojke, C., Gravelle, H., Wilkin, D., 2001. Is bigger better for primary care groups and trusts? BMJ 322, 599–602. doi:10.1136/bmj.322.7286.599

Booth, A.L., Frank, J., 1999. Earnings, Productivity, and Performance Related Pay. J. Labor Econ. 17, 447–463. doi:10.1086/209927

Briesacher, B.A., Field, T.S., Baril, J., Gurwitz, J.H., 2009. Pay-for-performance in nursing homes. Health Care Financ. Rev. 30, 1–13.

Buchbinder, S.B., Wilson, M., Melick, C.F., Powe, N.R., 1999. Estimates of costs of primary care physician turnover. Am. J. Manag. Care 5, 1431–8.

Burgess, S., Ratto, M., 2003. The Role of Incentives in the Public Sector: Issues and Evidence. Oxford Rev. Econ. Policy 19, 285–300. doi:10.1093/oxrep/19.2.285

Cable, D., Judge, T., 1994. Pay preferences and job search decisions: A person-organization fit perspective. Pers. Psychol. 47, 317–348.

Cadsby, C.B., Song, F., Tapon, F., 2007. Sorting and incentive effects of pay for performance: an experimental investigation. Acad. Manag. J. 50, 387–405. doi:10.5465/AMJ.2007.24634448

Campbell, S., Hannon, K., Lester, H., 2011. Exception reporting in the Quality and Outcomes Framework: views of practice staff - a qualitative study. Br. J. Gen. Pract. 61, 183–9. doi:10.3399/bjgp11X567117

Campbell, S., Kontopantelis, E., Reeves, D., Valderas, J.M., Gaehl, E., Small, N., Roland, M., 2010. Changes in Patient Experiences of Primary Care During Health Service Reforms in England Between 2003 and 2007. Ann. Fam. Med. 8, 499–506. doi:10.1370/afm.1145

Campbell, S., Reeves, D., Kontopantelis, E., Sibbald, B., Roland, M., 2009. Effects of Pay for Performance on the Quality of Primary Care in England. N. Engl. J. Med. 361, 368–378. doi:10.1056/NEJMsa0807651

Campbell, S.M., McDonald, R., Lester, H., 2008. The experience of pay for performance in English family practice: a qualitative study. Ann. Fam. Med. 6, 228–34. doi:10.1370/afm.844

Candace, I., Buchan, J., Xavier, S., 2009. NHS workforce planning Limitations and possibilities. London.

Card, D., 1992. Using regional variation in wages to measure the effects of the federal minimum wage. Ind. Labor Relations Rev. 46, 22–37.

Carrell, S.E., Fullerton, R.L., West, J.E., 2009. Does Your Cohort Matter? Measuring Peer Effects in College Achievement. J. Labor Econ. 27, 439–464.

Checkland, K., Coleman, A., Harrison, S., Hiroeh, U., 2009. "We can"t get anything done because...': making sense of "barriers" to Practice-based Commissioning. J. Health Serv. Res. Policy 14, 20–6. doi:10.1258/jhsrp.2008.008043

Checkland, K., Harrison, S., McDonald, R., Grant, S., Campbell, S., Guthrie, B., 2008. Biomedicine, holism and general medical practice: responses to the 2004 General Practitioner contract. Sociol. Health Illn. 30, 788–803. doi:10.1111/j.1467-9566.2008.01081.x

Clark, A.E., Kristensen, N., Westergård-Nielsen, N., 2009. Job Satisfaction and Co-worker Wages: Status or Signal? Econ. J. 119, 430–447. doi:10.1111/j.1468-0297.2008.02236.x

Craig, P., Cooper, C., Gunnell, D., Haw, S., Lawson, K., Macintyre, S., Ogilvie, D., Petticrew, M., Reeves, B., Sutton, M., Thompson, S., 2012. Using natural experiments to evaluate population health interventions: new Medical Research Council guidance. J. Epidemiol. Community Health 66, 1182–6. doi:10.1136/jech-2011-200375

Deci, E.L., 1971. Effects of externally mediated rewards on intrinsic motivation. J. Pers. Soc. Psychol. 18, 105–115. doi:10.1037/h0030644

Deci, E.L., Koestner, R., Ryan, R.M., 1999. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. Psychol. Bull. 125, 627–68; discussion 692–700.

Deci, E.L., Ryan, R.M., 2000. The "What" and "Why" of Goal Pursuits: Human Needs and the Self-Determination of Behavior. Psychol. Inq. 11, 227–268. doi:10.1207/S15327965PLI1104_01

Deci, E.L., Ryan, R.M., 1980. The empirical exploration of intrinsic motivational processes. Adv. Exp. Soc. Psychol. 13, 39–80.

Department of Health, 2011. Payment by Results Guidance for 2011-12 [WWW Document]. URL http://webarchive.nationalarchives.gov.uk/20130507170152/https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/151911/dh_126157.pdf.pdf (accessed 10.29.14).

Department of Health, 2007. World Class Commissioning [WWW Document]. URL http://webarchive.nationalarchives.gov.uk/20130107105354/http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_080956?IdcService=GET_FILE&dID=155367&Rendition=Web (accessed 2.24.15).

Department of Health, 2003a. Investing in general practice The new General Medical Services Contract [WWW Document]. URL http://www.nhsemployers.org/~/media/Employers/Documents/SiteCollectionDocuments/gms_contract_cd_130209.pdf (accessed 4.20.15).

Department of Health, 2003b. Delivering investment in general practice: implementing the new GMS contract [WWW Document]. URL http://webarchive.nationalarchives.gov.uk/20130107105354/http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_4070231.pdf (accessed 8.30.15).

Department of Health, 2000. The NHS Plan: a plan for investment, a plan for reform [WWW Document]. URL http://dera.ioe.ac.uk/4423/1/04055783.pdf (accessed 8.30.15).

DeVoe, J., Fryer Jr, G.E., Hargraves, J.L., Phillips, R.L., Green, L.A., 2002. Does career dissatisfaction affect the ability of family physicians to deliver high-quality patient care? J. Fam. Pract. 51, 223–8.

DiMatteo, M.R., Sherbourne, C.D., Hays, R.D., Ordway, L., Kravitz, R.L., McGlynn, E.A., Kaplan, S., Rogers, W.H., 1993. Physicians' characteristics influence patients' adherence to medical treatment: results from the Medical Outcomes Study. Health Psychol. 12, 93–102.

Doran, T., 2015. Incentivising improvements in health care delivery: a response to Adam Oliver. Heal. Econ. Policy Law 1–6.

Doran, T., Campbell, S., Fullwood, C., Kontopantelis, E., Roland, M., 2010. Performance of small general practices under the UK's Quality and Outcomes Framework. Br. J. Gen. Pract. 60, e335–44. doi:10.3399/bjgp10X515340

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. doi:10.1056/NEJMsa055505

Doran, T., Kontopantelis, E., Fullwood, C., Lester, H., Valderas, J.M., Campbell, S., 2012. Exempting dissenting patients from pay for performance schemes: retrospective analysis of exception reporting in the UK Quality and Outcomes Framework. BMJ 344, e2405–e2405. doi:10.1136/bmj.e2405

Doran, T., Kontopantelis, E., Reeves, D., Sutton, M., Ryan, A.M., 2014. Setting performance targets in pay for performance programmes: what can we learn from QOF? BMJ 348, g1595–g1595. doi:10.1136/bmj.g1595

Doran, T., Kontopantelis, E., Valderas, J.M., Campbell, S., Roland, M., Salisbury, C., Reeves, D., 2011. Effect of financial incentives on incentivised and non-incentivised clinical activities: longitudinal analysis of data from the UK Quality and Outcomes Framework. BMJ 342. doi:10.1136/bmj.d3590

Dowrick, C., Leydon, G.M., McBride, A., Howe, A., Burgess, H., Clarke, P., Maisey, S., Kendrick, T., 2009. Patients' and doctors' views on depression severity

questionnaires incentivised in UK quality and outcomes framework: qualitative study. BMJ 338, b663. doi:10.1136/bmj.b663

Dusheiko, M., Doran, T., Gravelle, H., Fullwood, C., Roland, M., 2011a. Does Higher Quality of Diabetes Management in Family Practice Reduce Unplanned Hospital Admissions? Health Serv. Res. 46, 27–46. doi:10.1111/j.1475-6773.2010.01184.x

Dusheiko, M., Gravelle, H., Martin, S., Rice, N., Smith, P., 2011b. Does better disease management in primary care reduce hospital costs? Evidence from English primary care. J. Health Econ. 30, 919–932. doi:10.1016/j.jhealeco.2011.08.001

Dusheiko, M., Gravelle, H., Yu, N., Campbell, S., 2007. The impact of budgets for gatekeeping physicians on patient satisfaction: Evidence from fundholding. J. Health Econ. 26, 742–762. doi:10.1016/j.jhealeco.2006.12.003

Easterlin, R., 1974. Does economic growth improve the human lot? Some empirical evidence. Nations households Econ. growth.

Eggleston, K., 2005. Multitasking and mixed systems for provider payment. J. Health Econ. 24, 211–223. doi:10.1016/j.jhealeco.2004.09.001

Eijkenaar, F., 2012. Pay for Performance in Health Care: An International Overview of Initiatives . Med. Care Res. Rev. 69 , 251–276. doi:10.1177/1077558711432891

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 (New. York). null. doi:10.1016/j.healthpol.2013.01.008

Ellis, R.P., McGuire, T.G., 1986. Provider behavior under prospective reimbursement : Cost sharing and supply. J. Health Econ. 5, 129–151. doi:Doi: 10.1016/0167-6296(86)90002-0

Emmert, M., Eijkenaar, F., Kemter, H., Esslinger, A.S., Schöffski, O., 2012. Economic evaluation of pay-for-performance in health care: a systematic review. Eur. J. Health Econ. 13, 755–67. doi:10.1007/s10198-011-0329-8

Eriksson, T., Villeval, M.C., 2008. Performance-pay, sorting and social motivation. J. Econ. Behav. Organ. 68, 412–421. doi:10.1016/j.jebo.2007.10.003

Fichera, E., Nikolova, S., Sutton, M., 2014. Comparative Performance Evaluation: Information on Quality, in: Culyer, A.J. (Ed.), Encyclopedia of Health Economics. Elsevier, pp. 111–116. doi:10.1016/B978-0-12-375678-7.01313-4

Fichera, E., Sutton, M., 2011. State and self investments in health. J. Health Econ. 30, 1164–73. doi:10.1016/j.jhealeco.2011.09.002

Finkelstein, A., Taubman, S., 2015. Health care policy. Randomize evaluations to improve health care delivery. Science 347, 720–2. doi:10.1126/science.aaa2362

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J.P., Allen, H., Baicker, K., 2012. The Oregon Health Insurance Experiment: Evidence from the First Year. Q. J. Econ. 127, 1057–1106.

Freeborn, D.K., 2001. Satisfaction, commitment, and psychological well-being among HMO physicians. West. J. Med. 174, 13–8.

Frey, B.S., Jegen, R., 2001. Motivation Crowding Theory. J. Econ. Surv. 15, 589–611. doi:10.1111/1467-6419.00150

Frick, B., Prinz, J., Winkelmann, K., 2003. Pay inequalities and team performance. Int. J. Manpow. 24, 472–488. doi:10.1108/01437720310485942

Gagne, M., Deci, E., 2005. Self-determination theory and work motivation. J. Organ. Behav. 26, 331–362. doi:10.1002/job.322

Ganster, D.C., Kiersch, C.E., Marsh, R.E., Bowen, A., 2011. Performance-Based Rewards and Work Stress. J. Organ. Behav. Manage. 31, 221–235. doi:10.1080/01608061.2011.619388

Gaynor, M., Moreno-Serra, R., Propper, C., 2010. Death by Market Power: Reform, Competition and Patient Outcomes in the National Health Service. Natl. Bur. Econ. Res. Work. Pap. Ser. No. 16164.

Gené-Badia, J., Escaramis-Babiano, G., Sans-Corrales, M., Sampietro-Colom, L., Aguado-Menguy, F., Cabezas-Peña, C., de Puelles, P.G., 2007. Impact of economic incentives on quality of professional life and on end-user satisfaction in primary care. Health Policy 80, 2–10. doi:10.1016/j.healthpol.2006.02.008

Georgellis, Y., Lange, T., Ileana Petrescu, A., Simmons, R., 2008. Human resource management practices and workers' job satisfaction. Int. J. Manpow. 29, 651–667. doi:10.1108/01437720810908947

Gillam, S., Siriwardena, A.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, 461–468. doi:10.1370/afm.1377

Gillam, S., Steel, N., 2013. The Quality and Outcomes Framework--where next? BMJ 346, f659–f659. doi:10.1136/bmj.f659

Gosden, T., Forland, F., Kristiansen, I.S., Sutton, M., Leese, B., Giuffrida, A., Sergison, M., Pedersen, L., 2000. Capitation, salary, fee-for-service and mixed systems of payment: effects on the behaviour of primary care physicians. Cochrane database Syst. Rev. CD002215. doi:10.1002/14651858.CD002215

Gravelle, H., Dusheiko, M., Sutton, M., 2002. The demand for elective surgery in a public system: time and money prices in the UK National Health Service. J. Health Econ. 21, 423–449. doi:Doi: 10.1016/s0167-6296(01)00137-0

Gravelle, H., Hole, A.R., Hossaun, M.I., 2008. Response bias in job satisfaction surveys: English general practitioners [WWW Document]. Univeristy York

Discuss. Pap. Econ. URL https://www.york.ac.uk/media/economics/documents/discussionpapers/2008/0824.pdf (accessed 7.20.15).

Gravelle, H., Santos, R., Siciliani, L., 2014. Does a hospital's quality depend on the quality of other hospitals? A spatial econometrics approach. Reg. Sci. Urban Econ. 49, 203–216. doi:10.1016/j.regsciurbeco.2014.09.005

Gravelle, H., Sutton, M., Ma, A., 2010. Doctor Behaviour under a Pay for Performance Contract: Treating, Cheating and Case Finding? Econ. J. 120, 129–156. doi:10.1111/j.1468-0297.2009.02340.x

Grembowski, D., Ulrich, C.M., Paschane, D., Diehr, P., Katon, W., Martin, D., Patrick, D.L., Velicer, C., 2003. Managed Care and Primary Physician Satisfaction. J. Am. Board Fam. Med. 16, 383–393. doi:10.3122/jabfm.16.5.383

Guryan, J., Kroft, K., Notowidigdo, M., 2007. Peer Effects in the Workplace: Evidence from Random Groupings in Professional Golf Tournaments.

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, 836–841.

Guthrie, B., Saultz, J.W., Freeman, G.K., Haggerty, J.L., 2008. Continuity of care matters. BMJ 337, a867. doi:10.1136/bmj.a867

Haas, J.S., Cook, E.F., Puopolo, A.L., Burstin, H.R., Cleary, P.D., Brennan, T.A., 2000. Is the professional satisfaction of general internists associated with patient satisfaction? J. Gen. Intern. Med. 15, 122–8.

Hadley, J., Mitchell, J.M., Sulmasy, D.P., Bloche, M.G., 1999. Perceived financial incentives, HMO market penetration, and physicians' practice styles and satisfaction. Health Serv. Res. 34, 307–21.

Haggerty, J.L., Reid, R.J., Freeman, G.K., Starfield, B.H., Adair, C.E., McKendry, R., 2003. Continuity of care: a multidisciplinary review. BMJ 327, 1219–21. doi:10.1136/bmj.327.7425.1219

Hamilton, B.H., Nickerson, J.A., Owan, H., 2003. Team Incentives and Worker Heterogeneity: An Empirical Analysis of the Impact of Teams on Productivity and Participation. J. Polit. Econ. 111, 465–497. doi:10.1086/374182

Hann, M., Goudie, R., Sutton, M., Gravelle, H., Sibbald, B., 2009. Working Conditions and Job Satisfaction of GPs: Findings from the Fifth National GP Worklife Survey. National Primary Care Research and Development Centre.

Hann, M., Reeves, D., Sibbald, B., 2011. Relationships Between Job Satisfaction, Intentions to Leave Family Practice and Actually Leaving Among Family Physicians in England. Eur. J. Public Health 21, 499–503. doi:10.1093/eurpub/ckq005

Harrison, M.J., Dusheiko, M., Sutton, M., Gravelle, H., Doran, T., Roland, M., 2014. Effect of a national primary care pay for performance scheme on emergency hospital admissions for ambulatory care sensitive conditions: controlled longitudinal study. BMJ 349, g6423. doi:10.1136/bmj.g6423

Haubrich, J.G., 1994. Risk Aversion, Performance Pay, and the Principal-Agent Problem. J. Polit. Econ. 102, 258–276. doi:10.2307/2138661

Hausman, J.A., 1978. Specification Tests in Econometrics. Econometrica 46, 1251–1271. doi:10.2307/1913827

Health & Social Care Information Centre, 2015a. The Quality and Outcomes Framework [WWW Document]. URL http://www.qof.hscic.gov.uk/ (accessed 3.24.15).

Health & Social Care Information Centre, 2015b. QOF results for your GP practice [WWW Document]. URL http://qof.hscic.gov.uk/search/ (accessed 7.20.15).

Health & Social Care Information Centre, 2015c. Health & Social Care Information Centre [WWW Document]. URL http://www.hscic.gov.uk/ (accessed 7.1.15).

Health & Social Care Information Centre, 2006. GP Earnings and Expenses Enquiry 2004/05 [WWW Document]. URL http://www.hscic.gov.uk/pubs/gpearnex0405 (accessed 5.14.14).

Health & Social Care Information Centre, 2005a. National Quality and Outcomes Framework Statistics for England 2004/05 [WWW Document]. URL http://www.hscic.gov.uk/catalogue/PUB01946/qof-eng-04-05-rep.pdf (accessed 7.1.15).

Health & Social Care Information Centre, 2005b. National Quality and Outcomes Framework Exception Reporting Statistics for England 2005/06 [WWW Document]. URL http://www.hscic.gov.uk/catalogue/PUB01960/qual-outc-fram-exce-rep-05-06-rep.pdf (accessed 7.2.15).

Health & Social Care Information Centre, 2004. Annex A: Quality indicators - Summary of points [WWW Document]. URL http://www.hscic.gov.uk/catalogue/PUB01946/qof-eng-04-05-defi-indi-anx.pdf (accessed 8.14.13).

Holmstrom, B., 1982. Moral Hazard in Teams. Bell J. Econ. 13, 324–340.

Holmstrom, B., Milgrom, P., 1991. Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. J. Law, Econ. Organ. 7, 24–52.

Holt, C.A., Laury, S.K., 2002. Risk Aversion and Incentive Effects. Am. Econ. Rev. 92, 1644–1655.

Houle, S.K.D., McAlister, F.A., Jackevicius, C.A., Chuck, A.W., Tsuyuki, R.T., 2012. Does performance-based remuneration for individual health care practitioners

affect patient care?: a systematic review. Ann. Intern. Med. 157, 889–99. doi:10.7326/0003-4819-157-12-201212180-00009

House of Commons Health Committee, 2005. Changes to Primary Care Trusts. Second Report of Session 2005–06 [WWW Document]. URL http://www.publications.parliament.uk/pa/cm200506/cmselect/cmhealth/646/646.pdf (accessed 11.4.14).

Hoxby, C.M., Leigh, A., 2004. Pulled Away or Pushed Out? Explaining the Decline of Teacher Aptitude in the United States. Am. Econ. Rev. 94, 236–240. doi:10.1257/0002828041302073

Huang, J., Yin, S., Lin, Y., Jiang, Q., He, Y., Du, L., 2013. Impact of pay-for-performance on management of diabetes: a systematic review. J. Evid. Based. Med. 6, 173–84. doi:10.1111/jebm.12052

Iversen, T., Ma, C., 2011. Market conditions and general practitioners' referrals. Int. J. Health Care Finance Econ. 11, 245–265. doi:10.1007/s10754-011-9101-y

Jacob, B.A., Levitt, S.D., 2003. Rotten Apples: An Investigation of The Prevalence and Predictors of Teacher Cheating. Q. J. Econ. 118, 843–877. doi:10.1162/00335530360698441

Jacobs, R., Gutacker, N., Mason, A., Goddard, M., Gravelle, H., Kendrick, T., Gilbody, S., Aylott, L., Wainwright, J., 2015. Do higher primary care practice performance scores predict lower rates of emergency admissions for persons with serious mental illness? An analysis of secondary panel data. Heal. Serv Deliv Res 3.

Jensen, M.C., 2001. Paying People to Lie: The Truth About the Budgeting Process. SSRN Electron. J. doi:10.2139/ssrn.267651

Jensen, M.C., Meckling, W.H., 1976. Theory of the firm: Managerial behavior, agency costs and ownership structure. J. financ. econ. 3, 305–360. doi:10.1016/0304-405X(76)90026-X

Jensen, M.C., Murphy, K.J., 1999. CEO Incentives: It's Not How Much You Pay, But How. SSRN Electron. J. doi:10.2139/ssrn.146148

Kaarboe, O., Siciliani, L., 2011. Multi-tasking, quality and pay for performance. Health Econ. 20, 225–38. doi:10.1002/hec.1582

Kankaanranta, T., Nummi, T., Vainiomäki, J., Halila, H., Hyppölä, H., Isokoski, M., Kujala, S., Kumpusalo, E., Mattila, K., Virjo, I., Vänskä, J., Rissanen, P., 2007. The role of job satisfaction, job dissatisfaction and demographic factors on physicians' intentions to switch work sector from public to private. Health Policy 83, 50–64. doi:10.1016/j.healthpol.2006.11.010

Keele, L., Kelly, N., 2005. Dynamic Models for Dynamic Theories: The Ins and Outs of Lagged Dependent Variables. Polit. Anal. 14, 186–205. doi:10.1093/pan/mpj006

Knez, M., Simester, D., 2001. Firm-Wide Incentives and Mutual Monitoring at Continental Airlines. J. Labor Econ. 19, 743–772. doi:10.1086/322820

Kontopantelis, E., Buchan, I., Reeves, D., Checkland, K., Doran, T., 2013a. Relationship between quality of care and choice of clinical computing system: retrospective analysis of family practice performance under the UK's quality and outcomes framework. BMJ Open 3, e003190–. doi:10.1136/bmjopen-2013-003190

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, 1117–1136. doi:10.1111/j.1475-6773.2011.01362.x

Kontopantelis, E., Reeves, D., Valderas, J.M., Campbell, S., Doran, T., 2013b. Recorded quality of primary care for patients with diabetes in England before and after the introduction of a financial incentive scheme: a longitudinal observational study. BMJ Qual. Saf. 22, 53–64. doi:10.1136/bmjqs-2012-001033

Kontopantelis, E., Roland, M., Reeves, D., 2010. Patient experience of access to primary care: identification of predictors in a national patient survey. BMC Fam. Pract. 11, 1–15. doi:10.1186/1471-2296-11-61

Kontopantelis, E., Springate, D.A., Ashworth, M., Webb, R.T., Buchan, I.E., Doran, T., 2015. Investigating the relationship between quality of primary care and premature mortality in England: a spatial whole-population study. BMJ 350, h904–h904. doi:10.1136/bmj.h904

Langdown, C., Peckham, S., 2014. The use of financial incentives to help improve health outcomes: is the quality and outcomes framework fit for purpose? A systematic review. J. Public Health (Oxf). 36, 251–8. doi:10.1093/pubmed/fdt077

Lavy, V., Silva, O., Weinhardt, F., 2012. The Good, the Bad, and the Average: Evidence on Ability Peer Effects in Schools. J. Labor Econ. 30, 367–414.

Lazear, E.P., 2000. Performance Pay and Productivity. Am. Econ. Rev. 90, 1346–1361.

Le Grand, J., 2003. Motivation, Agency, and Public Policy: Of Knights and Knaves, Pawns and Queens. Oxford University Press, Oxford. doi:10.1093/0199266999.001.0001

Lester, H., Campbell, S., 2010. Developing Quality and Outcomes Framework (QOF) indicators and the concept of'QOFability'. Qual. Prim. Care 18, 103–109.

Lester, H., Matharu, T., Mohammed, M.A., Lester, D., Foskett-Tharby, R., 2013. Implementation of pay for performance in primary care: a qualitative study 8 years after introduction. Br. J. Gen. Pract. 63, e408–15. doi:10.3399/bjgp13X668203

Lewis, R., Curry, N., Dixon, M., 2007. Practice-based Commissioning: from good idea to effective practice. London.

Lindenauer, P.K., Remus, D., Roman, S., Rothberg, M.B., Benjamin, E.M., Ma, A., Bratzler, D.W., 2007. Public reporting and pay for performance in hospital quality improvement. N. Engl. J. Med. 356, 486–96. doi:10.1056/NEJMsa064964

Linn, L.S., Brook, R.H., Clark, V.A., Davies, A.R., Fink, A., Kosecoff, J., 1985. Physician and patient satisfaction as factors related to the organization of internal medicine group practices. Med. Care 23, 1171–8.

Llanwarne, N.R., Abel, G.A., Elliott, M.N., Paddison, C.A.M., Lyratzopoulos, G., Campbell, J.L., Roland, M., 2013. Relationship Between Clinical Quality and Patient Experience: Analysis of Data From the English Quality and Outcomes Framework and the National GP Patient Survey. Ann. Fam. Med. 11, 467–72. doi:10.1370/afm.1514

Ma, C.A., 1994. Health Care Payment Systems: Cost and Quality Incentives. J. Econ. Manag. Strateg. 3, 93–112. doi:10.1111/j.1430-9134.1994.00093.x

Maisey, S., Steel, N., Marsh, R., Gillam, S., Fleetcroft, R., Howe, A., 2008. Effects of payment for performance in primary care: qualitative interview study. J. Health Serv. Res. Policy 13, 133–9. doi:10.1258/jhsrp.2008.007118

Manski, C.F., 1993. Identification of Endogenous Social Effects: The Reflection Problem. Rev. Econ. Stud. 60, 531. doi:10.2307/2298123

Marmot, M., Banks, J., Blundell, R., Lessof, C., Nazroo, J., 2003. Health, wealth and lifestyles of the older population in England: ELSA 2002. London.

Maynard, A., Bloor, K., 2010. Will financial incentives and penalties improve hospital care? BMJ 340, c88–c88. doi:10.1136/bmj.c88

McCafferty, S., Williams, I., Hunter, D., Robinson, S., Donaldson, C., Bate, A., 2012. Implementing world class commissioning competencies. J. Health Serv. Res. Policy 17 Suppl 1, 40–8. doi:10.1258/jhsrp.2011.011104

McDonald, R., Harrison, S., Checkland, K., Campbell, S., Roland, M., 2007. Impact of financial incentives on clinical autonomy and internal motivation in primary care: ethnographic study. BMJ 334, 1357. doi:10.1136/bmj.39238.890810.BE

McGregor, W., Jabareen, H., O'Donnell, C.A., Mercer, S.W., Watt, G.C., 2008. Impact of the 2004 GMS contract on practice nurses: a qualitative study. Br. J. Gen. Pract. 58, 711–9. doi:10.3399/bjgp08X342183

McKinsey and Company, 2015. Evaluation Report of the Diabetes Care Project [WWW Document]. URL http://www.health.gov.au/internet/main/publishing.nsf/Content/302DF0372F537A43CA257E35000138E8/$File/DCP Evaluation Report.pdf (accessed 6.3.15).

McLean, G., Sutton, M., Guthrie, B., 2006. Deprivation and quality of primary care services: evidence for persistence of the inverse care law from the UK Quality and Outcomes Framework. J. Epidemiol. Community Health 60, 917–922. doi:10.1136/jech.2005.044628

Meacock, R., Kristensen, S.R., Sutton, M., 2014. The cost-effectiveness of using financial incentives to improve provider quality: a framework and application. Health Econ. 23, 1–13. doi:10.1002/hec.2978

Mehran, H., 1995. Executive compensation structure, ownership, and firm performance. J. financ. econ. 38, 163–184. doi:10.1016/0304-405X(94)00809-F

Melville, A., 1980. Job satisfaction in general practice implications for prescribing. Soc. Sci. Med. Part A Med. Psychol. Med. Sociol. 14, 495–499. doi:10.1016/S0271-7123(80)80055-1

Misra-Hebert, A.D., Kay, R., Stoller, J.K., 2004. A Review of Physician Turnover: Rates, Causes, and Consequences. Am. J. Med. Qual. 19, 56–66. doi:10.1177/106286060401900203

Monitor, 2014. Discussion document following Monitor's call for evidence on GP services [WWW Document]. URL https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/288429/GPDiscussionDocFinal_0.pdf (accessed 3.4.15).

Morris, S., Goudie, R., Sutton, M., Gravelle, H., Elliott, R., Hole, A.R., Sibbald, B., Ska, D., Ma, A., Sibbald, B., Skåtun, D., 2011. Determinants of general practitioners' wages in England. Health Econ. 20, 147–160. doi:10.1002/hec.1573

Mullen, K.J., Frank, R.G., Rosenthal, M.B., 2009. Can You Get What You Pay For? Pay-For-Performance and the Quality of Healthcare Providers.

Nair, H.S., Manchanda, P., Bhatia, T., 2006. Asymmetric Peer Effects in Physician Prescription Behavior: The Role of Opinion Leaders. Stanford University, Graduate School of Business.

National Audit Office, 2008. NHS Pay Modernisation: New contracts for general practice services in England [WWW Document]. URL http://www.nao.org.uk/wp-content/uploads/2008/02/0708307.pdf (accessed 11.3.14).

NATPACT, 2004. Rewarding you fairly for quality in primary care [WWW Document]. URL http://www.natpact.info/uploads/Prevalence_guide.pdf (accessed 4.20.15).

NHS, 2015. NHS Choices [WWW Document]. URL http://www.nhs.uk/Service-Search (accessed 3.31.15).

NHS, 2010. GMS statement of financial entitlements [WWW Document]. URL https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/216071/dh_112959.pdf (accessed 2.10.15).

NHS, 2006. Primary Care Trusts (Establishment and Dissolution) (England) Order 2006. England.

NHS Confederation, 2011. The legacy of primary care trusts [WWW Document]. URL http://www.nhsconfed.org/~/media/Confederation/Files/Publications/Documents/The_legacy_of_PCTs.pdf (accessed 3.3.15).

NHS Employers, 2015. Quality and Outcomes Framework [WWW Document]. URL http://www.nhsemployers.org/PayAndContracts/GeneralMedicalServicesContract/QOF/Pages/QualityOutcomesFramework.aspx (accessed 3.24.15).

NHS Employers, BMA, 2011. The Quality and Outcomes Framework frequently asked questions [WWW Document]. URL http://www.nhsemployers.org/~/media/Employers/Documents/Primary care contracts/QOF/2012-13/UK QOF Frequently asked questions - June 2012.pdf (accessed 2.10.15).

NHS Employers, BMA, 2009. Prevalence arrangements from 2009/10 [WWW Document]. URL http://www.nhsemployers.org/~/media/Employers/Documents/Primary care contracts/QOF/2009-10/Prevalence arrangements from 200910 Joint letter to PCTs.pdf (accessed 4.20.15).

NHS England, BMA, NHS Employers, 2014. 2013/14 general medical services (GMS) contract quality and outcomes framework (QOF) [WWW Document]. URL http://www.nhsemployers.org/~/media/Employers/Documents/Primary care contracts/QOF/2013-14/2013-14 general medical services contract quality and outcomes framework - Guidance for GMS contract 201314.pdf (accessed 5.1.15).

Norbury, M., Fawkes, N., Guthrie, B., 2011. Impact of the GP contract on inequalities associated with influenza immunisation: a retrospective population-database analysis. Br. J. Gen. Pract. 61, e379–85. doi:10.3399/bjgp11X583146

Norton, E.C., Lindrooth, R.C., Ennett, S.T., 1998. Controlling for the endogeneity of peer substance use on adolescent alcohol and tobacco use. Health Econ. 7, 439–53.

O'Brien, R.M., 2007. A Caution Regarding Rules of Thumb for Variance Inflation Factors. Qual. Quant. 41, 673–690. doi:10.1007/s11135-006-9018-6

Oliver, A., 2015. Incentivising improvements in health care delivery. Heal. Econ. Policy Law FirstView, 1–17.

Pathman, D.E., Konrad, T.R., Williams, E.S., Scheckler, W.E., Linzer, M., Douglas, J., 2002. Physician job satisfaction, dissatisfaction, and turnover. J. Fam. Pract. 51, 593.

Pfeffer, J., Langton, N., 1993. The Effect of Wage Dispersion on Satisfaction, Productivity, and Working Collaboratively: Evidence from College and University Faculty. Adm. Sci. Q. 38, 382–407. doi:10.2307/2393373

Porter, L., Lawler, E., 1968. Managerial attitudes and performance. R.D. Irwin, Homewood Ill.

Prendergast, C., 1999. The Provision of Incentives in Firms. J. Econ. Lit. 37, 7–63.

Propper, C., Sutton, M., Whitnall, C., Windmeijer, F., 2008. Did "Targets and Terror" Reduce Waiting Times in England for Hospital Care? B. E. J. Econom. Anal. Policy 8. doi:10.2202/1935-1682.1863

Raleigh, V.S., Klazinga, N., 2013. Future proofing the Quality and Outcomes Framework. BMJ 346.

Ratto, M., Burgess, S., Croxson, B., Jewitt, I., Propper, C., 2001. Team-Based Incentives in the NHS: An Economic Analysis.

Rice, N., Sutton, M., 1998. Drinking patterns within households: the estimation and interpretation of individual and group variables. Health Econ. 7, 689–99.

Rittenhouse, D.R., Mertz, E., Keane, D., Grumbach, K., 2004. No exit: an evaluation of measures of physician attrition. Health Serv. Res. 39, 1571–88. doi:10.1111/j.1475-6773.2004.00304.x

Roland, M., 2004. Linking Physicians' Pay to the Quality of Care — A Major Experiment in the United Kingdom. N. Engl. J. Med. 351, 1448–1454. doi:10.1056/NEJMhpr041294

Roland, M., Campbell, S., 2014. Successes and Failures of Pay for Performance in the United Kingdom. N. Engl. J. Med. 370, 1944–1949. doi:10.1056/NEJMhpr1316051

Rosenthal, M.B., Frank, R.G., Li, Z., Epstein, A.M., 2005. Early experience with pay-for-performance: from concept to practice. JAMA 294, 1788–93. doi:10.1001/jama.294.14.1788

Roski, J., Jeddeloh, R., An, L., Lando, H., Hannan, P., Hall, C., Zhu, S.-H., 2003. The impact of financial incentives and a patient registry on preventive care quality: increasing provider adherence to evidence-based smoking cessation practice guidelines. Prev. Med. (Baltim). 36, 291–299. doi:10.1016/S0091-7435(02)00052-X

Ryan, A.M., 2009. Effects of the Premier Hospital Quality Incentive Demonstration on Medicare Patient Mortality and Cost. Health Serv. Res. 44, 821–842. doi:10.1111/j.1475-6773.2009.00956.x

Ryan, R.M., Deci, E.L., 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am. Psychol. 55, 68–78.

Sacerdote, B., 2001. Peer Effects with Random Assignment: Results for Dartmouth Roommates. Q. J. Econ. 116, 681–704. doi:10.1162/00335530151144131

Santos, R., Gravelle, H., Propper, C., 2015. Does quality affect patients' choice of Doctor? Evidence from England. Econ. J. Online only. doi:10.1111/ecoj.12282

Scott, A., Coote, W., 2010. Do regional primary-care organisations influence primary-care performance? A dynamic panel estimation. Health Econ. 19, 716–729. doi:10.1002/hec.1509

Scott, A., Gravelle, H., Simoens, S., Bojke, C., Sibbald, B., 2006. Job Satisfaction and Quitting Intentions: A Structural Model of British General Practitioners. Br. J. Ind. Relations 44, 519–540. doi:10.1111/j.1467-8543.2006.00511.x

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, in: Cochrane Database of Systematic Reviews. John Wiley & Sons, Ltd.

Serumaga, B., Ross-Degnan, D., Avery, A.J., Elliott, R.A., Majumdar, S.R., Zhang, F., Soumerai, S.B., 2011. Effect of pay for performance on the management and outcomes of hypertension in the United Kingdom: interrupted time series study. BMJ Br. Med. J. 342. doi:10.1136/bmj.d108

Sibbald, B., Enzer, I., Cooper, C., Rout, U., Sutherland, V., 2000. GP Job Satisfaction in 1987, 1990 and 1998: Lessons for the Future? Fam. Pract. 17, 364–371. doi:10.1093/fampra/17.5.364

Siciliani, L., 2009. Paying for performance and motivation crowding out. Econ. Lett. 103, 68–71. doi:DOI: 10.1016/j.econlet.2009.01.022

Sicsic, J., Le Vaillant, M., Franc, C., 2012. Intrinsic and extrinsic motivations in primary care: an explanatory study among French general practitioners. Health Policy 108, 140–8. doi:10.1016/j.healthpol.2012.08.020

Smith, J., Mays, N., 2005. Primary care trusts: do they have a future? BMJ 331, 1156–7. doi:10.1136/bmj.331.7526.1156

Smith, P., York, N., 2004. Quality Incentives: The Case Of U.K. General Practitioners. Health Aff. 23, 112–118. doi:10.1377/hlthaff.23.3.112

Sobanja, M., 2009. What is world class commissioning [WWW Document]. URL http://www.medicine.ox.ac.uk/bandolier/painres/download/whatis/What_is_WC_Comm.pdf (accessed 12.16.14).

Soljak, M., Calderon-Larrañaga, A., Sharma, P., Cecil, E., Bell, D., Abi-Aad, G., Majeed, A., 2011. Does higher quality primary health care reduce stroke admissions? A national cross-sectional study. Br. J. Gen. Pract. 61, e801–7. doi:10.3399/bjgp11X613142

Steel, N., Bachmann, M., Maisey, S., Shekelle, P., Breeze, E., Marmot, M., Melzer, D., 2008. Self reported receipt of care consistent with 32 quality indicators: national population survey of adults aged 50 or more in England. BMJ 337, a957–a957. doi:10.1136/bmj.a957

Steel, N., Willems, S., 2010. Research learning from the UK Quality and Outcomes Framework: a review of existing research. Qual. Prim. Care 18, 117–125.

Stevenson, B., Wolfers, J., 2008. Economic Growth and Subjective Well-Being: Reassessing the Easterlin Paradox.

Strong, K., Mathers, C., Leeder, S., Beaglehole, R., 2005. Preventing chronic diseases: how many lives can we save? Lancet 366, 1578–82. doi:10.1016/S0140-6736(05)67341-2

Sutton, M., Elder, R., Guthrie, B., Watt, G., 2010. Record rewards: the effects of targeted quality incentives on the recording of risk factors by primary care providers. Health Econ. 19, 1–13. doi:10.1002/hec.1440

Sutton, M., Godfrey, C., 1995. A grouped data regression approach to estimating economic and social influences on individual drinking behaviour. Health Econ. 4, 237–47.

Teljeur, C., Thomas, S., O'Kelly, F.D., O'Dowd, T., 2010. General practitioner workforce planning: assessment of four policy directions. BMC Health Serv. Res. 10, 148. doi:10.1186/1472-6963-10-148

Twardella, D., Brenner, H., 2007. Effects of practitioner education, practitioner payment and reimbursement of patients' drug costs on smoking cessation in primary care: a cluster randomised trial. Tob. Control 16, 15–21. doi:10.1136/tc.2006.016253

Van Herck, P., Smedt, D. De, 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. doi:10.1186/1472-6963-10-247

Vernooij-Dasssen, M.J., Faber, M.J., Rikkert, M.G.O., Koopmans, R.T., Achterberg, T. van, Braat, D.D., Raas, G.P., Wollersheim, H., 2009. Dementia care and labour market: The role of job satisfaction. Aging Ment. Heal.

Walshe, K., Smith, J., Dixon, J., Edwards, N., Hunter, D.J., Mays, N., Normand, C., Robinson, R., 2004. Primary care trusts. BMJ 329, 871–2. doi:10.1136/bmj.329.7471.871

Whalley, D., Bojke, C., Gravelle, H., Sibbald, B., 2006a. GP job satisfaction in view of contract reform: a national survey. Br. J. Gen. Pract. 56, 87–92.

Whalley, D., Bojke, C., Gravelle, H., Sibbald, B., 2005. 2004 National Survey of General Practitioner Job Satisfaction. National Primary Care Research and Development Centre.

Whalley, D., Gravelle, H., Sibbald, B., 2008. Effect of the new contract on GPs' working lives and perceptions of quality of care: a longitudinal survey. Br. J. Gen. Pract. 58, 8–14. doi:10.3399/bjgp08X263758

Whalley, D., Gravelle, H., Sibbald, B., 2006b. 2005 National Survey of General Practitioner Job Satisfaction. National Primary Care Research and Development Centre.

Williams, E.S., Konrad, T.R., Scheckler, W.E., Pathman, D.E., Linzer, M., McMurray, J.E., Gerrity, M., Schwartz, M., 2001. Understanding physicians: Intentions to withdraw from practice: The role of job satisfaction, job stress, mental and physical health, in: Advances in Health Care Management, Advances in Health Care Management. Emerald Group Publishing Limited, pp. 243–262. doi:doi:10.1016/S1474-8231(01)02029-8

Witter, S., Fretheim, A., Kessy, F.L., Lindahl, A.K., 2012. Paying for performance to improve the delivery of health interventions in low- and middle-income countries . Cochrane database Syst. Rev. 2, CD007899. doi:10.1002/14651858.CD007899.pub2

Wooldridge, J., 2010. Econometric analysis of cross section and panel data, 2nd ed. The MIT Press.

Wooldridge, J., 2009. Introductory Econometrics. A Modern Approach, 4th ed. South-Western Cengage Learning.

Yang, M., Lien, H.-Mi., Chou, S.-Y., 2014. Is there a physician peer effect? Evidence from new drug prescriptions. Econ. Inq. 52, 116–137. doi:10.1111/ecin.12022

Young, G.J., Beckman, H., Baker, E., 2012. Financial incentives, professional values and performance: A study of pay-for-performance in a professional organization. J. Organ. Behav. 33, 964–983. doi:10.1002/job.1770

Young, G.J., Meterko, M., Beckman, H., Baker, E., White, B., Sautter, K.M., Greene, R., Curtin, K., Bokhour, B.G., Berlowitz, D., Burgess, J.F., 2007. Effects of paying physicians based on their relative performance for quality. J. Gen. Intern. Med. 22, 872–6. doi:10.1007/s11606-007-0185-5

Young, R., Leese, B., 1999. Recruitment and retention of general practitioners in the UK: what are the problems and solutions? Br. J. Gen. Pract.

TABLES

Table 3.1: Description of key variables from the QOF

Variable name	*Description*	*Source*
Points	The amount of points available for each QOF indicator. Ranged from 1 to 57 and varied by indicator, disease and time but not by practices.	Freely available online (Health & Social Care Information Centre, 2015a)
Value of a point	The monetary value of a point. Ranged from £75 to 133.76. Varied only by year.	Detailed in the GMS contract documents (BMA, 2011; National Audit Office, 2008; NHS Employers and BMA, 2011; NHS, 2010)
Thresholds	The upper and lower thresholds for performance. Points were not earned for patients treated below the lower thresholds. Additional points were not earned above the upper threshold. Varied by indicator, disease and time but not by practices.	Detailed in the GMS contract documents for each year that changes were made (BMA and NHS Employers, 2012, 2011, 2009, 2008, 2006; Department of Health, 2003a)
Prevalence rates	The number of practice patients with a disease divided by the practice list size. Value was specific to each practice and each disease.	Freely available online (Health & Social Care Information Centre, 2015a)
Treated patients	The number of patients in a practice who received the appropriate treatment on a given indicator.	Freely available online (Health & Social Care Information Centre, 2015a)
Eligible patients	The number of patients in a practice who were eligible for treatment on a given indicator. This is the number of practice practices with the disease with Exempted and excluded patients removed. Exempted if they did not meet the requirements of the indicator (e.g. age) and excluded if they did not adhere to treatment or if there were conflicts (e.g. comorbidities)	Freely available online (Health & Social Care Information Centre, 2015a)
QOF performance	Treated patients as a percentage of eligible patients.	Derived from the online data
List size	Number of patients registered in each practice	Freely available online (Health & Social Care Information Centre, 2015a)
Contractor Population Index	List size divided by average list size in 2003. An index of practice size used to adjust payments	Derived from the online data
Adjusted disease prevalence factor	Prevalence divided by average prevalence. Additional adjustments made in some years, see full text for details. An index of prevalence used to adjust payments	Derived from the online data
Revenue per indicator	Measures the revenue available for each indicator for each practice. Consists of value of a point*CPI*ADPF*points.	Derived from the online data

Table 3.2: Equations for key variables relating to the QOF

Element	*Equation*
ADPF 2004/5 to 2008/9	$\sqrt{\text{trunc}(p_{idt})} \Big/ \overline{\sqrt{\text{trunc}(p_{dt})}}$
ADPF 2009/10	$\text{trunc}(p_{idt}) \Big/ \overline{\text{trunc}(p_{dt})}$
ADPF 2010/11 to 2012/13	$ADPF_{idt} = {}^{p_{idt}}\Big/_{\overline{p_{idt}}}$
Truncation of prevalence	$trunc(p_{idt}) = \min(p_{idt},\ 0.05\ (max(p_{dt}) - \min(p_{dt}))$
Contractor Population Index	$CPI_{it} = l_{it}/\mu$
Revenue per clinical indicator	$Revenue_{kidt} = \pi_{kidt} * ADPF_{idt} * CPI_{it} * \alpha_t$
Practice performance measure	$Y_{kidt} = {}^{\tau_{kidt}}\Big/_{(r_{idt} - [E'_{kidt} - E''_{kidt}])}$

Note: these equations are explained in Section 3.2

Table 4.1: Descriptive statistics for practice characteristics

	2004-05	*2005-06*	*2006-07*	*2007-08*	*2008-09*	*2009-10*	*2010-11*	*2011-12*	*2012-13*
Performance (N/D)%	83.798	88.961	89.354	89.629	89.789	88.858	89.103	88.104	87.905
Revenue per patient	0.405	0.361	0.737	0.662	0.669	0.727	0.555	0.837	0.836
Points per patient	0.613	0.531	1.182	0.978	0.930	1.143	0.989	1.233	1.175
Revenue per indicator	602.447	1015.080	1149.575	1204.992	1181.752	1204.430	1228.022	1184.423	1133.933
Adjusted disease prevalence factor	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Contractor Population Index	1.057	1.075	1.088	1.105	1.120	1.121	1.136	1.160	1.186
Value per point (£)	75.000	124.600	124.600	124.600	124.600	126.770	126.770	130.510	133.760
Points available	7.569	7.569	8.717	8.717	8.467	8.530	8.530	7.866	7.208
Lower Threshold (%)	25.000	25.000	39.750	39.750	39.500	39.697	39.697	39.701	44.653
Upper Threshold (%)	81.385	81.385	82.083	82.083	82.083	81.288	81.288	80.119	80.708
Disease Prevalence	3.161	3.301	3.235	3.399	3.539	3.718	3.813	3.634	3.248
list size (000's)	6.226	6.330	6.412	6.512	6.600	6.603	6.691	6.836	6.984
Number of GPs in practice	4.046	4.190	4.287	4.288	4.592	4.804	4.904	5.028	5.156
Patients per GP (000's)	1.802	1.771	1.758	1.756	1.674	1.621	1.585	1.579	1.561
Index of Multiple Deprivation score (04, 07 & 10)	25.959	25.984	26.431	26.336	26.288	26.545	26.594	26.548	26.451
Average age of GPs in practice	47.925	48.157	48.311	48.418	48.230				
Proportion of UK qualified GPs in practice	0.691	0.685	0.667	0.671	0.672				
Prop of GP partners	0.888	0.883	0.836	0.824	0.785				
Proportion of female GPs in practice	0.352	0.360	0.368	0.376	0.389				
Prop male patients aged 0 to 4	0.028	0.028	0.029	0.029	0.030				
Prop female patients aged 0 to 4	0.027	0.027	0.027	0.028	0.028				
Prop male patients aged 5 to 14	0.063	0.062	0.061	0.060	0.059				
Prop female patients aged 5 to 14	0.060	0.059	0.058	0.057	0.056				
Prop male patients aged 15 to 44	0.224	0.224	0.224	0.222	0.219				
Prop female patients aged 15 to 44	0.209	0.209	0.209	0.209	0.208				
Prop male patients aged 45 to 64	0.123	0.124	0.125	0.127	0.128				
Prop female patients aged 45 to 64	0.115	0.116	0.117	0.119	0.120				
Prop male patients aged 65 to 74	0.038	0.038	0.038	0.038	0.039				
Prop female patients aged 65 to 74	0.041	0.041	0.040	0.040	0.041				
Prop male patients aged 75 to 84	0.022	0.022	0.022	0.022	0.022				
Prop female patients aged 75 to 84	0.032	0.031	0.030	0.030	0.030				
Prop male patients aged 85+	0.005	0.005	0.006	0.006	0.006				
Prop female patients aged 85+	0.012	0.013	0.013	0.014	0.014				

Note: practice characteristics were not available for all years

Table 4.2: Detailed descriptive statistics for the ADPF

Adjusted disease prevalence factor	*Mean*	*sd*	*p5*	*p95*
2004-05	1.000	0.180	0.715	1.298
2005-06	1.000	0.167	0.762	1.277
2006-07	1.000	0.162	0.745	1.268
2007-08	1.000	0.184	0.724	1.304
2008-09	1.000	0.134	0.858	1.232
2009-10	1.000	0.397	0.605	1.533
2010-11	1.000	0.551	0.371	1.683
2011-12	1.000	0.605	0.375	1.696
2012-13	1.000	0.577	0.356	1.733

Note: for 2009/10 the square rooting was removed followed by the removal of the truncation in 2010/11.

Table 4.3: Detailed descriptive statistics for reputation and revenue payoff variables

	Mean	*sd*	*p5*	*p95*
Points per patient				
2004-05	0.613	1.714	0.017	2.469
2005-06	0.531	1.445	0.016	2.222
2006-07	1.182	3.738	0.021	5
2007-08	0.978	3.286	0.02	4
2008-09	0.93	2.737	0.019	4
2009-10	1.143	3.689	0.018	4.762
2010-11	0.989	3.187	0.018	4
2011-12	1.233	3.251	0.02	6
2012-13	1.175	3.028	0.02	5.357
Revenue per patient				
2004-05	0.405	1.142	0.023	1.570
2005-06	0.361	0.922	0.021	1.381
2006-07	0.737	2.329	0.027	2.891
2007-08	0.662	1.960	0.026	2.460
2008-09	0.669	1.809	0.024	2.538
2009-10	0.727	2.134	0.023	2.789
2010-11	0.555	1.457	0.023	1.811
2011-12	0.837	2.230	0.023	3.733
2012-13	0.836	2.396	0.025	3.176

Table 4.4: Correlations between payoff variables by year

	Points Payoff	
Revenue payoff	2004-05	0.711
	2005-06	0.754
	2006-07	0.764
	2007-08	0.743
	2008-09	0.674
	2009-10	0.607
	2010-11	0.526
	2011-12	0.627
	2012-13	0.560

Table 4.5: Fixed effects regressions of the effect of reputational and revenue incentives on practice performance

	Performance (N/D) %			
	Years 1-5	*Years 6-9*	*Years 1-9*	*Years 1-9 (unbalanced)*
Points per patient	0.205***	0.198***	0.183***	0.207***
	(6.04)	(8.04)	(10.13)	(18.14)
Revenue per patient	0.279***	0.245***	0.282***	0.241***
	(6.90)	(8.87)	(12.79)	(15.07)
Upper Threshold (%)	0.0264***	0.450***	0.135***	0.158***
	(5.52)	(104.22)	(33.60)	(42.33)
List size (000's)	-0.763***	-0.102***	-0.359***	-0.337***
	(-27.77)	(-5.26)	(-29.69)	(-32.13)
Patients per GP (000's)	-0.395***	-0.130***	-0.368***	-0.325***
	(-14.47)	(-5.78)	(-20.23)	(-21.23)
2005-06	5.908***		5.893***	4.930***
	(230.45)		(229.24)	(256.63)
2006-07	7.745***		7.481***	6.610***
	(245.95)		(239.94)	(254.63)
2007-08	8.140***		7.804***	6.690***
	(247.66)		(243.15)	(252.62)
2008-09	8.047***		7.684***	6.795***
	(235.76)		(234.53)	(253.02)
2009-10		0.420***	7.414***	6.635***
		(21.68)	(221.36)	(240.41)
2010-11		0.654***	7.664***	6.939***
		(36.48)	(227.23)	(250.07)
2011-12		0.792***	7.919***	7.131***
		(49.85)	(227.13)	(241.98)
2012-13			7.426***	6.877***
			(201.57)	(217.18)
Constant	83.81***	52.27***	72.72***	72.21***
	(200.82)	(135.87)	(222.19)	(236.53)
Observations	1699188	1356804	3055992	4670094
Groups	358415	350063	369290	822290
R^2 Within	0.120	0.0158	0.0868	0.0670
rho	0.634	0.603	0.574	0.644

Robust *t* statistics in parentheses (Std. Err. adjusted for practice-indicator clusters)
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Note: difference between payoffs was statistically significant in model 3 (F prop=0.2985, 0.3447, 0.0097, 0.1966)
Models presented are variations of Equation 4.3

Table 4.6: Fixed effect regressions of the year specific effect of reputational and revenue incentives on practice performance

	Performance (N/D)%			
	Years 1-9		*Years 1-9 (unbalanced)*	
Points per patient * year dummies				
2004-05	-0.0778	(-1.47)	-0.200***	(-4.21)
2005-06	0.0923	(1.85)	0.106*	(2.28)
2006-07	0.0870*	(2.51)	0.114***	(4.97)
2007-08	0.196***	(5.39)	0.145***	(7.02)
2008-09	0.210***	(5.69)	0.218***	(9.95)
2009-10	0.231***	(6.04)	0.249***	(13.08)
2010-11	0.262***	(6.13)	0.287***	(16.54)
2011-12	0.251***	(7.84)	0.304***	(17.42)
2012-13	0.217***	(6.92)	0.287***	(15.58)
Revenue per patient * year dummies				
2004-05	0.662***	(8.53)	0.610***	(9.18)
2005-06	0.638***	(8.22)	0.633***	(8.69)
2006-07	0.391***	(10.03)	0.235***	(6.74)
2007-08	0.405***	(8.27)	0.173***	(5.94)
2008-09	0.319***	(6.94)	0.239***	(7.79)
2009-10	0.209***	(4.47)	0.262***	(10.23)
2010-11	0.0967	(1.23)	0.290***	(11.70)
2011-12	0.160***	(5.24)	0.275***	(14.70)
2012-13	0.112***	(3.56)	0.272***	(12.96)
Year dummies				
2005-06	5.800***	(190.18)	4.754***	(207.59)
2006-07	7.481***	(220.92)	6.618***	(235.26)
2007-08	7.721***	(225.48)	6.705***	(235.56)
2008-09	7.634***	(217.37)	6.699***	(234.86)
2009-10	7.405***	(201.06)	6.500***	(222.64)
2010-11	7.671***	(208.56)	6.762***	(227.78)
2011-12	7.919***	(211.28)	6.928***	(223.22)
2012-13	7.462***	(190.19)	6.707***	(203.13)
Upper Threshold (%)	0.142***	(33.54)	0.159***	(40.87)
list size (000's)	-0.348***	(-28.65)	-0.324***	(-30.90)
Patients per GP (000's)	-0.366***	(-20.12)	-0.322***	(-21.02)
Constant	72.13***	(208.10)	72.10***	(224.70)
Observations	3055992		4670094	
Groups	369290		822290	
R^2 Within	0.0872		0.0678	
rho	0.572		0.644	

Robust *t* statistics in parentheses (Std. Err. adjusted for practice-indicator clusters)
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Note: models presented are variations of Equation 4.4

Table 4.7: Fixed effects regressions comparing the impact of independent variables on the effect of reputational and revenue incentives on practice performance

	Performance (N/D) %						
	Points only	*Revenue only*	*Points + revenue*	*+ upper threshold*	*+ controls (1-9)*	*+ controls (1-5)*	*+ balanced set of indicators*
Points per patient	0.316***		0.192***	0.193***	0.179***	0.182***	0.230***
	(25.57)		(8.60)	(8.63)	(7.67)	(7.75)	(6.73)
Revenue per patient		0.468***	0.237***	0.243***	0.260***	0.256***	0.283***
		(27.35)	(8.01)	(8.18)	(8.44)	(8.25)	(6.91)
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Upper threshold				Yes	Yes	Yes	Yes
Year 1-9 control variables					Yes	Yes	Yes
Year 1-5 control variables						Yes	Yes
Observations	2531666	2531666	2531666	2531666	2496234	2483030	2016920
Groups	640456	640456	640456	640456	635130	631117	424100
R^2 Within	0.0968	0.0967	0.0970	0.0971	0.0990	0.101	0.114
rho	0.649	0.648	0.649	0.639	0.657	0.684	0.674

Robust *t* statistics in parentheses (Std. Err. adjusted for practice-indicator clusters)
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Note: models presented are variations of Equation 4.3

Table 4.8: Fixed effects regressions comparing the specification of fixed effects on the effect of reputational and revenue incentives on practice performance

	Performance (N/D) %		
	i(pratice)	*i(practice-disease)*	*i(practice-indicator)*
Points per patient	0.218***	0.175***	0.183***
	(9.55)	(7.39)	(10.13)
Revenue per patient	0.292***	0.349***	0.282***
	(7.63)	(10.09)	(12.79)
Year dummies	Yes	Yes	Yes
Upper threshold	Yes	Yes	Yes
Control variables	Yes	Yes	Yes
Disease dummies	Yes		
Indicator dummies	Yes	Yes	
Observations	3055992	3055992	3055992
Groups	8819	96891	369290
R^2 Within	0.372	0.369	0.0868
rho	0.226	0.580	0.574

Robust *t* statistics in parentheses (Std. Err. adjusted for different clusters)
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Note: models presented are variations of Equation 4.3

Table 4.9: Fixed effects dummy variable model regressions comparing the specification of fixed effects on the effect of reputational and revenue incentives on practice performance

	Performance (N/D) %		
	i(practice)	*i(practice-disease)*	*i(practice-indicator)*
Points per patient	0.218***	0.186***	0.193***
	(59.39)	(47.53)	(40.30)
Revenue per patient	0.296***	0.344***	0.277***
	(47.99)	(53.64)	(40.14)
Year dummies	Yes	Yes	Yes
Upper threshold	Yes	Yes	Yes
Control variables	Yes	Yes	Yes
Disease dummies	Yes		
Indicator dummies	Yes	Yes	
Observations	3055992	3055992	3055992
R^2 (adjusted)	0.433	0.478	0.547

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Note: models presented are variations of Equation 4.3

Table 4.10: Comparison of fixed and random effects regressions of the effect of reputational and revenue incentives on practice performance

	Performance (N/D) %	
	FE	*RE*
Points per patient	0.183***	0.132***
	(33.80)	(28.08)
Revenue per patient	0.282***	0.384***
	(39.20)	(57.75)
Upper Threshold (%)	0.135***	0.455***
	(62.03)	(399.94)
list size (000's)	-0.359***	0.0104**
	(-40.92)	(3.12)
Patients per GP (000's)	-0.368***	-0.530***
	(-29.74)	(-48.39)
2005-06	5.893***	5.936***
	(275.56)	(278.17)
2006-07	7.481***	6.730***
	(333.41)	(307.35)
2007-08	7.804***	7.047***
	(350.99)	(325.52)
2008-09	7.684***	6.904***
	(343.28)	(317.58)
2009-10	7.414***	6.610***
	(329.54)	(303.50)
2010-11	7.664***	6.852***
	(339.32)	(314.18)
2011-12	7.919***	6.967***
	(344.64)	(317.21)
2012-13	7.426***	6.153***
	(311.71)	(276.56)
Constant	72.72***	45.27***
	(399.57)	(469.93)
Observations	3055992	3055992
Groups	369290	369290
R^2 Within	0.0868	0.0801
rho	0.574	0.458

t statistics in parentheses (Std. Err. adjusted for practice-indicator clusters)
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Chi squared statistic for Hausman test = 38802.4
Note: models presented are variations of Equation 4.3

Table 5.1: Comparison of practices dropped from peer effects analysis due to attrition

	Dropped	*Retained*
Total QOF points	863.101	973.652
List size ('000s)	3.063	6.530
List size per GP ('000s)	1.717	1.781
Average age (years) of GPs	52.706	48.008
Proportion of UK qualified GPs	0.535	0.685
Proportion of GPs partners	0.897	0.841
Proportion of male GPs	0.688	0.631
Number of FTE GPs	1.889	3.956
PMS practice	0.387	0.398
Dispensing practice	0.454	0.155
Rural/urban practice, ONS 07	0.071	0.154
Low Income Scheme Index	18.422	12.336

Note: practices are dropped if they do not appear in all years

Table 5.2: Descriptive statistics for practices in peer effects sample

	2004/5	*2005/6*	*2006/7*	*2007/8*	*2008/9*
Number of practices	8041	8041	8041	8041	8041
Number of PCTs	303	303	152	152	152
Number of practices per PCT	33	33	63	63	63
Total QOF points	969.645	1014.220	958.312	970.538	955.604
Total QOF points %	92.347	96.592	91.268	97.054	95.560
List size ('000s)	6.412	6.455	6.524	6.594	6.669
List size per GP ('000s)	1.831	1.810	1.794	1.754	1.715
Average age (years) of GPs	47.553	47.872	48.077	48.329	48.208
Proportion of UK qualified GPs	0.704	0.696	0.675	0.676	0.675
Proportion of GPs partners	0.885	0.879	0.833	0.823	0.786
Proportion of male GPs	0.648	0.640	0.632	0.622	0.614
Number of FTE GPs	3.709	3.818	4.069	4.011	4.177
PMS practice	0.358	0.340	0.432	0.432	0.432
Dispensing practice	0.155	0.155	0.155	0.155	0.155
Rural/urban practice, ONS 07	0.154	0.154	0.154	0.154	0.154
Lower Income Scheme Index	12.556	12.625	12.575	12.442	11.474
Proportion male patients aged 0 to 4	0.028	0.028	0.029	0.029	0.030
Proportion female patients aged 0 to 4	0.027	0.027	0.027	0.028	0.028
Proportion male patients aged 5 to 14	0.063	0.062	0.061	0.060	0.059
Proportion female patients aged 5 to 14	0.060	0.059	0.058	0.057	0.056
Proportion male patients aged 15 to 44	0.223	0.223	0.222	0.221	0.219
Proportion female patients aged 15 to 44	0.209	0.209	0.209	0.208	0.207
Proportion male patients aged 45 to 64	0.123	0.124	0.126	0.127	0.128
Proportion female patients aged 45 to 64	0.115	0.117	0.118	0.119	0.121
Proportion male patients aged 65 to 74	0.039	0.038	0.038	0.038	0.039
Proportion female patients aged 65 to 74	0.041	0.041	0.040	0.040	0.041
Proportion male patients aged 75 to 84	0.022	0.022	0.022	0.022	0.022
Proportion female patients aged 75 to 84	0.032	0.031	0.031	0.030	0.030
Proportion male patients aged 85+	0.005	0.005	0.006	0.006	0.006
Proportion female patients aged 85+	0.013	0.013	0.013	0.014	0.014

Table 5.3: Fixed effects regressions of peer effects for merged and non-merged practices

	Total QOF points %			
	Merged		*Non-merged*	
Old peer group	0.723***	(18.18)	0.701***	(14.30)
Old peer group * post-period	-0.102**	(-3.02)	0.000777	(0.02)
2005/6	1.090***	(6.95)	1.313***	(5.71)
2006/7	9.223**	(2.94)	-0.345	(-0.08)
2007/8	11.43***	(3.48)	1.489	(0.32)
2008/9	10.96***	(3.36)	1.114	(0.24)
Number of practices per PCT	-0.000303	(-0.14)	-0.0711	(-1.26)
List size ('000s)	-0.480***	(-3.97)	-0.294*	(-2.25)
List size per GP ('000s)	-0.337	(-1.85)	-0.729***	(-3.72)
Total GPs	-0.0880	(-1.33)	-0.275**	(-3.13)
Average age (years) of GPs	-0.0970***	(-4.80)	-0.128***	(-6.32)
Proportion of UK qualified GPs in practice	-0.484	(-1.07)	-0.420	(-0.99)
Proportion of GP partners	-0.158	(-0.40)	0.742	(1.73)
Proportion of male GPs	0.510	(1.21)	-0.741	(-1.41)
Number of FTE GPs in practice	-0.128*	(-2.56)	-0.140*	(-2.31)
PMS practice	-0.283*	(-1.96)	-0.568*	(-2.28)
Lower Income Scheme Index	-0.00963	(-0.53)	0.0226	(1.07)
Prop male patients aged 0 to 4	38.87	(0.96)	-1.025	(-0.02)
Prop female patients aged 0 to 4	47.10	(1.16)	-19.49	(-0.39)
Prop male patients aged 5 to 14	53.96	(1.38)	7.925	(0.17)
Prop female patients aged 5 to 14	52.15	(1.38)	-7.250	(-0.15)
Prop male patients aged 15 to 44	53.00	(1.52)	17.51	(0.41)
Prop female patients aged 15 to 44	56.79	(1.61)	36.21	(0.81)
Prop male patients aged 45 to 64	60.45	(1.67)	14.95	(0.33)
Prop female patients aged 45 to 64	31.44	(0.80)	15.90	(0.36)
Prop male patients aged 65 to 74	52.37	(1.22)	32.93	(0.56)
Prop female patients aged 65 to 74	46.30	(1.09)	-8.768	(-0.17)
Prop male patients aged 75 to 84	51.81	(1.06)	-5.508	(-0.08)
Prop female patients aged 75 to 84	8.698	(0.23)	4.524	(0.09)
Prop male patients aged 85+	34.48	(0.51)	6.003	(0.07)
Constant	-13.07	(-0.38)	25.55	(0.59)
Observations	23736		15231	
R^2	0.362		0.315	

Robust *t* statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Note: omitted year dummy for 2004/5 and Prop female patients aged 85+
Models presented are of Equation 5.3

Table 5.4: Fixed effects regression of peer effects on a combined sample of merged and non-merged practices

	Total QOF points %
Old peer group	0.709***
	(14.58)
Merged * Old peer group	0.0195
	(0.31)
Old peer group * post-period	-0.00333
	(-0.07)
Merged * Old peer group * post-period	-0.111
	(-1.87)
Observations	38979
R^2	0.339

Robust *t* statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Note: control variables included in all models and interacted with merged
Models presented are variations of Equation 5.3

Table 5.5: Fixed effects regression of peer effects for old and new peers

	Total QOF points %
Old peer group	0.684***
	(15.95)
New peers only	0.130***
	(3.67)
Old peer group * post-period	-0.100**
	(-2.75)
New peers * post-period	0.0354
	(0.99)
Observations	23736
R^2	0.363

Robust *t* statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Note: control variables included in all models
Models presented are variations of Equation 5.4

Table 5.6: OLS regressions of peer effects with lagged peer performance

	Total QOF points %			
	t=2005/6	*t=2006/7*	*t=2007/8*	*t=2008/9*
Own performance t-1	-7.380	-15.60	15.68*	-13.15*
	(-1.55)	(-1.58)	(2.26)	(-1.98)
Own performance t-1^2	0.157	0.326	-0.299	0.311*
	(1.54)	(1.49)	(-1.90)	(1.96)
Own performance t-1^3	-0.00135	-0.00290	0.00257	-0.00303*
	(-1.45)	(-1.43)	(1.69)	(-1.96)
Own performance t-1^4	0.00000419	0.00000964	-0.00000815	0.0000107*
	(1.35)	(1.41)	(-1.53)	(2.00)
Old peer performance t-1	-0.0595*	0.0127	0.0310	0.0468
	(-2.17)	(0.33)	(1.05)	(1.36)
New peer performance t-1	0.00833	-0.0220	0.0996***	0.00422
	(0.54)	(-0.63)	(3.35)	(0.13)
Observations	4767	4759	4746	4719
R^2	0.557	0.542	0.512	0.427

Robust *t* statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Note: control variables included in all models
Models presented are of Equation 5.5

Table 5.7: Fixed effects regressions of peer effects for old and new peers stratified by PCT management quality

	Total QOF points %	
	Bottom 50%	*Top 50%*
Old peer group	0.526***	0.736***
	(9.67)	(12.46)
New peers	0.181**	0.106*
	(3.21)	(2.21)
Old peer group * post-period	-0.0671	-0.0899
	(-1.34)	(-1.71)
New peers * post-period	0.0679	0.00300
	(1.21)	(0.06)
Observations	11340	10543
R^2	0.362	0.376

Robust *t* statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Note: control variables included in all models
Difference between interacted coefficients is not statistically significant (t-ratio 0.34 and 0.81)
Models presented are variations of Equation 5.4

Table 5.8: Fixed effects regressions of peer effects for old and new peers stratified by PCT size

	Total QOF points %	
	Bottom 50%	*Top 50%*
Old peer group	0.746***	0.588***
	(13.29)	(9.03)
New peers	0.116**	0.177**
	(2.64)	(3.14)
Old peer group * post-period	-0.114*	-0.101
	(-2.34)	(-1.87)
New peers * post-period	0.0750	0.00957
	(1.18)	(0.22)
Observations	11880	11856
R^2	0.399	0.340

Robust *t* statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Note: control variables included in all models
PCT size measures the size of a practices old PCT as a percentage of the new PCT
Difference between interacted coefficients is not statistically significant (t-ratio -0.04 and 0.89)
Models presented are variations of Equation 5.4

Table 5.9: OLS regressions of peer effects with lagged peer performance stratified by PCT management quality

	Total QOF points %							
	t=2005/6		*t=2006/7*		*t=2007/8*		*t=2008/9*	
	Bottom 50%	*Top 50%*	*Bottom 50%*	*Top 50%*	*Bottom 50%*	*Top 50%*	*Bottom 50%*	*Top 50%*
Old peer performance t-1	-0.0117	-0.0767*	-0.00894	0.0321	0.0648	0.0278	-0.0200	0.0996
	(-0.41)	(-2.16)	(-0.17)	(0.62)	(1.62)	(0.74)	(-0.41)	(1.93)
New peer performance t-1	0.0465	0.000229	0.0673	-0.0689	0.128**	0.0517	0.104*	0.0174
	(1.83)	(0.01)	(1.12)	(-1.64)	(2.83)	(1.64)	(2.06)	(0.39)
Observations	2276	2118	2272	2115	2267	2107	2263	2092
R^2	0.590	0.569	0.579	0.538	0.586	0.511	0.439	0.508

Robust *t* statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Note: control variables and lag of own performance included in all models. Own performance is modelled with a quartic specification.
Statistically significant difference found between top and bottom 50% for old peer performance in 2008/9 model (t-ratio -2.11)
Models presented are variations of Equation 5.5

Table 5.10: OLS regressions of peer effects with lagged peer performance stratified by PCT size

	Total QOF points %							
	t=2005/6		*t=2006/7*		*t=2007/8*		*t=2008/9*	
	Bottom 50%	*Top 50%*	*Bottom 50%*	*Top 50%*	*Bottom 50%*	*Top 50%*	*Bottom 50%*	*Top 50%*
Old peer performance t-1	-0.0762*	-0.0145	0.0567	-0.0578	-0.0237	0.105*	0.101*	0.0266
	(-2.07)	(-0.48)	(0.88)	(-1.20)	(-0.58)	(2.52)	(2.06)	(0.49)
New peer performance t-1	-0.00844	0.00857	-0.00839	-0.0147	0.173**	0.0641	-0.0806	0.0377
	(-0.40)	(0.35)	(-0.13)	(-0.38)	(3.24)	(1.85)	(-1.23)	(0.91)
Observations	2503	2386	2504	2378	2500	2368	2490	2351
R^2	0.478	0.588	0.480	0.554	0.509	0.507	0.391	0.416

Robust *t* statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Note: control variables and lag of own performance included in all models. Own performance is modelled with a quartic specification
Statistically significant difference found between top and bottom 50% for old peer performance in 2007/8 model (t-ratio -3.86) and between top and bottom 50% for new peer performance 2008/9 model (t-ratio -2.67)
Models presented are variations of Equation 5.5

Table 5.11: Fixed effects regressions of peer effects for old and new peers using alternative peer performance measures

	Total QOF points %		
	Low performers	*High performers*	*2nd lowest score*
Old peer group	-0.149***	0.00535	0.320***
	(-15.37)	(1.39)	(15.29)
New peers only	-0.000174	0.0107***	0.0188
	(-0.16)	(4.31)	(1.24)
Old peer group * post-period	0.0481***	-0.0119	-0.0207
	(4.06)	(-1.76)	(-1.14)
New peers * post-period	-0.000633	-0.00760	-0.0472**
	(-0.30)	(-1.71)	(-2.71)
Observations	23736	23736	23729
R^2	0.347	0.305	0.363

Robust *t* statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Note: control variables included in all model
Low performers measures the percentage of practices within the PCT who scored below the 5th percentile of national performance.
High performers measures the percentage of practices within the PCT who score 100%
2nd lowest score measures the 2nd lowest score in the PCT. For practices that scored this value, the 3rd lowest is used.
Models presented are variations of Equation 5.4

Table 6.1: Descriptive statistics for GP WLS

	2004	*2005*	*2008*
Life satisfaction (1 to 7)	4.649	5.095	5.008
Overall satisfaction with job (1 to 7)	4.567	5.201	4.728
Satisfaction with physical working conditions (1 to 7)	4.862	5.044	5.129
Satisfaction with freedom to choose own method of working (1 to 7)	4.636	4.892	4.640
Satisfaction with colleagues & fellow workers (1 to 7)	5.515	5.599	5.602
Satisfaction with recognition you get for good work (1 to 7)	4.224	4.726	4.495
Satisfaction with amount of responsibility you are given (1 to 7)	4.976	5.406	5.276
Satisfaction with remuneration (1 to 7)	4.376	5.387	4.849
Satisfaction with opportunity to use abilities (1 to 7)	4.787	5.147	5.074
Satisfaction with hours of work (1 to 7)	3.914	4.802	4.205
Satisfaction with amount of variety in job (1 to 7)	5.011	5.269	5.276
High likelihood of quitting	0.256	0.218	0.251
Hours per week typically work as a GP	44.540	40.509	42.738
Male	0.662	0.636	0.633
Married/living vs no spouse	0.920	0.914	0.911
Number of children under 18	1.418	1.284	1.302
Non-white GP	0.155	0.121	0.123
Age (years)	47.034	47.977	48.777
No partner/partner not working	0.261	0.304	0.203
Partner working part-time	0.360	0.336	0.417
Partner working full-time	0.379	0.360	0.381
Income <25K	0.009	0.003	
25-50K	0.164	0.099	
50-70K	0.261	0.140	
70-85K	0.266	0.166	
85-100K	0.178	0.210	
100-120K	0.086	0.232	
120-150K	0.027	0.114	
150+K	0.008	0.037	
<25K			0.004
25-50K			0.049
50-75K			0.154
75-100K			0.284
100-125K			0.310
125-150K			0.127
150-175K			0.042
175+K			0.031
Total number of GP partners	4.776	4.962	4.610
Practice list size	8975.922	9091.339	9357.244
Patients per GP/1000*	1.850	1.831	1.591
Time in current practice (years)	14.561	15.417	16.298
PMS practice	0.456	0.451	0.447
Rural practice*	0.183	0.187	0.188
Low income scheme index*	10.433	10.618	9.599
Non-white population*	0.122	0.116	0.113
Dispensing practice*	0.194	0.189	0.209

* Variables not from the GP WLS

Notes: statistics shown are for the full GP WLS sample

Missing values for income are due to changes made to income bands

Table 6.2: Maximum achievable QOF income for practices with GPs who responded to the GP WLS

	Practices	*Mean (£)*	*sd*	*p10*	*p90*
2005 QOF income	1736	116612	56493	49487	183062
2008 QOF income	1811	192919	90305	83060	300750

Notes: figures are for gross income calculated using Equation 3.8

Table 6.3: Interval regressions for the determinants of income for GP WLS respondents

	2004		*2005*		*2008*	
Male	20891.1***	(19.29)	26685.4***	(20.75)	26117.2***	(19.42)
Age (years)	796.4	(1.11)	3043.1***	(3.61)	3018.1**	(2.97)
Age squared	-7.166	(-0.95)	-30.55***	(-3.45)	-30.50**	(-2.82)
Patients per GP/1000	8749.8***	(7.00)	12841.7***	(5.86)	17835.7***	(8.73)
Partnership size: 2	-6100.2	(-1.89)	1031.2	(0.27)	-7609.4	(-1.57)
Partnership size: 3	-2949.1	(-0.92)	1384.6	(0.38)	-5705.2	(-1.26)
Partnership size: 4	-684.8	(-0.21)	1742.3	(0.49)	-5624.8	(-1.27)
Partnership size: 5	-1420.5	(-0.44)	2731.9	(0.77)	-7507.2	(-1.70)
Partnership size: 6	-1107.9	(-0.34)	2196.6	(0.60)	-6075.0	(-1.35)
Partnership size: 7	738.2	(0.22)	3940.0	(1.02)	-7479.5	(-1.55)
Partnership size: 8	441.8	(0.11)	6632.0	(1.56)	-9101.0	(-1.70)
Partnership size: 9+	6007.8	(1.41)	3255.9	(0.77)	-1504.1	(-0.28)
Dispensing practice	12623.0***	(7.60)	12901.6***	(6.81)	15705.3***	(7.37)
Non-white GP	-192.6	(-0.12)	-3581.6	(-1.83)	-7187.5**	(-3.15)
PMS practice	7414.7***	(6.95)	9403.8***	(7.33)	10180.6***	(7.42)
Non-white population	9596.6*	(2.35)	13174.7**	(2.98)	19003.6***	(3.40)
Low income scheme index	-255.1**	(-2.75)	-373.7***	(-3.51)	-291.9*	(-2.17)
Rural practice	3139.8*	(2.04)	1700.9	(0.94)	2415.8	(1.17)
Constant	19348.0	(1.14)	-29748.5	(-1.47)	-16746.9	(-0.70)
Observations	1867		1898		1990	

t statistics in parentheses (Standard errors clustered by practice)
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Notes: omitted category Partnership size=1
Models presented are of Equations 6.3 and 6.4

Table 6.4: GP incomes predicted from interval regressions and QOF exposure for GPs and practices

	Practices	*Mean*	*sd*	*p10*	*p90*
Predicted income 2004	1918	73827	13584	53670	90028
Predicted income 2005	1956	73584	13814	53390	90211
Predicted income 2008	2071	71360	13194	52198	87256
GP QOF income exposure 2005	1956	14.591	4.053	9.830	20.144
Practice QOF income exposure 2005	2065	10.784	0.641	10.012	11.522
GP QOF income exposure 2008	2069	25.644	8.936	16.764	35.770
Practice QOF income exposure 2008	2194	17.170	1.145	15.717	18.496

Notes: predicted incomes are based on determinants of income from before the QOF in 2004 (the first model in Table 6.3
Practice exposure is calculated using Equation 6.1 and GP exposure from Equation 6.2

Table 6.5: OLS regressions of the effect of GP P4P exposure on job satisfaction using cross-sections of the GP WLS

	Job satisfaction 2005		*Job satisfaction 2008*	
QOF income exposure (GP)	-0.0215*	(-2.00)	-0.00795	(-1.90)
50-70K	0.131	(1.09)		
70-85K	0.0783	(0.60)		
85-100K	0.408**	(3.19)		
100-120K	0.457***	(3.50)		
120-150K	0.667***	(4.56)		
150+K	0.697***	(3.40)		
50-75K			0.267	(1.77)
75-100K			0.200	(1.32)
100-125K			0.424**	(2.73)
125-150K			0.675***	(3.91)
150-175K			0.680**	(3.22)
175+K			0.941***	(3.77)
Male	-0.211**	(-2.61)	-0.319***	(-4.23)
Non-white GP	-0.225*	(-2.08)	-0.0186	(-0.18)
Age (years)	-0.167***	(-4.25)	-0.182***	(-4.37)
Age squared	0.00174***	(4.25)	0.00190***	(4.44)
Patients per GP/1000	-0.204*	(-2.38)	-0.274**	(-3.15)
Time in current practice (years)	0.00124	(0.21)	0.00728	(1.22)
PMS practice	0.0426	(0.70)	0.0369	(0.60)
Rural practice	0.0236	(0.32)	0.000879	(0.01)
Low income scheme index	-0.000111	(-0.02)	0.00149	(0.26)
Constant	9.580***	(10.18)	9.296***	(9.01)
Observations	1888		1970	
R^2	0.049		0.039	

t statistics in parentheses (Standard errors clustered by practice)
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Notes: omitted category Income <50K
Models presented are variations of Equation 6.5

Table 6.6: OLS regressions of the effect of GP P4P exposure on changes in job satisfaction using a longitudinal sample of the GP WLS

	Change in job satisfaction 2004-2005		*Change in job satisfaction 2004-2008*		*Change in job satisfaction 2004-2008 (Balanced panel)*	
QOF income exposure (GP)	0.0328*	(2.49)	0.000794	(0.11)	0.00183	(0.25)
Change in income (000s)	0.00747	(0.67)	-0.00495	(-0.57)	-0.00498	(-0.56)
Male	0.0978	(0.91)	0.0304	(0.23)	0.0641	(0.50)
Non-white GP	0.119	(0.79)	-0.218	(-0.99)	-0.256	(-1.21)
Age (years)	-0.0108	(-0.14)	0.123	(1.00)	0.0915	(0.74)
Age squared	0.0000337	(0.04)	-0.00118	(-0.88)	-0.000819	(-0.61)
Patients per GP/1000	-0.143	(-1.19)	0.237	(1.55)	0.194	(1.27)
Time in current practice (years)	0.0146	(1.71)	-0.00371	(-0.27)	-0.00218	(-0.16)
PMS practice	-0.00792	(-0.09)	-0.139	(-1.25)	-0.115	(-1.04)
Rural practice	0.0454	(0.42)	-0.105	(-0.66)	-0.0805	(-0.51)
Low income scheme index	-0.00817	(-1.36)	0.00658	(0.66)	0.00658	(0.65)
Constant	0.440	(0.26)	-3.241	(-1.14)	-2.557	(-0.90)
Observations	1201		743		729	
R^2	0.012		0.018		0.016	

t statistics in parentheses (Standard errors clustered by practice)
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Notes: omitted category Income <50K
Models presented are variations of Equation 6.6

Table 6.7: OLS regressions of the effect of GP P4P exposure on job satisfaction using a longitudinal sample of the GP WLS and the continuous DID approach

	Job satisfaction 2004 & 2005		*Job satisfaction 2004 & 2008*	
QOF income exposure * year	0.0177	(1.81)	-0.00368	(-0.55)
QOF income exposure (GP)	-0.0401***	(-3.48)	-0.00252	(-0.34)
Year (2005 or 2008)	0.250	(1.50)	-0.0311	(-0.13)
50-70K * 2005	-0.117	(-0.74)		
70-85K * 2005	-0.157	(-0.98)		
85-100K * 2005	-0.00910	(-0.06)		
100-120K * 2005	-0.0831	(-0.50)		
120-150K * 2005	0.140	(0.55)		
150+K * 2005	1.318*	(2.27)		
50-70K	0.259*	(2.04)	0.198	(1.26)
70-85K	0.264	(1.92)	0.288	(1.76)
85-100K	0.470**	(3.24)	0.378*	(2.10)
100-120K	0.602***	(3.82)	0.412*	(2.05)
120-150K	0.591*	(2.41)	0.519	(1.75)
150+K	-0.542	(-0.92)	-1.347	(-1.81)
50-75K			0.266	(1.76)
75-100K			0.206	(1.36)
100-125K			0.422**	(2.73)
125-150K			0.671***	(3.93)
150-175K			0.672**	(3.20)
175+K			0.920***	(3.68)
Male	-0.268***	(-3.64)	-0.290***	(-4.09)
Non-white GP	-0.214*	(-2.16)	-0.0387	(-0.38)
Age (years)	-0.167***	(-4.25)	-0.181***	(-4.79)
Age squared	0.00177***	(4.31)	0.00187***	(4.80)
Patients per GP/1000	-0.233**	(-3.01)	-0.358***	(-4.81)
Time in current practice (years)	-0.00308	(-0.56)	0.00641	(1.15)
PMS practice	0.0714	(1.25)	0.102	(1.77)
Rural practice	-0.0259	(-0.36)	0.0246	(0.31)
Low income scheme index	0.00161	(0.35)	0.00118	(0.22)
Constant	9.366***	(9.95)	9.393***	(10.10)
Observations	3079		2722	
R^2	0.098		0.044	

t statistics in parentheses (Standard errors clustered by practice)
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Notes: omitted category Income <50K
Models presented are variations of Equation 6.7

Table 6.8: OLS regressions of the effect of practice P4P exposure on job satisfaction using cross-sections of the GP WLS

	Job satisfaction 2005		*Job satisfaction 2008*	
QOF income exposure (Practice)	-0.0480	(-1.01)	-0.0509	(-1.87)
50-70K	0.101	(0.84)		
70-85K	0.00788	(0.06)		
85-100K	0.324**	(2.66)		
100-120K	0.394**	(3.18)		
120-150K	0.605***	(4.34)		
150+K	0.625**	(3.21)		
50-75K			0.279	(1.84)
75-100K			0.191	(1.26)
100-125K			0.407**	(2.63)
125-150K			0.616***	(3.60)
150-175K			0.627**	(3.03)
175+K			0.852***	(3.49)
Male	-0.147*	(-2.11)	-0.263***	(-3.77)
Non-white GP	-0.222*	(-2.08)	-0.0355	(-0.35)
Age (years)	-0.149***	(-3.67)	-0.176***	(-4.27)
Age squared	0.00157***	(3.72)	0.00184***	(4.36)
Patients per GP/1000	-0.302***	(-4.36)	-0.324***	(-3.81)
Time in current practice (years)	-0.00000759	(-0.00)	0.00699	(1.18)
PMS practice	0.0768	(1.32)	0.0448	(0.74)
Rural practice	0.0447	(0.60)	0.0319	(0.38)
Low income scheme index	-0.00225	(-0.50)	-0.00107	(-0.19)
Constant	9.546***	(8.57)	9.893***	(8.82)
Observations	1934		2016	
R^2	0.045		0.037	

Omitted categories: Income <50K
t statistics in parentheses (Standard errors clustered by practice)
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Models presented are variations of Equation 6.5

Table 6.9: OLS regressions of the effect of practice P4P exposure on changes in job satisfaction using a longitudinal sample of the GP WLS

	Change in job satisfaction 2004-2005		*Change in job satisfaction 2004-2008*		*Change in job satisfaction 2004-2008 (Balanced panel)*	
QOF income exposure (Practice)	0.0542	(0.81)	0.0847	(1.60)	0.0840	(1.57)
Change in income (000s)	0.00646	(0.58)	-0.00414	(-0.48)	-0.00424	(-0.49)
Male	0.0195	(0.19)	0.00612	(0.05)	0.0340	(0.28)
Non-white GP	0.129	(0.85)	-0.196	(-0.90)	-0.235	(-1.12)
Age (years)	-0.00572	(-0.08)	0.130	(1.07)	0.0987	(0.80)
Age squared	-0.0000247	(-0.03)	-0.00124	(-0.93)	-0.000881	(-0.66)
Patients per GP/1000	0.0307	(0.31)	0.274*	(2.13)	0.239	(1.85)
Time in current practice (years)	0.0153	(1.78)	-0.00442	(-0.32)	-0.00280	(-0.20)
PMS practice	-0.0627	(-0.78)	-0.144	(-1.30)	-0.119	(-1.08)
Rural practice	-0.0147	(-0.14)	-0.147	(-0.93)	-0.124	(-0.78)
Low income scheme index	-0.00689	(-1.15)	0.00685	(0.70)	0.00688	(0.69)
Constant	-0.00964	(-0.01)	-4.937	(-1.66)	-4.221	(-1.40)
Observations	1201		743		729	
R^2	0.008		0.022		0.020	

Omitted categories: Income <50K
t statistics in parentheses (Standard errors clustered by practice)
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Models presented are variations of Equation 6.6

Table 6.10: OLS regressions of the effect of practice P4P exposure on job satisfaction using a longitudinal sample of the GP WLS and the continuous DID approach

	Job satisfaction 2004 & 2005		*Job satisfaction 2004 & 2008*	
QOF income exposure * 2005	0.00856	(0.13)	-0.0117	(-0.25)
QOF income exposure (Practice)	-0.0656	(-0.99)	-0.0476	(-1.06)
Year (2005 or 2008)	0.411	(0.56)	0.0471	(0.06)
50-70K * 2005	-0.0554	(-0.36)		
70-85K * 2005	-0.127	(-0.81)		
85-100K * 2005	-0.000376	(-0.00)		
100-120K * 2005	-0.0459	(-0.28)		
120-150K * 2005	0.130	(0.51)		
150+K * 2005	1.320*	(2.30)		
50-70K	0.164	(1.34)	0.149	(0.99)
70-85K	0.158	(1.19)	0.263	(1.69)
85-100K	0.366*	(2.58)	0.324	(1.83)
100-120K	0.488**	(3.16)	0.372	(1.95)
120-150K	0.523*	(2.14)	0.457	(1.56)
150+K	-0.629	(-1.09)	-1.393	(-1.87)
50-75K			0.283	(1.86)
75-100K			0.205	(1.35)
100-125K			0.419**	(2.72)
125-150K			0.628***	(3.70)
150-175K			0.638**	(3.09)
175+K			0.855***	(3.50)
Male	-0.173**	(-2.62)	-0.252***	(-3.85)
Non-white GP	-0.227*	(-2.32)	-0.0616	(-0.62)
Age (years)	-0.155***	(-3.91)	-0.182***	(-4.91)
Age squared	0.00165***	(3.99)	0.00188***	(4.92)
Patients per GP/1000	-0.375***	(-5.91)	-0.397***	(-5.50)
Time in current practice (years)	-0.00316	(-0.59)	0.00623	(1.13)
PMS practice	0.121*	(2.22)	0.107	(1.89)
Rural practice	0.0136	(0.19)	0.0466	(0.61)
Low income scheme index	-0.0000514	(-0.01)	-0.000382	(-0.07)
Constant	9.454***	(7.86)	10.26***	(8.50)
Observations	3171		2798	
R^2	0.095		0.044	

t statistics in parentheses (Standard errors clustered by practice)
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Notes: omitted category Income <50K
Models presented are variations of Equation 6.7

Table 6.11: OLS regressions of the effect of GP P4P exposure on GP working lives in 2005 using a longitudinal sample of the GP WLS and the continuous DID approach

	High likelihood of quitting	*Hours per week*	*Life satisfaction*	*Physical working conditions*	*Choose method of working*	*Colleagues*	*Recognition for good work*	*Amount of responsibility*	*Re-muneration*	*Opportunity to use abilities*	*Hours of work*	*Variety in job*
QOF income exposure * 2005	-0.00479	0.0147	0.0131	0.0254	0.0131	0.0163	0.0201	0.0108	0.0147	0.00583	0.00531	0.0150
	(-1.42)	(0.14)	(1.03)	(1.66)	(0.98)	(1.29)	(1.49)	(0.78)	(1.12)	(0.45)	(0.35)	(1.17)
QOF income exposure (GP)	0.00120	0.650***	-0.0423***	-0.0637***	-0.0305*	-0.00614	-0.0236	-0.0155	-0.0407**	-0.0126	-0.0413**	-0.00355
	(0.38)	(6.31)	(-3.62)	(-4.35)	(-2.40)	(-0.50)	(-1.80)	(-1.13)	(-3.16)	(-0.97)	(-2.78)	(-0.29)
2005	-0.00625	-4.672**	0.143	-0.157	-0.0161	-0.124	0.260	0.201	0.838***	0.124	0.620*	0.00378
	(-0.12)	(-3.16)	(0.71)	(-0.61)	(-0.08)	(-0.61)	(1.17)	(0.93)	(3.76)	(0.60)	(2.55)	(0.02)
50-70K * 2005	0.00979	-5.129***	-0.0366	-0.207	0.0251	-0.116	-0.141	0.0820	-0.156	0.213	0.187	-0.261
	(0.22)	(-4.12)	(-0.21)	(-0.92)	(0.13)	(-0.67)	(-0.74)	(0.44)	(-0.83)	(1.21)	(0.90)	(-1.46)
70-85K * 2005	0.0807	-1.134	-0.0888	-0.138	-0.0889	-0.196	-0.320	-0.149	-0.458*	-0.110	0.0541	-0.222
	(1.78)	(-0.95)	(-0.51)	(-0.63)	(-0.48)	(-1.16)	(-1.70)	(-0.80)	(-2.45)	(-0.63)	(0.26)	(-1.27)
85-100K * 2005	0.0628	-0.605	-0.0242	-0.114	-0.0129	-0.153	-0.228	-0.0699	-0.516**	-0.0240	0.163	-0.00879
	(1.41)	(-0.51)	(-0.14)	(-0.52)	(-0.07)	(-0.92)	(-1.21)	(-0.38)	(-2.72)	(-0.14)	(0.79)	(-0.05)
100-120K * 2005	0.0438	0.00511	0.0546	-0.151	0.0888	-0.253	-0.292	-0.153	-0.515**	-0.164	0.321	-0.124
	(0.85)	(0.00)	(0.28)	(-0.66)	(0.45)	(-1.45)	(-1.41)	(-0.73)	(-2.61)	(-0.89)	(1.39)	(-0.66)
120-150K * 2005	0.0453	3.525	0.317	0.526	-0.0354	0.252	-0.153	0.123	-0.279	0.420	0.567	0.222
	(0.60)	(1.72)	(0.93)	(1.39)	(-0.12)	(0.76)	(-0.48)	(0.38)	(-0.78)	(1.17)	(1.62)	(0.64)
150+K * 2005	-0.167	-0.117	0.641	0.757	1.443*	0.872	0.906	1.120	0.234	1.149*	0.468	0.662
	(-1.19)	(-0.02)	(1.04)	(1.19)	(2.44)	(1.33)	(1.83)	(1.65)	(0.38)	(2.24)	(0.84)	(1.29)
50-70K	-0.0637	10.70***	0.149	0.343*	0.111	0.0862	0.245	0.0717	0.333*	-0.00739	-0.188	0.275*
	(-1.87)	(10.44)	(1.17)	(2.07)	(0.82)	(0.66)	(1.73)	(0.49)	(2.36)	(-0.06)	(-1.20)	(2.03)
70-85K	-0.0720*	11.87***	0.0777	0.481**	0.319*	0.0830	0.386**	0.257	0.875***	0.239	-0.104	0.309*
	(-1.97)	(11.71)	(0.58)	(2.88)	(2.23)	(0.63)	(2.64)	(1.69)	(5.86)	(1.76)	(-0.64)	(2.21)
85-100K	-0.0685	12.75***	0.304*	0.555**	0.453**	0.202	0.408**	0.328*	1.146***	0.368*	-0.0412	0.218
	(-1.77)	(11.98)	(2.13)	(3.08)	(3.01)	(1.46)	(2.61)	(2.05)	(7.15)	(2.54)	(-0.24)	(1.49)
100-120K	-0.100*	12.27***	0.357*	0.625***	0.341*	0.311*	0.600***	0.584**	1.494***	0.554***	-0.00959	0.361*
	(-2.12)	(9.80)	(2.14)	(3.30)	(1.98)	(2.11)	(3.32)	(3.05)	(8.69)	(3.44)	(-0.05)	(2.20)
120-150K	-0.116	8.820***	0.175	0.173	0.666*	-0.170	0.660*	0.401	1.466***	0.0850	-0.113	0.0456
	(-1.65)	(4.56)	(0.55)	(0.50)	(2.41)	(-0.55)	(2.22)	(1.31)	(4.34)	(0.25)	(-0.35)	(0.14)
Observations	3069	3050	3083	3082	3080	3073	3077	3078	3081	3083	3083	3082
R^2	0.288	0.390	0.058	0.039	0.040	0.032	0.048	0.051	0.197	0.054	0.113	0.034

t statistics in parentheses (Standard errors clustered by practice)

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

Notes: control variables included in all models. Omitted category income <50K

Models presented are variations of Equation 6.7

Table 6.12: OLS regressions of the effect of practice P4P exposure on GP working lives in 2005 using a longitudinal sample of the GP WLS and the continuous DID approach

	High likelihood of quitting	*Hours per week*	*Life satisfaction*	*Physical working conditions*	*Choose method of working*	*Colleagues*	*Recognition for good work*	*Amount of responsibility*	*Re-muneration*	*Opportunity to use abilities*	*Hours of work*	*Variety in job*
QOF income exposure * 2005	0.0171	0.0671	0.0532	0.0558	-0.0540	-0.0735	-0.0399	0.0281	0.0584	0.0468	0.0463	-0.0144
	(0.78)	(0.10)	(0.68)	(0.58)	(-0.63)	(-0.97)	(-0.47)	(0.32)	(0.69)	(0.61)	(0.49)	(-0.18)
QOF income exposure (Practice)	-0.0302	0.956	-0.0740	-0.257***	-0.0430	0.0343	0.00994	-0.0733	-0.0317	-0.0472	-0.178*	-0.0179
	(-1.68)	(1.64)	(-1.17)	(-3.34)	(-0.62)	(0.58)	(0.15)	(-1.05)	(-0.44)	(-0.76)	(-2.32)	(-0.29)
2005	-0.258	-5.556	-0.229	-0.457	0.769	0.892	0.977	0.0285	0.411	-0.281	0.205	0.347
	(-1.07)	(-0.74)	(-0.27)	(-0.43)	(0.81)	(1.08)	(1.04)	(0.03)	(0.44)	(-0.33)	(0.20)	(0.40)
50-70K * 2005	0.0114	-5.690***	0.0242	-0.0571	0.0737	-0.0880	-0.104	0.134	-0.121	0.213	0.261	-0.214
	(0.27)	(-4.76)	(0.14)	(-0.26)	(0.41)	(-0.52)	(-0.56)	(0.75)	(-0.66)	(1.26)	(1.30)	(-1.25)
70-85K * 2005	0.0682	-0.975	-0.0732	-0.0437	-0.102	-0.175	-0.312	-0.158	-0.436*	-0.126	0.0832	-0.178
	(1.54)	(-0.85)	(-0.43)	(-0.21)	(-0.57)	(-1.07)	(-1.70)	(-0.88)	(-2.37)	(-0.75)	(0.42)	(-1.06)
85-100K * 2005	0.0577	-0.554	-0.0247	-0.0114	-0.0235	-0.137	-0.195	-0.0663	-0.498**	-0.0281	0.178	0.0246
	(1.31)	(-0.47)	(-0.14)	(-0.05)	(-0.13)	(-0.84)	(-1.05)	(-0.37)	(-2.66)	(-0.17)	(0.88)	(0.14)
100-120K * 2005	0.0433	0.546	0.0293	-0.151	0.0132	-0.232	-0.260	-0.132	-0.525**	-0.195	0.279	-0.140
	(0.85)	(0.42)	(0.16)	(-0.68)	(0.07)	(-1.37)	(-1.28)	(-0.64)	(-2.69)	(-1.09)	(1.24)	(-0.76)
120-150K * 2005	0.0492	3.943	0.293	0.617	-0.0702	0.227	-0.173	0.164	-0.271	0.412	0.563	0.267
	(0.65)	(1.91)	(0.86)	(1.66)	(-0.23)	(0.69)	(-0.54)	(0.51)	(-0.76)	(1.16)	(1.61)	(0.77)
150+K * 2005	-0.179	0.0167	0.620	0.809	1.448*	0.891	0.942	1.192	0.282	1.163*	0.435	0.704
	(-1.29)	(0.00)	(1.02)	(1.31)	(2.47)	(1.36)	(1.94)	(1.77)	(0.46)	(2.30)	(0.80)	(1.39)
50-70K	-0.0649*	12.01***	0.0497	0.180	0.0344	0.0703	0.184	0.0194	0.251	-0.0209	-0.295*	0.243
	(-2.05)	(12.69)	(0.41)	(1.16)	(0.27)	(0.58)	(1.37)	(0.15)	(1.87)	(-0.17)	(-2.00)	(1.95)
70-85K	-0.0729*	13.61***	-0.0301	0.278	0.263	0.0919	0.362**	0.226	0.776***	0.229	-0.220	0.279*
	(-2.09)	(14.29)	(-0.23)	(1.75)	(1.94)	(0.73)	(2.61)	(1.61)	(5.43)	(1.80)	(-1.43)	(2.15)
85-100K	-0.0718	14.67***	0.205	0.350*	0.391**	0.197	0.376*	0.292	1.056***	0.346*	-0.186	0.207
	(-1.93)	(14.26)	(1.46)	(2.03)	(2.68)	(1.50)	(2.49)	(1.93)	(6.78)	(2.50)	(-1.11)	(1.48)
100-120K	-0.107*	13.84***	0.286	0.485**	0.337*	0.307*	0.557**	0.538**	1.424***	0.560***	-0.0898	0.405**
	(-2.31)	(11.62)	(1.78)	(2.67)	(2.02)	(2.19)	(3.19)	(2.94)	(8.51)	(3.69)	(-0.45)	(2.58)
120-150K	-0.131	10.40***	0.106	-0.0278	0.615*	-0.154	0.659*	0.359	1.390***	0.0782	-0.240	0.0429
	(-1.86)	(5.33)	(0.33)	(-0.08)	(2.24)	(-0.50)	(2.22)	(1.19)	(4.14)	(0.23)	(-0.74)	(0.13)
Observations	3162	3141	3176	3175	3172	3164	3170	3170	3174	3176	3176	3175
R^2	0.287	0.370	0.055	0.038	0.038	0.030	0.048	0.051	0.198	0.054	0.112	0.033

t statistics in parentheses (Standard errors clustered by practice)
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Notes: control variables included in all models. Omitted category income <50K
Models presented are variations of Equation 6.7

Table 6.13: OLS regressions of the effect of GP P4P exposure on GP working lives in 2008 using a longitudinal sample of the GP WLS and the continuous DID approach

	High likelihood of quitting	*Hours per week*	*Life satisfaction*	*Physical working conditions*	*Choose method of working*	*Colleagues*	*Recognition for good work*	*Amount of responsibility*	*Re-muneration*	*Opportunity to use abilities*	*Hours of work*	*Variety in job*
QOF income exposure * 2008	-0.00282	0.0992	-0.00307	0.00726	-0.00246	0.00329	-0.00120	-0.00910	0.00488	0.00230	-0.00349	-0.00309
	(-1.50)	(1.82)	(-0.47)	(0.96)	(-0.34)	(0.51)	(-0.16)	(-1.29)	(0.63)	(0.35)	(-0.41)	(-0.45)
QOF income exposure (GP)	0.00203	0.0398	-0.00634	-0.0113	0.000250	-0.0107	-0.00441	0.000512	-0.0124	-0.00670	-0.00720	-0.000274
	(1.20)	(0.86)	(-1.13)	(-1.75)	(0.04)	(-1.81)	(-0.65)	(0.08)	(-1.80)	(-1.13)	(-0.93)	(-0.04)
2008	0.0844	-5.086**	0.288	0.108	-0.0750	-0.251	0.157	0.372	-0.264	-0.131	0.359	0.251
	(1.28)	(-2.89)	(1.17)	(0.37)	(-0.29)	(-1.12)	(0.58)	(1.44)	(-0.94)	(-0.54)	(1.20)	(1.04)
50-70K	-0.0637	11.32***	0.159	0.217	-0.0455	0.0890	0.218	0.0753	-0.00400	0.0690	-0.272	0.333*
	(-1.72)	(9.26)	(1.02)	(1.05)	(-0.27)	(0.58)	(1.24)	(0.44)	(-0.02)	(0.43)	(-1.39)	(2.03)
70-85K	-0.0704	13.34***	0.146	0.306	0.201	0.0821	0.373*	0.236	0.456*	0.310*	-0.151	0.389*
	(-1.72)	(10.77)	(0.89)	(1.46)	(1.19)	(0.53)	(2.09)	(1.31)	(2.50)	(1.97)	(-0.76)	(2.39)
85-100K	-0.0856	12.87***	0.337	0.351	0.279	0.105	0.439*	0.231	0.909***	0.448**	0.0868	0.282
	(-1.96)	(9.92)	(1.88)	(1.53)	(1.50)	(0.61)	(2.28)	(1.18)	(4.53)	(2.59)	(0.40)	(1.56)
100-120K	-0.105	13.38***	0.356	0.663**	0.317	0.276	0.450*	0.401	1.207***	0.580**	-0.126	0.383
	(-1.81)	(8.44)	(1.81)	(2.90)	(1.54)	(1.71)	(1.99)	(1.60)	(5.57)	(2.87)	(-0.47)	(1.87)
120-150K	-0.141	10.19***	0.359	0.259	0.720*	-0.177	0.465	0.305	1.324**	0.230	-0.186	-0.0609
	(-1.93)	(4.45)	(0.98)	(0.60)	(2.07)	(-0.48)	(1.34)	(0.84)	(3.25)	(0.63)	(-0.46)	(-0.16)
50-75K	-0.0529	5.350***	0.181	0.290	0.234	0.219	0.162	0.138	0.587**	0.419**	0.0271	0.197
	(-1.18)	(4.86)	(1.12)	(1.64)	(1.49)	(1.53)	(0.94)	(0.88)	(3.08)	(2.65)	(0.15)	(1.36)
75-100K	-0.0390	12.30***	0.165	0.178	0.0860	0.239	0.0819	0.154	0.689***	0.332*	-0.310	0.134
	(-0.87)	(11.22)	(1.04)	(1.03)	(0.56)	(1.72)	(0.48)	(1.00)	(3.65)	(2.13)	(-1.78)	(0.94)
100-125K	-0.0873	13.79***	0.216	0.242	0.260	0.201	0.380*	0.296	1.087***	0.593***	-0.157	0.297*
	(-1.91)	(12.25)	(1.34)	(1.38)	(1.64)	(1.41)	(2.18)	(1.87)	(5.69)	(3.77)	(-0.89)	(2.06)
125-150K	-0.0800	14.26***	0.450**	0.168	0.373*	0.210	0.497**	0.295	1.371***	0.576***	0.110	0.297
	(-1.60)	(11.27)	(2.59)	(0.87)	(2.11)	(1.30)	(2.63)	(1.71)	(6.73)	(3.35)	(0.57)	(1.86)
150-175K	-0.0650	12.36***	0.579**	0.256	0.580*	0.231	0.433	0.289	1.540***	0.647**	0.0802	0.285
	(-1.02)	(8.21)	(2.69)	(1.05)	(2.57)	(1.13)	(1.80)	(1.26)	(6.15)	(2.88)	(0.34)	(1.33)
175+K	-0.118	15.20***	0.245	0.284	0.563*	0.220	0.570*	0.380	1.959***	0.776**	0.404	0.0959
	(-1.68)	(8.01)	(0.93)	(1.06)	(2.15)	(0.94)	(2.13)	(1.36)	(7.17)	(2.97)	(1.52)	(0.37)
Observations	2709	2687	2722	2725	2717	2722	2716	2721	2723	2724	2721	2712
R^2	0.267	0.308	0.035	0.026	0.020	0.027	0.034	0.029	0.094	0.040	0.044	0.032

t statistics in parentheses (Standard errors clustered by practice)
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Notes: control variables included in all models. Omitted category income <50K
Models presented are variations of Equation 6.7

Table 6.14: OLS regressions of the effect of practice P4P exposure on GP working lives in 2008 using a longitudinal sample of the GP WLS and the continuous DID approach

	High likelihood of quitting	*Hours per week*	*Life satisfaction*	*Physical working conditions*	*Choose method of working*	*Colleagues*	*Recognition for good work*	*Amount of responsibility*	*Re-muneration*	*Opportunity to use abilities*	*Hours of work*	*Variety in job*
QOF income	0.0171	0.0309	0.0305	0.00239	-0.0308	-0.0665	0.0210	-0.0241	-0.0179	0.0594	0.0513	0.0187
exposure * 2008	(1.33)	(0.07)	(0.61)	(0.04)	(-0.57)	(-1.30)	(0.38)	(-0.42)	(-0.31)	(1.21)	(0.83)	(0.38)
QOF income	-0.00133	0.634	-0.0846*	-0.0851	-0.0557	0.0490	-0.0325	-0.0552	-0.0224	-0.0891*	-0.137*	-0.0685
exposure (GP)	(-0.13)	(1.74)	(-2.00)	(-1.51)	(-1.20)	(1.07)	(-0.69)	(-1.09)	(-0.44)	(-2.09)	(-2.50)	(-1.60)
2008	-0.262	-3.033	-0.364	0.207	0.366	0.987	-0.244	0.541	0.102	-1.156	-0.612	-0.187
	(-1.16)	(-0.42)	(-0.41)	(0.18)	(0.38)	(1.09)	(-0.25)	(0.54)	(0.10)	(-1.32)	(-0.56)	(-0.22)
50-70K	-0.0573	11.79***	0.109	0.178	-0.0995	0.0882	0.167	0.0391	-0.0827	-0.00766	-0.305	0.286
	(-1.61)	(10.19)	(0.72)	(0.88)	(-0.61)	(0.59)	(0.99)	(0.24)	(-0.49)	(-0.05)	(-1.64)	(1.80)
70-85K	-0.0726	13.82***	0.110	0.280	0.178	0.0749	0.336	0.220	0.403*	0.255	-0.203	0.350*
	(-1.83)	(11.70)	(0.70)	(1.36)	(1.08)	(0.49)	(1.96)	(1.26)	(2.28)	(1.66)	(-1.05)	(2.21)
85-100K	-0.0817	13.83***	0.272	0.292	0.257	0.0867	0.377*	0.216	0.819***	0.383*	0.0166	0.236
	(-1.92)	(11.07)	(1.55)	(1.30)	(1.43)	(0.51)	(2.02)	(1.14)	(4.21)	(2.28)	(0.08)	(1.35)
100-120K	-0.104	14.43***	0.234	0.578*	0.249	0.204	0.404	0.337	1.169***	0.516**	-0.188	0.345
	(-1.88)	(9.71)	(1.21)	(2.56)	(1.21)	(1.23)	(1.88)	(1.39)	(5.62)	(2.70)	(-0.74)	(1.70)
120-150K	-0.135	11.29***	0.257	0.148	0.665	-0.202	0.398	0.243	1.211**	0.124	-0.332	-0.132
	(-1.88)	(4.91)	(0.71)	(0.36)	(1.93)	(-0.55)	(1.15)	(0.68)	(3.04)	(0.34)	(-0.83)	(-0.34)
50-75K	-0.0635	5.449***	0.215	0.299	0.233	0.217	0.145	0.152	0.598**	0.432**	0.0198	0.209
	(-1.43)	(5.11)	(1.32)	(1.71)	(1.50)	(1.53)	(0.85)	(0.98)	(3.23)	(2.73)	(0.11)	(1.46)
75-100K	-0.0584	12.79***	0.168	0.167	0.0847	0.212	0.0514	0.139	0.693***	0.339*	-0.340*	0.132
	(-1.32)	(12.30)	(1.05)	(0.99)	(0.56)	(1.56)	(0.30)	(0.92)	(3.82)	(2.18)	(-1.99)	(0.93)
100-125K	-0.0994*	14.45***	0.206	0.229	0.254	0.175	0.351*	0.270	1.074***	0.594***	-0.236	0.288*
	(-2.20)	(13.68)	(1.27)	(1.33)	(1.63)	(1.25)	(2.04)	(1.75)	(5.87)	(3.80)	(-1.37)	(2.04)
125-150K	-0.0817	15.19***	0.417*	0.110	0.333	0.148	0.422*	0.233	1.315***	0.544**	0.00659	0.263
	(-1.66)	(12.68)	(2.40)	(0.58)	(1.92)	(0.94)	(2.27)	(1.37)	(6.70)	(3.19)	(0.03)	(1.67)
150-175K	-0.0802	13.57***	0.525*	0.225	0.563*	0.158	0.392	0.236	1.506***	0.654**	-0.0218	0.281
	(-1.30)	(9.66)	(2.50)	(0.94)	(2.56)	(0.80)	(1.67)	(1.07)	(6.32)	(2.98)	(-0.10)	(1.36)
175+K	-0.130	17.18***	0.148	0.194	0.496	0.134	0.524*	0.280	1.889***	0.744**	0.170	0.0653
	(-1.91)	(9.53)	(0.57)	(0.74)	(1.94)	(0.59)	(2.00)	(1.04)	(7.19)	(2.91)	(0.65)	(0.26)
Observations	2785	2763	2798	2801	2793	2797	2792	2797	2799	2800	2797	2788
R^2	0.262	0.308	0.036	0.028	0.025	0.024	0.033	0.031	0.093	0.042	0.046	0.034

t statistics in parentheses (Standard errors clustered by practice)
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Notes: control variables included in all models. Omitted category income <50K
Models presented are variations of Equation 6.7

Table 6.15: OLS regressions of the effect of GP and practice P4P exposure on job satisfaction using a longitudinal sample of the GP WLS and the continuous DID approach stratified by GP characteristics

	Male	*Female*	*<40*	*40-49*	*50-59*	*60+*
QOF income exposure * 2005	0.00729	0.0331*	0.0314	0.0337*	-0.0178	0.0282
	(0.39)	(2.01)	(1.11)	(2.02)	(-0.80)	(0.48)
QOF income exposure (GP)	-0.0259	-0.0492***	-0.0294	-0.0551***	-0.0171	-0.105
	(-1.35)	(-3.44)	(-1.18)	(-3.70)	(-0.75)	(-1.39)
2005	0.347	-0.0785	-0.0472	0.0279	0.709*	0.00756
	(1.33)	(-0.29)	(-0.11)	(0.11)	(2.15)	(0.01)
Observations	1992	1087	431	1435	1057	156
R^2	0.113	0.071	0.077	0.089	0.123	0.253
QOF income exposure * 2005	0.0262	-0.0306	0.153	-0.0547	0.0221	0.376
	(0.28)	(-0.21)	(0.77)	(-0.51)	(0.15)	(1.06)
QOF income exposure	-0.0349	-0.126	-0.180	-0.0162	-0.0920	-0.368
(Practice)	(-0.45)	(-1.06)	(-1.06)	(-0.18)	(-0.75)	(-1.25)
2005	0.171	0.812	-1.237	1.122	0.222	-3.578
	(0.17)	(0.52)	(-0.57)	(0.96)	(0.14)	(-0.95)
Observations	2051	1120	449	1474	1086	162
R^2	0.112	0.068	0.074	0.085	0.119	0.253
QOF income exposure * 2008	-0.00832	0.00131	-0.00163	0.00425	-0.00215	-0.0918***
	(-0.86)	(0.11)	(-0.09)	(0.37)	(-0.18)	(-3.64)
QOF income exposure (GP)	0.00216	-0.00723	-0.00210	-0.0126	-0.000914	0.0946***
	(0.26)	(-0.62)	(-0.14)	(-1.26)	(-0.08)	(3.67)
2008	-0.0792	-0.135	-0.182	-0.313	0.547	3.732***
	(-0.14)	(-0.37)	(-0.30)	(-0.83)	(1.22)	(4.19)
Observations	1746	976	354	1177	1059	132
R^2	0.049	0.051	0.074	0.044	0.062	0.324
QOF income exposure * 2008	0.0492	-0.124	0.316*	-0.0555	-0.129	-1.398
	(0.77)	(-1.43)	(2.24)	(-0.78)	(-1.47)	(-1.56)
QOF income exposure	-0.0875	0.0205	-0.435***	-0.0199	0.0589	1.613
(Practice)	(-1.59)	(0.26)	(-3.51)	(-0.33)	(0.76)	(1.82)
2008	-1.304	2.042	-5.677*	0.775	2.564	26.37
	(-1.06)	(1.35)	(-2.32)	(0.62)	(1.64)	(1.65)
Observations	1796	1002	365	1209	1090	134
R^2	0.049	0.055	0.110	0.044	0.064	0.340

t statistics in parentheses (Standard errors clustered by practice)
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Notes: control variables included in all models.
Models presented are variations of Equation 6.7

Table 6.16: OLS regressions of the effect of GP and practice P4P exposure on job satisfaction using a longitudinal sample of the GP WLS and the continuous DID approach stratified by practice characteristics

	Large practices	*Small practices*	*PMS*	*GMS*	*Single partner*	*2-5 partners*	*6+ partners*
QOF income exposure * 2005	-0.00462	0.0306*	-0.000420	0.0246	0.0259	0.0253	-0.00407
	(-0.25)	(1.99)	(-0.02)	(1.56)	(0.45)	(1.76)	(-0.17)
QOF income exposure (GP)	-0.0344*	-0.0324*	-0.0461**	-0.0315*	-0.0554	-0.0330*	-0.0366
	(-2.03)	(-2.21)	(-2.60)	(-2.15)	(-0.66)	(-2.45)	(-1.73)
2005	0.517	0.0253	0.391	0.152	-0.207	0.155	0.427
	(1.96)	(0.11)	(1.47)	(0.61)	(-0.20)	(0.69)	(1.35)
Observations	1515	1564	1403	1676	172	1764	1143
R^2	0.097	0.103	0.090	0.097	0.169	0.100	0.084
QOF income exposure * 2005	-0.0410	0.0428	-0.000744	0.0244	-0.0944	0.0291	-0.0470
	(-0.37)	(0.38)	(-0.01)	(0.22)	(-0.23)	(0.30)	(-0.33)
QOF income exposure	-0.0297	-0.0605	-0.0450	-0.0830	0.147	-0.0960	-0.0329
(Practice)	(-0.31)	(-0.66)	(-0.49)	(-0.87)	(0.42)	(-1.22)	(-0.26)
2005	0.918	0.0288	0.419	0.284	1.309	0.225	0.904
	(0.77)	(0.02)	(0.34)	(0.23)	(0.30)	(0.21)	(0.59)
Observations	1569	1602	1442	1729	176	1804	1191
R^2	0.095	0.098	0.082	0.097	0.178	0.099	0.075
QOF income exposure * 2008	-0.0121	0.00357	-0.0106	0.00436	0.0111	0.00192	-0.0191
	(-1.13)	(0.39)	(-1.14)	(0.41)	(0.38)	(0.22)	(-1.40)
QOF income exposure (GP)	0.00292	-0.00560	0.00153	-0.00744	0.00409	-0.00494	0.00510
	(0.30)	(-0.69)	(0.18)	(-0.80)	(0.15)	(-0.64)	(0.43)
2008	0.0994	-0.155	-0.185	0.0114	-0.194	-0.0342	0.178
	(0.26)	(-0.47)	(-0.53)	(0.03)	(-0.16)	(-0.11)	(0.36)
Observations	1427	1295	1228	1494	148	1691	883
R^2	0.059	0.042	0.048	0.045	0.140	0.041	0.070
QOF income exposure * 2008	-0.0589	0.0496	-0.00247	-0.00142	-0.0282	-0.00996	-0.0181
	(-0.86)	(0.61)	(-0.03)	(-0.02)	(-0.08)	(-0.16)	(-0.19)
QOF income exposure	-0.0222	-0.0811	-0.0476	-0.0658	-0.0583	-0.0544	-0.0724
(Practice)	(-0.36)	(-1.17)	(-0.72)	(-1.03)	(-0.18)	(-1.00)	(-0.82)
2008	0.891	-1.036	-0.400	0.0887	0.501	0.181	-0.0874
	(0.74)	(-0.72)	(-0.29)	(0.07)	(0.08)	(0.16)	(-0.05)
Observations	1476	1322	1262	1536	151	1725	922
R^2	0.058	0.044	0.047	0.045	0.132	0.043	0.069

t statistics in parentheses (Standard errors clustered by practice)
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Notes: control variables included in all models.
Models presented are variations of Equation 6.7

Table 6.17: OLS regressions of the effect of GP and practice P4P exposure on job satisfaction using a longitudinal sample of the GP WLS and the continuous DID approach stratified by practice characteristics in 2004

	Large practices	*Small practices*	*PMS*	*GMS*	*Single partner*	*2-5 partners*	*6+ partners*
QOF income exposure * 2005	-0.000390	0.0263	0.00300	0.0221	-0.0349	0.0299	-0.000834
	(-0.02)	(1.52)	(0.14)	(1.24)	(-0.56)	(1.94)	(-0.03)
QOF income exposure (GP)	-0.0337	-0.0255	-0.0372*	-0.0328*	0.0145	-0.0360**	-0.0281
	(-1.93)	(-1.64)	(-2.00)	(-2.12)	(0.17)	(-2.61)	(-1.28)
2005	0.458	0.0644	0.379	0.144	0.772	0.0956	0.377
	(1.60)	(0.24)	(1.33)	(0.51)	(0.70)	(0.40)	(1.05)
Observations	1224	1188	1112	1300	127	1403	882
R^2	0.102	0.106	0.092	0.091	0.267	0.093	0.085
QOF income exposure * 2005	-0.0701	0.0794	0.0291	-0.00307	-0.123	0.0429	-0.0425
	(-0.59)	(0.60)	(0.23)	(-0.02)	(-0.29)	(0.41)	(-0.25)
QOF income exposure	-0.0441	-0.0529	-0.0466	-0.0940	0.221	-0.0989	-0.0235
(Practice)	(-0.48)	(-0.56)	(-0.51)	(-0.98)	(0.61)	(-1.25)	(-0.19)
2005	1.224	-0.393	0.120	0.534	1.571	0.0775	0.843
	(0.96)	(-0.28)	(0.09)	(0.39)	(0.35)	(0.07)	(0.46)
Observations	1268	1216	1140	1344	134	1433	917
R^2	0.094	0.105	0.085	0.089	0.258	0.093	0.073
QOF income exposure * 2008	-0.00373	0.00522	-0.00849	0.00914	0.00386	0.0113	-0.0137
	(-0.32)	(0.47)	(-0.80)	(0.74)	(0.10)	(1.07)	(-1.01)
QOF income exposure (GP)	0.00851	-0.00464	0.00407	-0.00518	0.0357	-0.00345	0.0135
	(0.86)	(-0.54)	(0.45)	(-0.56)	(1.01)	(-0.42)	(1.14)
2008	0.421	-0.481	0.464	-0.383	2.113	-0.149	-0.106
	(0.85)	(-1.00)	(1.01)	(-0.82)	(1.46)	(-0.35)	(-0.18)
Observations	805	708	714	799	80	846	587
R^2	0.073	0.069	0.067	0.069	0.385	0.061	0.094
QOF income exposure * 2008	0.00203	0.0914	0.0607	0.0245	0.0767	0.0419	0.0451
	(0.03)	(0.86)	(0.64)	(0.29)	(0.21)	(0.54)	(0.40)
QOF income exposure	-0.0379	-0.0907	-0.0533	-0.0889	-0.175	-0.0568	-0.0991
(Practice)	(-0.62)	(-1.27)	(-0.80)	(-1.39)	(-0.57)	(-1.03)	(-1.12)
2008	0.326	-2.074	-0.977	-0.613	0.409	-0.703	-1.134
	(0.24)	(-1.10)	(-0.58)	(-0.42)	(0.06)	(-0.50)	(-0.58)
Observations	835	721	735	821	82	864	610
R^2	0.070	0.070	0.064	0.070	0.365	0.062	0.091

t statistics in parentheses (Standard errors clustered by practice)
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Notes: control variables included in all models.
Models presented are variations of Equation 6.7

Table 6.18: OLS regression of the effect of practice P4P exposure on job satisfaction using a longitudinal sample of salaried GPs from the GP WLS and the continuous DID approach

	Job satisfaction 2004 & 2005		*Job satisfaction 2004 & 2008*	
P4P exposure * year	0.297	(1.23)	0.0746	(0.43)
P4P exposure (Practice)	-0.203	(-0.89)	0.0296	(0.21)
Year	-3.350	(-1.43)	-1.299	(-0.50)
Observations	118		208	
R^2	0.123		0.132	

Control variables included in all models.
t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Models presented are variations of Equation 6.7

Table 7.1: QOF indicators measuring the quality of diabetes care at practice level

Indicator	*2004/5*	*2005/6*	*2006/7*	*2007/8*	*2008/9*
Diabetes 1 = The practice can produce a register of all patients with diabetes Mellitus	✓	✓	×	×	×
Diabetes 2 = Patients with diabetes whose notes record BMI in the previous 15 months	✓	✓	✓	✓	✓
Diabetes 3 = Patients with diabetes in whom there is a record of smoking status in the previous 15 months except those who have never smoked where smoking status should be recorded once	✓	✓	×	×	×
Diabetes 4 = Patients with diabetes who smoke and whose notes contain a record that smoking cessation advice has been offered in the last 15 months	✓	✓	×	×	×
Diabetes 5 = Diabetic patients who have a record of HbA1c or equivalent in the previous 15 months	✓	✓	✓	✓	✓
Diabetes 6 = Patients with diabetes in whom the last HbA1C is 7.4 or less in last 15 months	✓	✓	×	×	×
Diabetes 7 = Patients with diabetes in whom the last HbA1C is 10 or less in last 15 months	✓	✓	✓	✓	✓
Diabetes 8 = Patients with diabetes who have a record of retinal screening in the previous 15 months	✓	✓	×	×	×
Diabetes 9 = Patients with diabetes with a record of presence or absence of peripheral pulses in the previous 15 months	✓	✓	✓	✓	✓
Diabetes 10 = Patients with diabetes with a record of neuropathy testing in the previous 15 months	✓	✓	✓	✓	✓
Diabetes 11 = Patients with diabetes who have a record of the blood pressure in the past 15 months	✓	✓	✓	✓	✓
Diabetes 12 = Patients with diabetes in whom the last blood pressure is 145/85 or less	✓	✓	✓	✓	✓
Diabetes 13 = Patients with diabetes who have a record of micro-albuminuria testing in the previous 15 months	✓	✓	✓	✓	✓
Diabetes 14 = Patients with diabetes who have a record of serum creatinine testing in the previous 15 months	✓	✓	×	×	×
Diabetes 15 = Patients with diabetes with proteinuria or micro-albuminuria who are treated with ACE inhibitors (or A2 antagonists)	✓	✓	✓	✓	✓
Diabetes 16 = Patients with diabetes who have a record of total cholesterol in the previous 15 months	✓	✓	✓	✓	✓
Diabetes 17 = Patients with diabetes whose last measured total cholesterol within previous 15 months is 5 or less	✓	✓	✓	✓	✓
Diabetes 18 = Patients with diabetes who have had influenza immunisation in the preceding 1 September to 31 March	✓	✓	✓	✓	✓
Diabetes 19 = The practice can produce a register of all patients aged 17 years and over with diabetes mellitus, which specifies whether the patient has Type 1 or Type 2 diabetes	×	×	✓	✓	✓
Diabetes 20 = Patients with diabetes in whom the last HbA1c is 7.5 or less (or equivalent test/reference range depending on local laboratory) in the previous 15 months	×	×	✓	✓	✓
Diabetes 21 = Patients with diabetes who have a record of retinal screening in the previous 15 months	×	×	✓	✓	✓
Diabetes 22 = Patients with diabetes who have a record of estimated glomerular filtration rate (eGFR) or serum creatinine testing in the previous 15 months	×	×	✓	✓	✓

Notes: Bold rows are variables included in analysis

Table 7.2: QOF indicators measuring the quality of hypertension care at practice level

Indicator	*2004/5*	*2005/6*	*2006/7*	*2007/8*	*2008/9*
Hypertension 1 = The practice can produce a register of patients with established hypertension	✓	✓	✓	✓	✓
Hypertension 2 = Patients with hypertension whose notes record smoking status at least once	✓	✓	×	×	×
Hypertension 3 = Patients with hypertension who smoke, whose notes contain a record that smoking cessation advice has been offered at least once	✓	✓	×	×	×
Hypertension 4 = Patients with hypertension in which there is a record of the blood pressure in the past 9 months	✓	✓	✓	✓	✓
Hypertension 5 = Patients with hypertension in whom the last blood pressure (measured in last 9 months) is 150/90 or less	✓	✓	✓	✓	✓

Notes: Bold rows are variables included in analysis

Table 7.3: ELSA questions on aspects of care for diabetes and hypertension

	Variable in ELSA	*Present in Wave 2*	*Present in Wave 3*	*Present in Wave 4*
1	**Blood pressure: whether checked by doctor or nurse in past year**	✓	✓	✓
2	High blood pressure: whether doctor or nurse explained it in understandable way	✓	×	×
3	High blood pressure: whether doctor or nurse gave choice of treatment	✓	×	×
4	**High blood pressure: whether doctor or nurse suggested medication to lower it**	✓	✓	✓
5	**High blood pressure: whether taking medication**	✓	✓	✓
6	High blood pressure: whether taking medication to prevent high level returning	×	✓	✓
7	MI: did your doctor ever tell that you should take a medication called a beta-blocker	✓	×	×
8	MI: nurse checks if respondent is taking beta-blockers	✓	×	×
9	**Diabetes: whether currently injects insulin**	✓	✓	✓
10	**Diabetes: whether taking medication (swallow)**	✓	✓	✓
11	**Diabetes: whether ACE inhibitor or A2 receptor blocker discussed by doctor**	✓	✓	✓
12	**Diabetes: whether taking ACE inhibitor or A2 receptor blocker**	✓	✓	✓
13	**Diabetes: whether been checked for protein in urine in past year by doctor**	✓	✓	✓
14	**Diabetes: whether been told has protein in urine by doctor**	✓	✓	✓
15	**Diabetes: whether been told has kidney trouble by doctor**	✓	✓	✓
16	**Diabetes: whether ever had glycosylated haemoglobin (A1c) test**	✓	✓	✓
17	**Diabetes: whether had A1c test in last 12 months**	✓	✓	✓
18	**Diabetes: whether had feet examined in past year by doctor or nurse**	✓	✓	✓
19	**Diabetes: whether received training to manage it**	✓	✓	✓
20	**Diabetes: how much knows about managing it (self-reported)**	✓	✓	✓

Notes: bold rows are variables included in analysis

Table 7.4: ELSA questions on quality of care compared to their QOF equivalent

ELSA	*QOF*
Blood pressure: whether checked by doctor or nurse in past year	Hypertension 4 = Patients with hypertension in which there is a record of the blood pressure in the past 9 months And Combined practice score for hypertension
High blood pressure: whether doctor or nurse suggested medication to lower it	Combined practice score for hypertension
High blood pressure: whether taking medication	Combined practice score for hypertension
Diabetes: whether currently injects insulin	Combined practice score for diabetes
Diabetes: whether taking medication (swallow)	Combined practice score for diabetes
Diabetes: whether ACE inhibitor or A2 receptor blocker discussed by doctor	Combined practice score for diabetes
Diabetes: whether taking ACE inhibitor or A2 receptor blocker	Diabetes 15 = Patients with diabetes with proteinuria or micro-albuminuria who are treated with ACE inhibitors (or A2 antagonists) Note: the requirement for a diagnosis of proteinuria or micro-albuminuria is removed from this indicator And Combined practice score for diabetes
Diabetes: whether been checked for protein in urine in past year by doctor	Combined practice score for diabetes
Diabetes: whether been told has protein in urine by doctor	Combined practice score for diabetes
Diabetes: whether been told has kidney trouble by doctor	Combined practice score for diabetes
Diabetes: whether ever had glycosylated haemoglobin (A1c) test	Combined practice score for diabetes
Diabetes: whether had A1c test in last 12 months	Diabetes 5 = Diabetic patients who have a record of HbA1c or equivalent in the previous 15 months And Combined practice score for diabetes
Diabetes: whether had feet examined in past year by doctor or nurse	Diabetes 9 = Patients with diabetes with a record of presence or absence of peripheral pulses in the previous 15 months Diabetes 10 = Patients with diabetes with a record of neuropathy testing in the previous 15 months And Combined practice score for diabetes
Diabetes: whether received training to manage it	Combined practice score for diabetes
Diabetes: how much knows about managing it (self-reported)	Combined practice score for diabetes

Notes: combined diabetes score is the practice performance on diabetes 2, 5, 7, 9-13, 15-18 from Table 7.1. Combined hypertension score is the practice performance on hypertension 4 and 5 from Table 7.2 [p.194]

Table 7.5: Details of the matching procedure and results for the 2004 wave of ELSA

Matching method	*Stage*	*Matching details*	*Number matched*
Reclink	1	Postcode only (no manual matches, all exact)	2214
Reclink	1	Postcode and practice name (no exact matched, all taken above score= 91%)	206
Reclink	1	Practice name, address 1 and postcode (no exact, manually changed 6 mistaken matches)	430
Reclink	1	Address 1 and postcode (no exact, manually changed 10 mistaken matches)	144
Reclink	1	Practice name, address 1 and address 2 (no exact, no mistaken matches found above score=65%)	521
Reclink	1	Postcode and practice name (repeated 2nd reclink, no exact matches, manually changed 59 mistaken matches)	92
Reclink	1	Practice name and address 2 (no exact, manually changed 19 mistaken matches, manually changes any match where the practice name was "the surgery", no mistakes above score=86%)	49
Reclink	1	Practice name and address 2 (no exact, manually changed 54 mistaken matches, manually changed 4 genuine matches with low scores, manually changed any match where the practice name was "the surgery", no mistakes above score=82%)	91
Reclink	1	Address 1, address 2 and address 3 (no exact, manually changed 15 mistaken matches, no mistakes above score=71%)	222
Reclink	1	Non-unique postcodes and names (no exact, manually changed 31 postcodes and practices names that were not precise enough)	74
Reclink	1	Non-unique postcodes, practice names and address 1 (no exact, manually changed 44 postcodes and practices names that were not precise enough)	27
Reclink	1	Postcode and address 1 [when practice name is renamed to address 1] (no exact, , manually changed 13 mistaken matches)	103
Reclink	1	Postcode, address 1 and address 2 [when practice name is renamed to address 1] (no exact, manually changed 2 mistaken matches, manually changed 5 genuine matches with low scores)	90
Reclink	1	Practice name, address 1 and address 2 [all text made lower case] (no exact, , manually changed 10 mistaken matches)	504
Reclink	1	Practice name and address 1 [all text made lower case] (no exact, manually changed 10 mistaken matches, manually changed any match where the practice name was "the surgery", manually changed 18 mistaken matches)	59
Excel	1	Matches made in Excel (fully manual)	842
Reclink	2	Postcode and practice address 1(29 manual changes and all over 80.1% score taken)	196
Reclink	2	Postcode and practice name (7 manual changes)	7
Excel	2	Manual matches to unique practices	274
Excel	2	Manual matches to non-unique practices	277
Imputation	3	Across couples	665
Imputation	3	Across waves	687
Correction	4	Removed matches due to moving home	-80
		Total	7,694

Table 7.6: Details of the matching procedure and results for the 2008 wave of ELSA

Matching method	*Stage*	*Matching details*	*Number matched*
Reclink	1	Postcode only (no manual matches, all exact)	3739
Reclink	1	Postcode and practice name (no exact matched, all taken above score= 93%)	581
Reclink	1	Practice name, address 1 and postcode (no exact, manually checked for mistaken matches)	91
Reclink	1	Address 1 and postcode (no exact, , manually checked for mistaken matches)	18
Reclink	1	Practice name and address 1 (no exact, manually changed 17 mistaken matches, no mistaken matches found above score=89%)	41
Reclink	1	Postcode and practice name (repeated 2nd reclink, no exact matches, manually changed 93 mistaken matches, no mistaken matches found above score=85%)	157
Reclink	1	Practice name and address 2 (no exact, manually changed 10 mistaken matches, manually changed 7 genuine matches with low scores, no mistakes above score=93%)	15
Reclink	1	Non-unique postcodes, practice names (no exact, manually changed 48 mistaken matches)	271
Reclink	1	Non-unique postcodes, practice names and address 1 (no exact, manually changed 8 genuine matches with low scores)	8
Reclink	1	Non-unique Practice name, address 1 and address 2 (no exact, no mistakes above score=90%)	2
Reclink	1	Practice name and address 1 [all text made lower case] (no exact, manually changed 127 mistaken matches)	580
Excel	1	Matches made in Excel (fully manual)	1071
Reclink	2	Postcode and practice address 1(18 manual changes and all over 88.9% score taken)	245
Reclink	2	Postcode and practice name (3 manual changes)	3
Excel	2	Manual matches to unique practices	171
Excel	2	Manual matches to non-unique practices	145
Imputation	3	Across couples	787
Imputation	3	Across waves	1031
Correction	4	Removed matches due to moving home	-128
		Total	8,828

Table 7.7: Statistics on the number of ELSA respondents matched to practices and the number of practices with an ELSA respondent

	2004 wave	*2006 wave*	*2008 wave*
Total interviewed	9,432 (100%)	9,779 (100%)	11,050 (100%)
Nurse Visit	7,666 (81.3%)	None	8,643 (78.2%)
Provided GP address	7,332 (77.7%)	None	8,138 (73.6)
Matched GP address	7,694 (81.5%)	6,680 (68.3%)	8,828 (75%)
Total Practices in QOF	8,921 (100%)	8,887 (100%)	8,961 (100%)
Practices with at least one matched ELSA respondent	3,082 (34.5%)	2,806 (31.6%)	3,439 (38.4%)
Respondents per practice:			
Mean	2.62	2.48	2.68
Std. Dev.	2.24	2.04	2.69
Min	1	1	1
Max	20	19	18

Notes: wave 3 of ELSA did note include a nurse visit to collect GP information, matches were imputed

Table 7.8: OLS regression of the probability of an ELSA respondent being matched to a practice alongside the descriptive statistics for matched and unmatched respondents

	Coefficients		*Mean Values*		
			Matched	Unmatched	Total
Age	0.0346***	(0.00377)	65.05	63.71	64.75
Age squared	-0.000231***	(0.0000284)	4330	4182	4297
Gender (Male=1)	0.000534	(0.00665)	0.44	0.43	0.44
Same house as last interview	0.249***	(0.0122)	0.96	0.89	0.95
Excellent health	0.0109	(0.00901)	0.17	0.18	0.17
Very good health	0.00737	(0.00647)	0.33	0.33	0.33
Good health			0.3	0.27	0.29
Fair health	-0.0220**	(0.00809)	0.15	0.15	0.15
Poor health	-0.0707***	(0.0135)	0.05	0.07	0.06
No long-standing illness			0.45	0.47	0.45
Non-limiting Long-standing illness	0.0240***	(0.0157)	0.22	0.19	0.21
Limiting Long-standing illness	0.0250**	(0.0168)	0.33	0.35	0.34
Not married or cohabiting			0.68	0.62	0.66
Cohabiting couple	0.0548***	(0.0156)	0.05	0.05	0.05
Married couple	0.0465***	(0.00785)	0.28	0.33	0.29
Memory function 1 (lowest)	-0.0592***	(0.0117)	0.09	0.13	0.1
Memory function 2	-0.0118	(0.00789)	0.17	0.18	0.17
Memory function 3			0.25	0.24	0.25
Memory function 4	0.0163*	(0.00686)	0.28	0.26	0.28
Memory function 5 (highest)	0.0343***	(0.00801)	0.21	0.19	0.21
Executive function 1 (lowest)	-0.0333***	(0.00924)	0.18	0.23	0.19
Executive function 2	0.00589	(0.00871)	0.14	0.14	0.14
Executive function 3			0.19	0.18	0.19
Executive function 4	0.0125	(0.00782)	0.21	0.18	0.2
Executive function 5	0.00606	(0.00865)	0.18	0.17	0.17
Executive function 6 (highest)	0.00899	(0.0108)	0.1	0.1	0.1
2004	0.0148***	(0.00431)			
2006	-0.115***	(0.00508)			
Constant	-0.721***	(0.124)			
Observations	28963		22433	6530	28963
R^2	0.06				

Standard errors in parentheses
*p< 0.05, **p< 0.01, ***p< 0.001

Table 7.9: OLS regression of the probability of a practice being included alongside the descriptive statistics for included and not included practices

	Coefficients		*Mean Values*		
			Included	Not included	Total
Rural/urban practice, ONS 07	-0.0829***	(0.0145)	0.15	0.15	0.15
Total QOF points achieved (%)	0.192***	(0.0490)	0.96	0.94	0.95
Low Income Scheme Index	-0.00317***	(0.000570)	10.34	13.08	12.05
PMS practice	-0.0245**	(0.00926)	0.39	0.41	0.4
Practice population size	-0.0000109**	(0.00000347)	8028.93	5785.53	6623.38
Number of GPs at practice	-0.000612	(0.00360)	5.02	3.6	4.13
Patients of ELSA age (45+)	0.000106***	(0.00000684)	3359.1	2216.66	2643.33
2004	-0.0286***	(0.00519)			
2006	-0.0613***	(0.00417)			
Constant	0.0789	(0.0480)			
Observations	23118		8634	14484	23118
R^2	0.11				

Standard errors in parentheses
*p< 0.05, **p< 0.01, ***p< 0.001

Table 7.10: Descriptive statistics on the quality of care reported by patients and by practices over time

Year	*2004*	*2004*	*2008*	*2004*	*2006*	*2008*
ELSA	Mean			N		
Insulin (diabetes)	0.22	0.2	0.2	576	575	790
Taking medication (diabetes)	0.62	0.67	0.69	576	575	790
Taking ACE inhibitor (diabetes)	0.46	0.5	0.53	576	584	790
Taking medication (Hypertension)	0.8	0.86	0.89	1381	2208	3027
Checked for protein (diabetes)	0.82	0.79	0.8	283	265	339
A1c test (diabetes)	0.86	0.86	0.82	567	562	761
Alc test within 12 months (diabetes)	0.93	0.94	0.92	489	484	624
Feet check (diabetes)	0.84	0.84	0.84	567	562	761
Blood pressure test (Hypertension)	0.97	0.98	0.97	765	1256	2944
Discussed ACE inhibitor (diabetes)	0.24	0.25	0.26	567	562	761
Protein in urine (diabetes)	0.18	0.18	0.16	231	210	270
Kidney trouble (diabetes)	0.14	0.15	0.17	567	562	761
Received training (diabetes)	0.27	0.24	0.25	567	562	761
Knowledge (diabetes)	0.8	0.81	0.83	567	559	757
Suggested medications (Hypertension)	0.85	0.83	0.87	1366	2710	3263
QOF diabetes A1c test	0.93	0.94	0.95	3073	2786	3403
QOF diabetes feet check	0.75	0.85	0.86	3073	2786	3403
QOF diabetes ACE inhibitor	0.06	0.09	0.09	3073	2786	3403
QOF diabetes combined	0.74	0.79	0.8	3073	2774	3401
QOF hypertension blood pressure test	0.9	0.92	0.91	3073	2787	3403
QOF hypertension combined	0.88	0.83	0.83	3073	2775	3401

Note: ELSA quality of care measures are sorted by the type of care measured (medication, tests, discussion with doctor)

Table 7.11: OLS regressions of four ELSA indicator performance on practice indicator performance

ELSA indicator performance	*No constant*	*Constant*	*Interview dummies*	*Indicator dummies*	*Interaction terms*
QOF indicator performance	1.042*** (0.003)	0.520*** (0.016)	0.520*** (0.017)	0.253*** (0.066)	
A1c				0.209*** (0.059)	0.370* (0.185)
Feet				0.149** (0.053)	0.180* (0.079)
Blood pressure				0.259*** (0.057)	0.402*** (0.060)
QOF*ACE					0.451** (0.167)
QOF*A1c					0.099 (0.196)
QOF*Feet					0.234** (0.090)
QOF*Blood pressure					0.112 (0.061)
Constant		0.465*** (0.015)	0.463*** (0.015)	0.481*** (0.015)	0.466*** (0.019)
Observations	10327	10327	10327	10327	10327
R^2	0.845				
R^2		0.235	0.235	0.247	0.248

Standard errors in parentheses (clustered by practice)
*$p< 0.05$, **$p< 0.01$, ***$p< 0.001$
Notes: ACE inhibitor is the base category
Models presented are variations of Equation 7.1 and 7.2

Table 7.12: OLS regressions of four ELSA indicator performance on practice indicator performance

ELSA indicator performance	*interaction terms*	*person controls*	*household controls*	*health controls*	*function controls*	*practice controls*
QOF*ACE	0.451**	0.429*	0.435**	0.463**	0.497**	0.505**
	(0.167)	(0.168)	(0.167)	(0.168)	(0.169)	(0.169)
QOF*A1c	0.099	0.107	0.101	0.081	0.079	0.076
	(0.196)	(0.198)	(0.196)	(0.196)	(0.197)	(0.213)
QOF*Feet	0.234**	0.212*	0.211*	0.208*	0.208*	0.209*
	(0.090)	(0.089)	(0.090)	(0.090)	(0.090)	(0.094)
QOF*Blood pressure	0.112	0.110	0.097	0.106	0.103	0.125
	(0.061)	(0.060)	(0.057)	(0.057)	(0.057)	(0.068)
A1c	0.370*	0.359	0.365*	0.387*	0.390*	0.394
	(0.185)	(0.188)	(0.186)	(0.186)	(0.187)	(0.202)
Feet	0.180*	0.197*	0.198*	0.204**	0.205**	0.206*
	(0.079)	(0.078)	(0.079)	(0.079)	(0.079)	(0.082)
Blood pressure	0.402***	0.402***	0.415***	0.418***	0.421***	0.406***
	(0.060)	(0.059)	(0.056)	(0.056)	(0.056)	(0.065)
2nd observation	0.005	0.004	0.005	0.008	0.009	0.009
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
3rd observation	-0.004	-0.006	-0.006	-0.004	-0.005	-0.006
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Age		0.012*	0.010	0.011*	0.009	0.011*
		(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Age squared		-0.000*	-0.000	-0.000	-0.000	-0.000
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Gender (Male=1)		0.017*	0.011	0.012	0.013	0.014
		(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Changed GP since last interview			-0.012	-0.010	-0.011	-0.012
			(0.018)	(0.018)	(0.017)	(0.018)
Married couple			0.028**	0.028**	0.027**	0.025**
			(0.009)	(0.009)	(0.009)	(0.009)
Cohabiting couple			-0.025	-0.024	-0.023	-0.022
			(0.024)	(0.024)	(0.024)	(0.024)
Excellent health				-0.009	-0.011	-0.010
				(0.013)	(0.013)	(0.013)
Very good health				-0.003	-0.004	-0.004
				(0.008)	(0.008)	(0.008)
Fair health				0.001	0.004	0.006
				(0.009)	(0.009)	(0.009)
Poor health				-0.001	0.005	0.008
				(0.013)	(0.013)	(0.013)
No long-standing illness				-0.036***	-0.035***	-0.035***
				(0.009)	(0.009)	(0.009)
Memory function 1 (lowest)					-0.028*	-0.027*
					(0.013)	(0.013)
Memory function 2					0.001	0.002
					(0.010)	(0.010)
Memory function 4					0.011	0.012
					(0.009)	(0.009)
Memory function 5 (highest)					0.012	0.009
					(0.011)	(0.011)
Executive function 1 (lowest)					-0.006	-0.006
					(0.011)	(0.011)
Executive function 2					-0.004	-0.002
					(0.010)	(0.010)
Executive function 4					-0.011	-0.012
					(0.011)	(0.011)
Executive function 5					-0.001	-0.000

					(0.012)	(0.012)
Executive function 6 (highest)					0.023 (0.015)	0.024 (0.015)
Rural Practice						0.006 (0.011)
Deprivation (LISI)						-0.001 (0.001)
PMS contract						-0.001 (0.008)
Disease register						-0.000 (0.000)
Constant	0.466*** (0.019)	0.033 (0.185)	0.060 (0.185)	0.050 (0.184)	0.077 (0.186)	0.036 (0.189)
Observations	10327	10237	10237	10181	10140	10025
R^2	0.248	0.249	0.250	0.252	0.253	0.254

Standard errors in parentheses (clustered by practice)
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Notes: ACE inhibitor is the base category
Models presented are variations of Equation 7.2 and 7.3

Table 7.13: First difference regressions of four ELSA indicator performance on practice indicator performance

ELSA indicator performance	*Combined indicators*	*Separate indicators*	*Separate indicators with all controls*
QOF performance	0.244 (0.126)		
QOF*ACE		1.098** (0.342)	1.172*** (0.354)
QOF*A1c		-0.551 (0.477)	-0.620 (0.484)
QOF*Feet		0.130 (0.155)	0.077 (0.162)
QOF*Blood pressure		0.123 (0.131)	0.179 (0.181)
Constant	0.009 (0.006)	0.007 (0.006)	0.007 (0.030)
Observations	3694	3694	3549
R^2	0.008	0.007	0.014

Standard errors in parentheses (clustered by practice)
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Notes: control variables are those from Table 7.12 [p.204]
Models presented are variations of Equation 7.4

Table 7.14: OLS regressions of four ELSA indicator performance on practice disease performance

ELSA indicator performance	*No constant*	*Constant*	*Interview dummies*	*Indicator dummies*	*Interaction terms*
QOF disease performance	1.055*** (0.005)	1.140*** (0.072)	1.122*** (0.072)	0.235** (0.079)	
A1c				0.429*** (0.015)	0.900*** (0.199)
Feet				0.341*** (0.016)	0.428* (0.213)
Blood pressure				0.457*** (0.015)	0.869*** (0.199)
Diabetes QOF*ACE					0.572* (0.241)
Diabetes QOF*A1c					-0.030 (0.127)
Diabetes QOF*Feet					0.461* (0.186)
Hypertension QOF*Blood pressure					0.058 (0.056)
Observations	10320	10320	10320	10320	10320
R^2	0.857	0.034	0.034	0.246	0.247

Standard errors in parentheses (clustered by practice)
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Notes: ACE inhibitor is the base category
Models presented are variations of Equation 7.1 and 7.2

Table 7.15: OLS regressions of four ELSA indicator performance on practice disease performance

ELSA indicator performance	*interaction terms*	*person controls*	*household controls*	*health controls*	*function controls*	*practice controls*
Diabetes QOF*ACE	0.572* (0.241)	0.553* (0.244)	0.553* (0.243)	0.580* (0.245)	0.583* (0.244)	0.629* (0.250)
Diabetes QOF*A1c	-0.030 (0.127)	-0.027 (0.130)	-0.025 (0.129)	-0.037 (0.129)	-0.041 (0.130)	-0.048 (0.138)
Diabetes QOF*Feet	0.461* (0.186)	0.397* (0.184)	0.398* (0.184)	0.385* (0.184)	0.380* (0.184)	0.338 (0.191)
Hypertension QOF* Blood pressure	0.058 (0.056)	0.047 (0.057)	0.027 (0.057)	0.039 (0.057)	0.031 (0.057)	0.013 (0.058)
Constant	0.053 (0.188)	-0.386 (0.267)	-0.359 (0.267)	-0.391 (0.267)	-0.367 (0.268)	-0.443 (0.273)
Observations	10320	10230	10230	10174	10133	10018
R^2	0.247	0.248	0.250	0.251	0.252	0.253

Standard errors in parentheses (clustered by practice)
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Notes: control variables are those from Table 7.12 [p.204]
Models presented are variations of Equation 7.2 and 7.3

Table 7.16: First difference regressions of four ELSA indicator performance on practice disease performance

ELSA indicator performance	*Combined indicators*	*Separate indicators*	*Separate indicators with all controls*
QOF performance	0.007		
	(0.139)		
QOF*ACE		0.435	0.195
		(0.377)	(0.411)
QOF*A1c		-0.340	-0.552
		(0.256)	(0.281)
QOF*Feet		0.061	-0.223
		(0.312)	(0.320)
QOF*Blood pressure		-0.151	-0.030
		(0.149)	(0.184)
Constant	0.013*	0.011	0.019
	(0.006)	(0.007)	(0.032)
Observations	3686	3686	3541
R^2	0.013	0.011	0.014

Standard errors in parentheses (clustered by practice)
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Notes: control variables are those from Table 7.12 [p.204]
Models presented are variations of Equation 7.4

Table 7.17: OLS and first difference regressions of 15 ELSA indicator performance on practice disease performance

ELSA indicator performance	*Interaction terms*	*All controls*	*First difference with all controls*
Diabetes QOF*Insulin	-0.116	-0.137	0.081
	(0.209)	(0.218)	(0.249)
Diabetes QOF*Taking medication (diabetes)	0.285	0.306	0.576*
	(0.236)	(0.246)	(0.249)
Diabetes QOF*Taking ACE inhibitor	0.517*	0.600*	0.306
	(0.240)	(0.250)	(0.249)
Hypertension QOF*Taking medication (Hypertension)	-0.051	-0.027	-0.099
	(0.102)	(0.103)	(0.172)
Diabetes QOF*Checked for protein	-0.249	-0.230	-1.259**
	(0.228)	(0.237)	(0.427)
Diabetes QOF*A1c test	0.096	0.126	-0.180
	(0.171)	(0.180)	(0.249)
Diabetes QOF*Alc test (12 months)	-0.086	-0.076	-0.446
	(0.126)	(0.137)	(0.285)
Diabetes QOF*Feet check	0.406*	0.311	-0.111
	(0.185)	(0.188)	(0.249)
Hypertension QOF*Blood pressure	0.092	0.061	-0.123
	(0.056)	(0.060)	(0.224)
Diabetes QOF*Discussed ACE inhibitor	0.339	0.294	0.259
	(0.191)	(0.199)	(0.249)
Diabetes QOF*Protein in urine	-0.221	-0.295	0.001
	(0.267)	(0.285)	(0.511)
Diabetes QOF*Kidney trouble	0.181	0.141	-0.139
	(0.173)	(0.186)	(0.249)
Diabetes QOF*Received training	-0.085	-0.111	-0.865***
	(0.213)	(0.225)	(0.249)
Diabetes QOF*Knowledge	0.311	0.309	0.269
	(0.189)	(0.197)	(0.251)
Hypertension QOF*Suggested medications	0.081	0.093	-0.083
	(0.101)	(0.102)	(0.152)
Constant	0.987***	0.137	-0.004
	(0.098)	(0.189)	(0.014)
Observations	38965	37732	14899
R^2	0.364	0.373	0.007

Standard errors in parentheses (clustered by practice)
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Notes: control variables are those from Table 7.12 [p.204]
Models presented are variations of Equation 7.2, 7.3 and 7.4

FIGURES

Figure 3.1: The relationship between the percent of treated patients for a QOF clinical indicator and the number of QOF points awarded

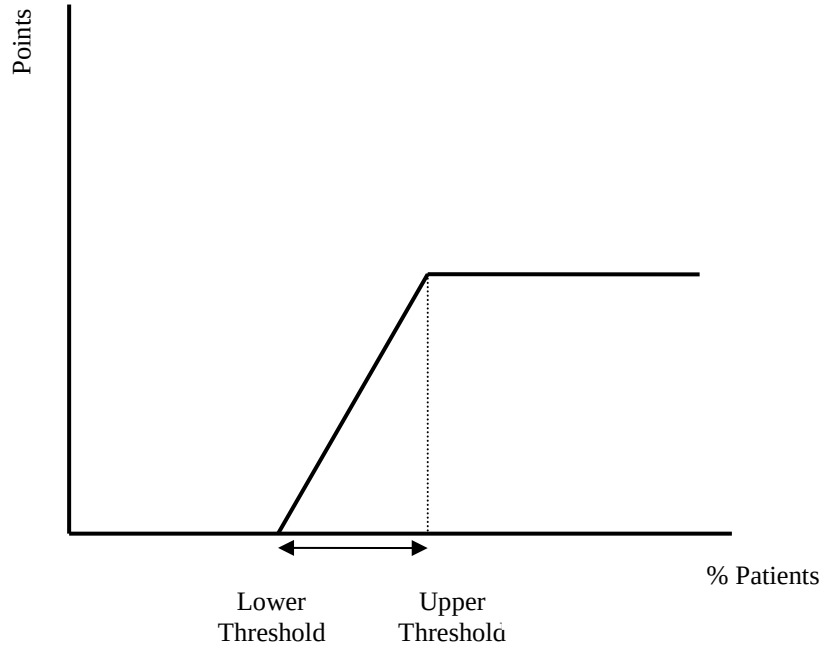
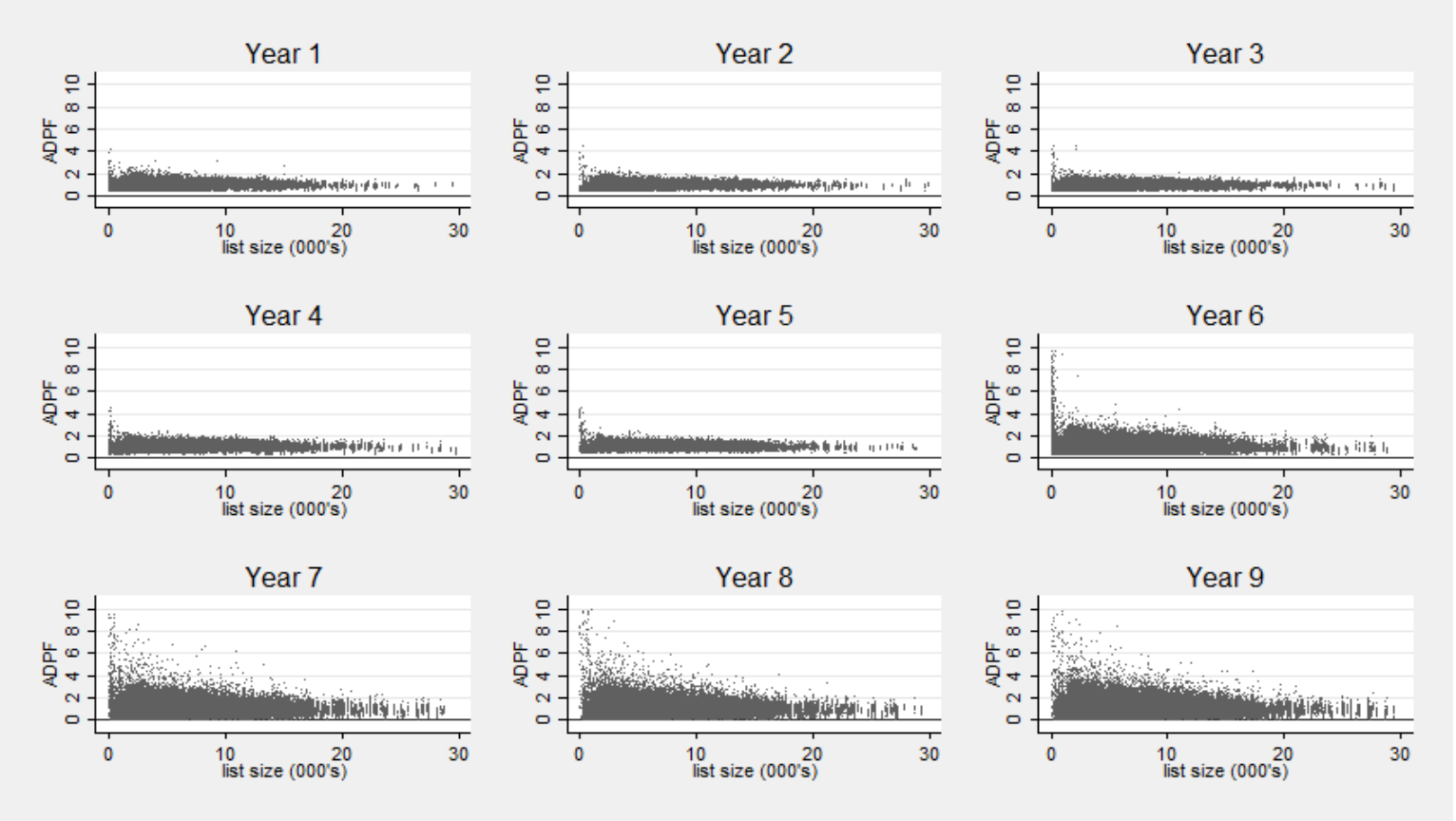
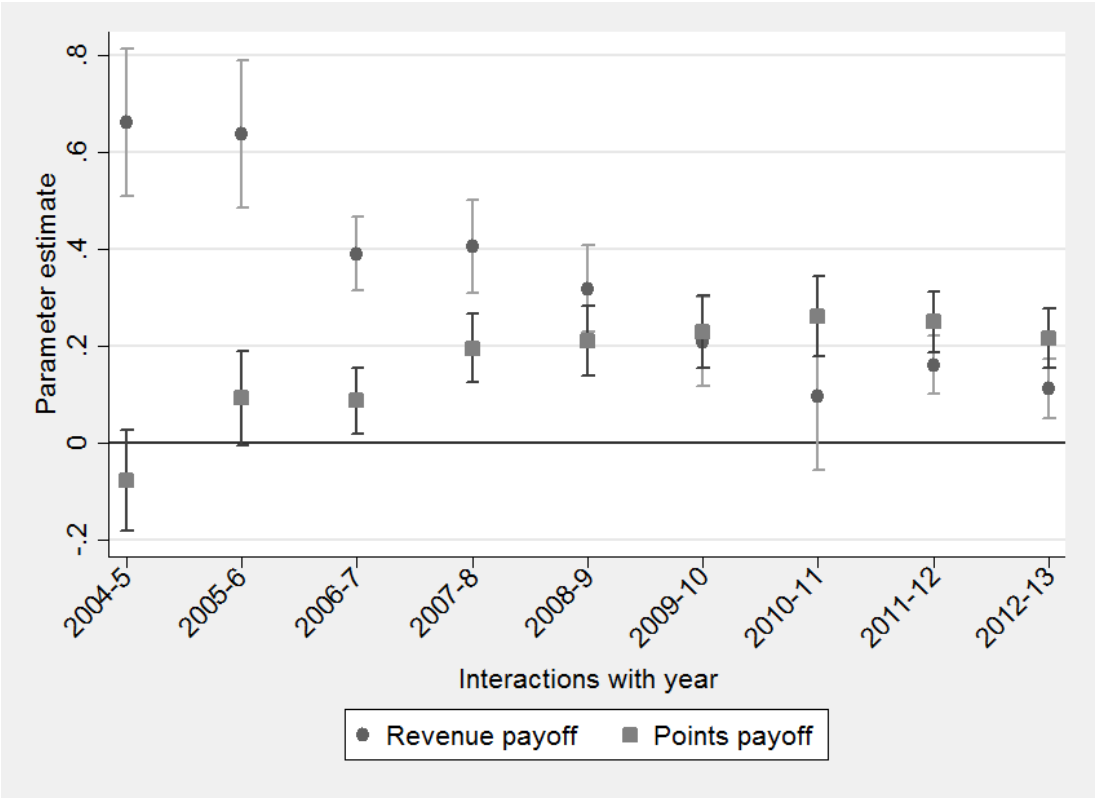


Figure 4.1: Annual scatter plot of ADPF and list size for 2004/5 to 2012/13



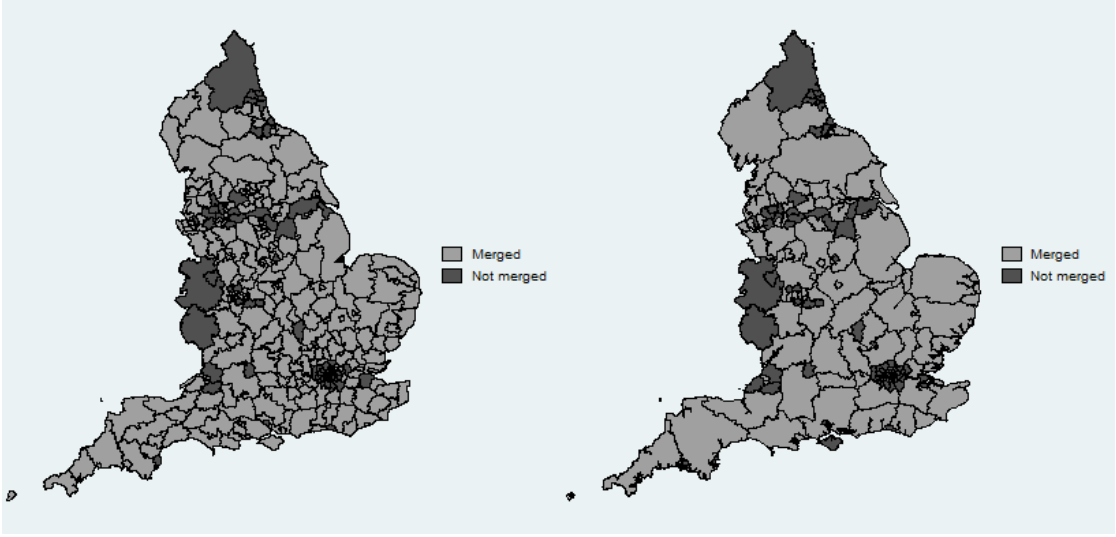
Note: each dot represents a practice-disease group. The sample is restricted to list sizes less than 30,000 and ADPFs less than 10 as the outliers skewed the plots. In year 6 (2009/10) the square rooting of prevalence rates was removed. In year 7 (2010/11) the 5% truncation was removed. Line at y=0 to show the effect of ADPF truncation in years 1-6.

Figure 4.2: Estimated coefficients and 95% CIs for revenue and reputation incentives



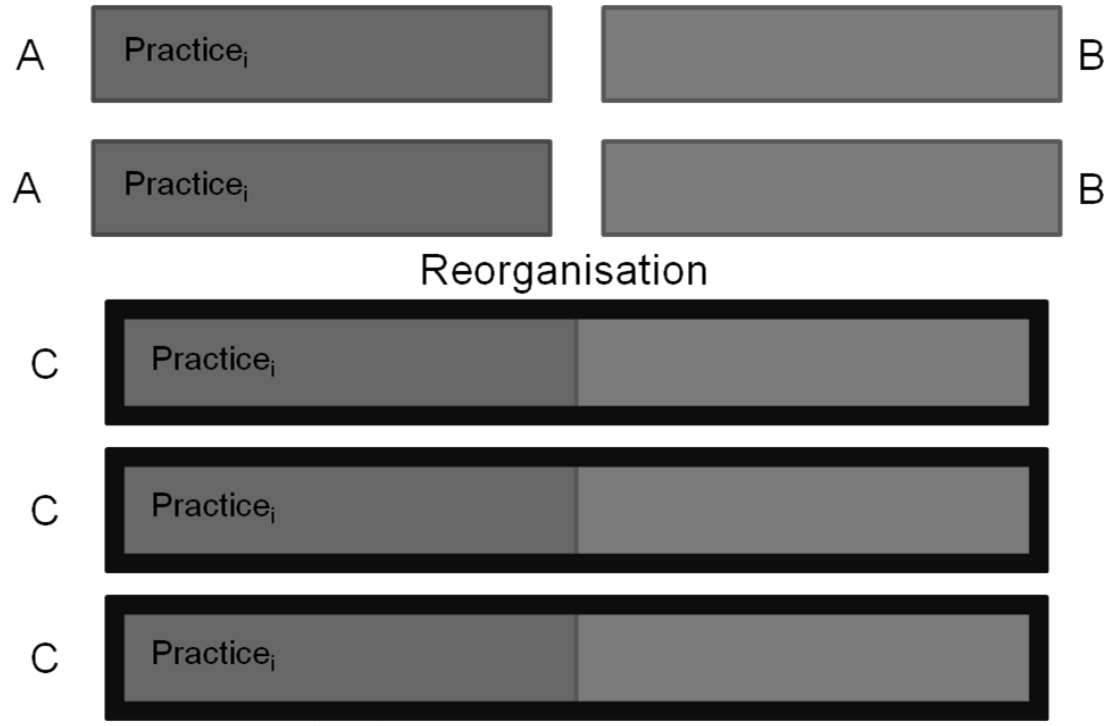
Note: coefficients are taken from incentives interacted with year dummies from model 1 in Table 4.6, Equation 4.4

Figure 5.1: PCT configuration for England before and after the PCT reorganisation



Note: left map shows the structure before the reorganisation. Right map shows the structure after the reorganisation

Figure 5.2: Structure of data on practice PCT peer groups 2004/5 to 2008/9



Note: each row represents a year of data, five in total. Practice$_i$ was in PCT A during the pre-reorganisation years and merged with B to form PCT C in the post-reorganisation period. In each year we were able to observe the peers in all three relevant groups.

Figure 6.1: Locally weighted regression plot of GP P4P exposure on practice P4P exposure for 2005 and 2008

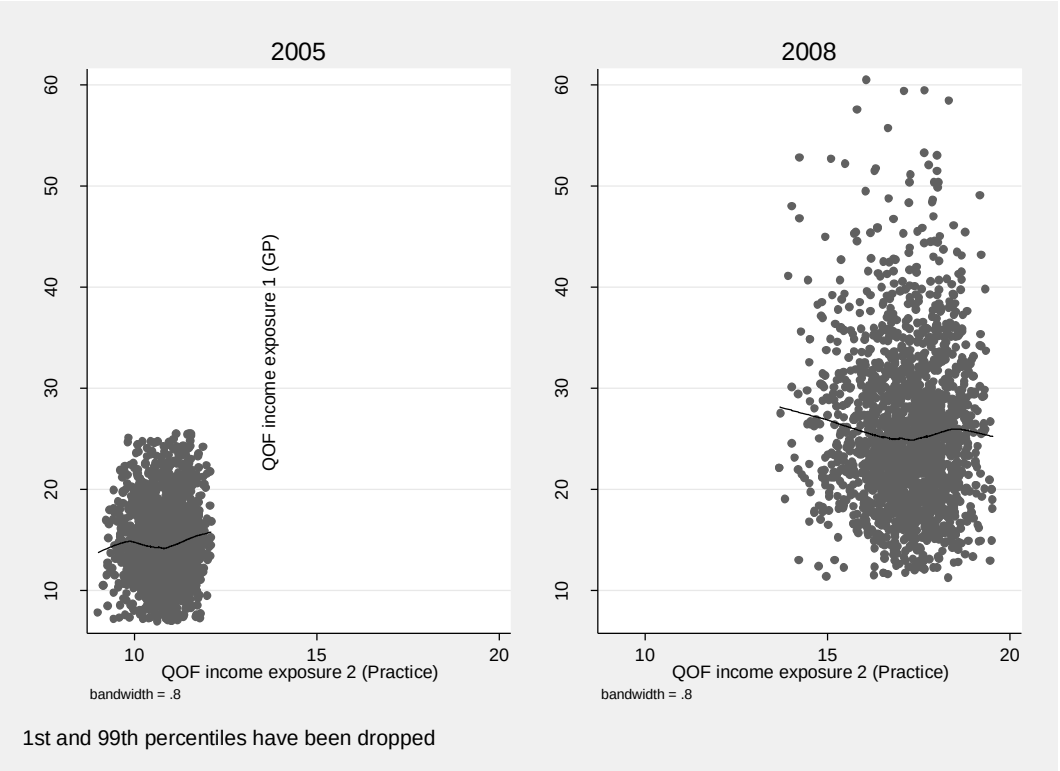


Figure 6.2: Locally weighted regression plot of GP and practice P4P exposure on changes in job satisfaction between 2004 and 2005

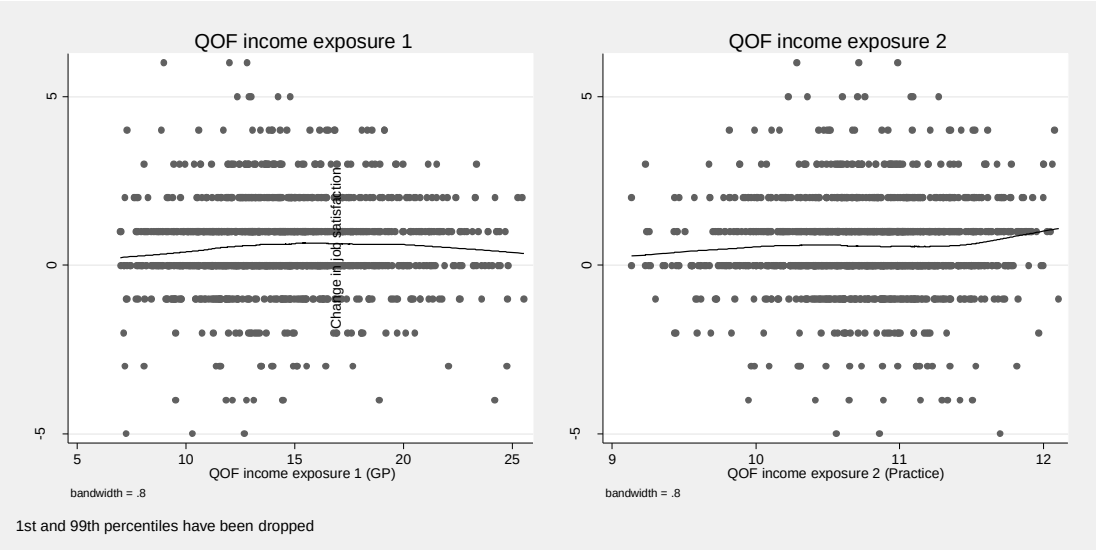


Figure 6.3: Locally weighted regression plot of GP and practice P4P exposure on changes in job satisfaction between 2004 and 2008

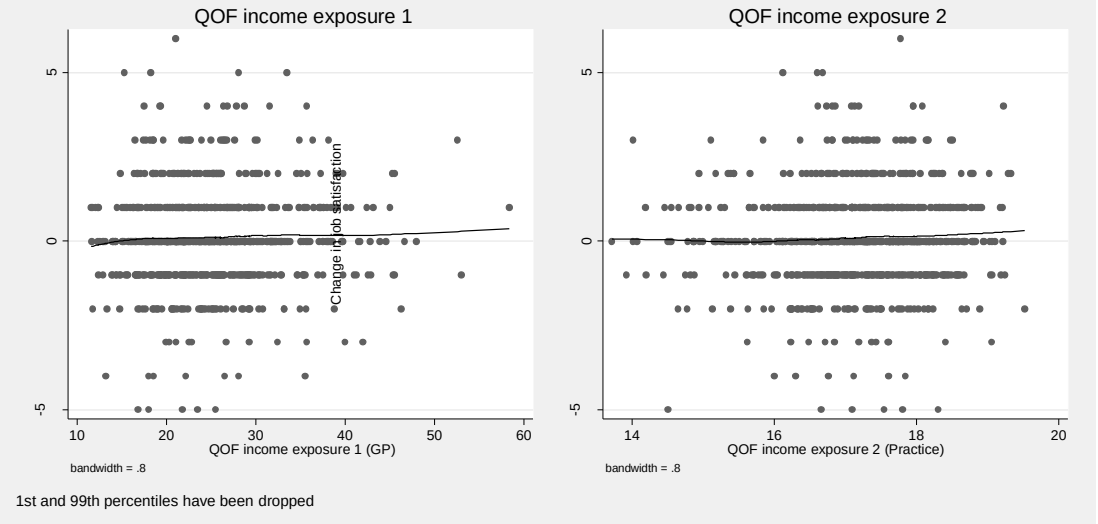


Figure 7.1: Predicted values for interaction terms measuring the association between patient-reported and practice-reported quality of care from Table 7.11 [p.203]

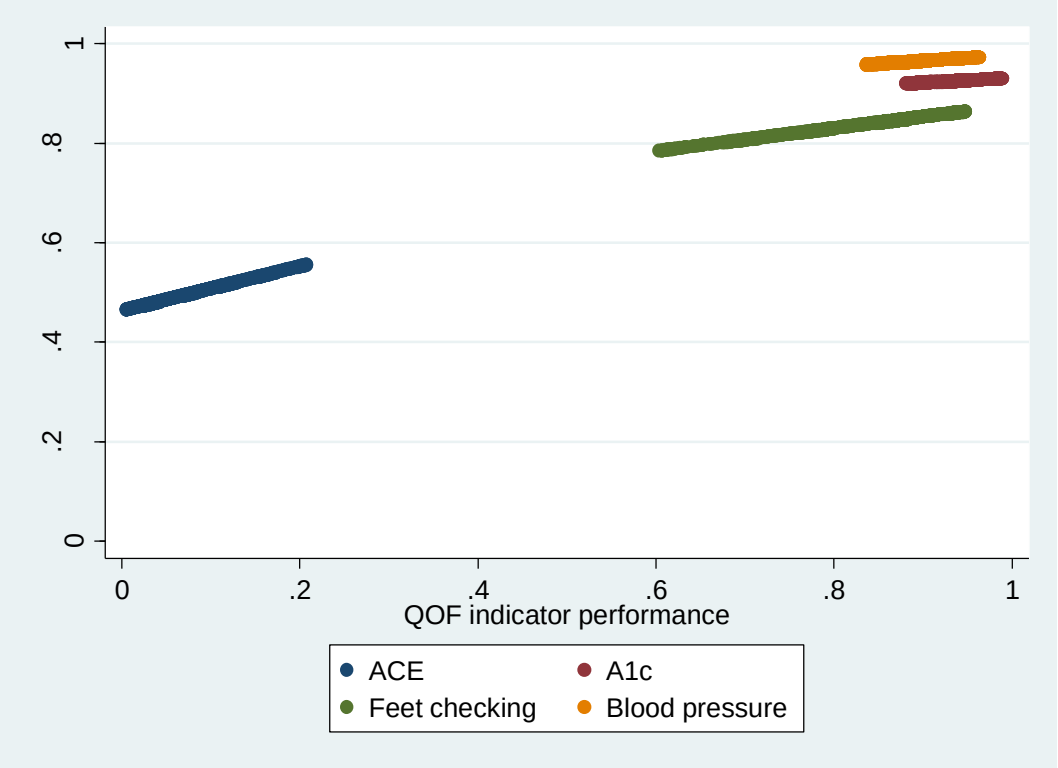


Figure 7.2: Predicted values for interaction terms measuring the association between patient-reported and practice-reported quality of care from Table 7.14 [p.206]

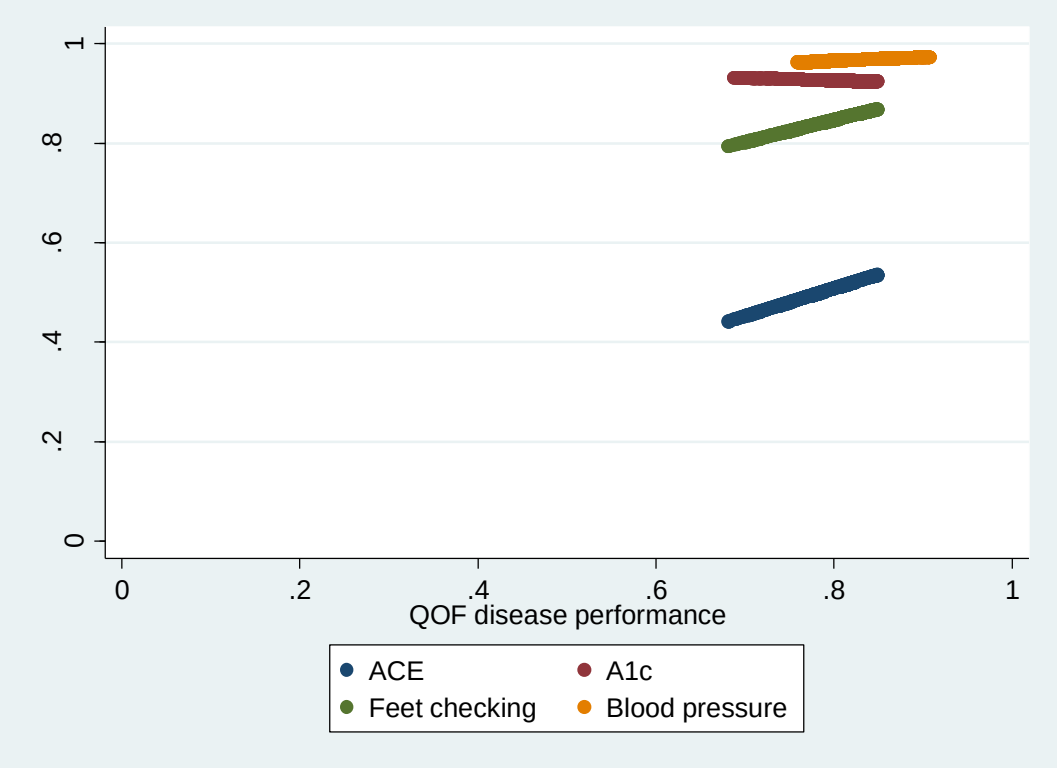


Figure 7.3: Predicted values for interaction terms measuring the association between patient-reported and practice-reported quality of care from Table 7.17 [p.208]