IFS

Institute for Fiscal Studies

Paul Rodríguez-Lesmes
Marcos Vera-Hernández

Working paper

25/08

Multitasking, two-part contracts, and bunching: an application to doctors' tasks

Economic
and Social
Research Council

Multitasking, Two-part Contracts, and Bunching: an Application to Doctors' Tasks [*]

Paul Rodríguez-Lesmes[†], Marcos Vera-Hernández[‡]

February 24, 2025

Abstract

The optimal design of incentive contracts critically depends on whether the tasks performed by agents are complementary or substitutable, yet empirical evidence on this remains limited. This paper develops a novel empirical strategy to identify complementarities and substitutabilities in tasks, even in the absence of contract variation across agents, provided the incentive contract is piecewise linear. We apply this method to data on the management of chronic diseases by UK family physicians and find evidence that some tasks are complements, while none are substitutes. These complementarities may explain the widespread adoption of incentive contracts in healthcare. Furthermore, our findings suggest that healthcare systems centred around family physicians, rather than specialists, could achieve significant efficiency gains by consolidating complementary tasks under a single provider.

JEL: I11, I18, M52

[*] We would like to thank Juan Pablo Atal, David Bardey, Pedro Pita Barros, Helmuth Cremer, Pierre Dubois, Jonathan Kolstad, Albert Ma, Suphanit Piyapromdee, Carol Propper, Rudi Rocha, Luigi Siciliani and participants at the 1st Toulouse Health Economics Conference, the 18th European Health Theory Workshop, iHEA World Meeting, Barcelona GSE Summer Forum, Latin American Health Economics Network workshop, LACEA-LAMES annual meeting, Institute for Fiscal Studies, and at department seminars at universities del Rosario, de los Andes, Javeriana, Warwick, Rennes, La Laguna, Munich, Oslo, Oxford, Lausanne, Kent, and Barcelona. Any errors are our own. Rodríguez-Lesmes acknowledges financial support from The World Bank under the program Fulbright - Colciencias and Colombia Cientifica - Alianza EFI # 60185 contract # FP44842 - 220 - 2018.

[†] E-mail: paul.rodriguez@urosario.edu.co. Universidad del Rosario.

[‡] E-mail: m.vera@ucl.ac.uk. University College London, Institute for Fiscal Studies, and Centre for Economic Policy Research.

1 Introduction

Principal-agent relationships are widespread in economics. Since decades ago, it is well-known that agents might shirk if the effort is not verifiable, and that economic incentives might be needed to achieve the second-best (Stiglitz, 1974; Harris and Raviv, 1979; Hölmstrom, 1979). However, there are also important potential limitations to the use of economic incentives such as crowding-out of intrinsic motivation (Deci and Ryan, 1985; Gneezy and Rustichini, 2000; Fehr and Falk, 2002; Bénabou and Tirole, 2003, 2006), imperfect knowledge of the production function by either the principal or agent (Prendergast, 2002; Mohanan et al., 2021) and the fact that many agents work on more than one task, the so-called "multitasking" problem (Hölmstrom and Milgrom, 1991). Our understanding of the practical importance of these possible limitations in real-life settings lags well behind the theoretical advances to knowledge.

Our paper addresses the "multitasking" problem, which extends beyond the common depiction of rewarding one task when another goes unrewarded. The issue is more general: if increasing effort on one task raises the marginal cost of effort on other tasks ("substitute tasks"), incentive schemes can become very costly when principals reward multiple tasks, potentially making it optimal for principals to forgo incentives altogether (Hölmstrom and Milgrom, 1991). Conversely, if tasks are "complements" (where increasing effort in one task reduces the marginal cost of other tasks), rewarding just one or a few tasks can increase output across all complementary tasks. Often overlooked in the literature, whether tasks are complements or substitutes also affects job design, as it is efficient to assign complementary tasks to the same agents (Hölmstrom and Milgrom, 1991)

The objective of this paper is to develop an empirical strategy to estimate whether tasks are complements or substitutes when agents face a piece-wise linear contract, essentially a contract with two or more different piece-rate levels. Our identification strategy combines the kinks embedded in these piece-wise linear contracts, with changes over time to such contracts. When agents face such contracts, some agents will be at the kink and others not. The effort choices of those at the kink are insensitive to small changes in the marginal benefit or marginal cost. This creates a "less-sensitive" group, which we exploit to identify the substitution/complementary pattern across tasks.

Our proposed strategy has three desirable features. First, we do not require variation in the contracts across the cross-section of agents. Although we require that the incentive rates change over time, these changes can be the same for the entire population of agents. Second, we are not restricted to assessing substitution patterns between rewarded and unrewarded tasks but can also estimate such patterns amongst rewarded tasks, and third, we do not require data on agents' performance before the introduction of the incentive contract when such data might not be available (although we need it before and after the change in the incentive rate). In turn, this allows us to assess the complementary/substitution pattern well after the programme has been implemented, allowing for capital and labour to adjust, and hence recovering steady-state estimates. This is potentially important because short-term capacity constraints could render tasks to be substitutes, but they might be complements once these constraints are relaxed.[1]

We apply our proposed strategy to identify whether different tasks that family doctors perform are complements or substitutes in their cost function. Examples of the tasks that we analyse include carrying out certain tests on diabetic patients, recording smoking history among *at-risk* patients, or reviewing asthmatic patients with some minimum frequency, among others. We exploit the *Quality Outcomes Framework* (QOF), the largest primary care pay-for-performance programme worldwide (Roland and Olesen, 2016), established in 2004 that remunerates all Primary Health Centres (PHCs) in England according to their performance in a battery of tasks (tests on diabetic patients, etc.) The programme was rolled out simultaneously across England, and any subsequent changes to the remuneration schedule also took effect simultaenously on all PHCs of the country.[2]

In 2011 we find that several tasks are complements and none are substitutes.

[1] Similarly, the introduction of incentive schemes might be accompanied by the introduction of new information systems leading to a change in the way agents' output is recorded.(Handel et al., 2020)

[2] The programme has been studied previously by comparing incentivised and unincentivised tasks before and after the introduction of the programme. Sutton et al. (2010) and Doran et al. (2011) found improvements in both incentivized and unincentivized tasks (which points in the direction of complementarities), which were higher for incentivized ones. Compared to this standard approach, ours takes advantage of the kinks in the incentive contracts, and hence we do not need to rely on any assumption of common trends across tasks, and we can estimate the effects long after the introduction of the programme.

It is important to note that we are measuring the effects seven years after QOF was implemented, and hence capital and labour have already adjusted, leading to the relaxation of short-term capacity constraints. Indeed, the number of nurses working on PHCs increased steadily until 2009, and then stabilised. This highlights the advantage of estimating the substitution patterns long after the pay-for-performance scheme was first implemented, and once the steady state has probably been reached.

Our paper contributes to the literatures on empirical contract theory, "bunching", and health care. We contribute to the empirical contract theory literature (see Chiappori and Salanié (2003); Chiappori (2000); Chiappori and Salanié (1997)) by proposing an empirical strategy that will expand the settings in which it is possible to estimate the complementary/substitution patterns (i.e. neither cross-section variation in the contracts nor pre-programme data are necessary), helping us to close the gap between the theoretical advances on contract theory and the available empirical evidence,[3] and indirectly contributing to our understanding on why we observe incentives schemes in some settings and not others.

Our paper also makes a contribution to the "bunching" literature (Saez, 2010; Kleven, 2016), which has used kinks in the payment function to measure the effect of economic incentives, especially in public finance. Although we use the agents "bunched" at the kinks in a different way to the bunching literature, we borrow from this literature the procedures required to test for bunching at the kinks.

We also contribute to the health care literature, and specifically to the optimal design of incentive contracts in health care, for which the pattern of complementary or substitution across tasks is of first order importance. Pay-for-performance schemes are ubiquitous in the health care system of high-income countries (i.e. Australia, Canada, Germany, Israel, New Zealand, Spain, Taiwan, United King-

[3] Besides studies on an educational setting (see Brickley and Zimmerman (2001); Jacob (2005); Atkinson et al. (2009); Neal (2011); De Philippis (2015)) most of the evidence comes from examining a very particular type of multitasking: the quality-quantity trade-off, which most of the evidence fails to find (Al-Ubaydli et al., 2015; Lazear, 2000; Paarsch and Shearer, 2000; Shearer, 2004; Kosfeld and Neckermann, 2011; Johnson et al., 2015; Bradler et al., 2016; Hong et al., 2018; Englmaier et al., 2016). Although the results are mixed (Finan et al., 2017), the extent of the evidence is broader in developing countries thanks to the proliferation of experiments that randomize agents across rewarded and unrewarded tasks (e.g. Glewwe et al. (2010); Sylvia et al. (2013); Olken et al. (2014); Celhay et al. (2019)).

dom, United States) and are increasingly implemented in low and middle-income countries (Chalkley et al., 2016; Eijkenaar, 2012).

We proceed as follows. Section 2 presents our theoretical model which links the complementarity/substitution between tasks to how output responds to changes in the incentives that agents face. Section 3 explains how we combine the kinks of the piece-wise linear contract with changes over time to such contract to estimate the complementarity/substitution patterns between tasks. Section 4 describes how we tailor our empirical strategy to the QOF. Section 5 reports our results, and their implications are discussed in section 6. Section 7 concludes. Several appendices provide proofs, further details on QOF, and additional results.

2 Model

We present a stylised model where agents have been hired to exert effort on two tasks: e_1 and e_2. The agent's cost function, $C(e_{1,}e_2; z)$, is characterised by an efficiency parameter z. Agents are identical, except for the efficiency parameter z, which is distributed in the population following a pdf $g(\cdot)$, with CDF $G(\cdot)$. Hence we assume that $\frac{\partial C}{\partial z} < 0$, as well as other standard assumptions on the cost function: $\frac{\partial C}{\partial e_i} = C_i > 0$, $\frac{\partial^2 C}{\partial e_{ii}^2} = C_{ii} > 0$ $(i \in 1, 2)$, and $C_{11}C_{22} - C_{12}^2 < 0$.

Concerning the cross-derivative $C_{12} = \frac{\partial^2 C}{\partial e_1 \partial e_2}$, we assume that its sign (positive or negative) is the same for all agents, independently of their value of z. The tasks are substitutes (complements) if increasing effort in one task, increases (reduces) the marginal cost of exerting effort on the other task. Our main goal in this paper is to estimate the sign C_{12} to ascertain whether the tasks are complements or substitutes (Hölmstrom and Milgrom, 1991).

The principal benefits increasingly from the output of the two tasks (x_1, x_2), where $x_1 = e_1 + \varepsilon_1$ and $x_2 = e_2 + \varepsilon_2$, and both ε_1 and ε_2 are independent random variables with zero mean and finite variance. The agent, with increasing and concave utility function $U(\cdot)$, is paid according to the contract $P(x_1, x_2)$, which is taken as given, and chooses effort levels e_1 and e_2 to maximises his expected utility, that is:

$$\begin{aligned} \max_{e_1,e_2} \quad & E_{\varepsilon_1,\varepsilon_2}\left[U\left(P(x_1,x_2)\right) - C(e_1,e_2;z)\right] \\ s.t. \quad & x_i = e_i + \varepsilon_i, \quad i = 1,2 \end{aligned} \tag{1}$$

2.1 Model without uncertainty

To derive the intuition behind our empirical strategy, we start by making a series of simplifying assumptions: (1) task 1 is paid according to a two-part linear contract, but task 2 is paid with respect to a linear contract, (2) effort is increasing in productivity $\frac{\partial e_1}{\partial z} > 0$, and (3) there is no uncertainty in either of the tasks ($x_1 = e_1$ and $x_2 = e_2$), which leads us to assume that $U(\cdot)$ is linear.[4] Hence, the agent is paid according to:

$$P(e_1,e_2;a_1^R,a_1^L,a_2,UL) = \begin{cases} a_2e_2 + a_1^L * e_1, & \text{if } e_1 < UL \\ a_2e_2 + a_1^R * e_1 + (a_1^L - a_1^R) * UL, & \text{if } e_1 \geq UL \end{cases} \tag{2}$$

which is depicted in Figure 1.

Proposition 1. *If there is no uncertainty, and if effort on task 2 is remunerated linearly at rate a_2, the agent's optimal choice (e_1^*, e_2^*) will satisfy one of the following:*

(i) $C_1(e_1^*,e_2^*;z) = a_1^L$ *and* $C_2(e_1^*,e_2^*;z) = a_2$ *if* $e_1^* < UL$,

(ii) $a_1^R < C_1(e_1^*,e_2^*;z) < a_1^L$ *and* $C_2(e_1^*,e_2^*;z) = a_2$ *if* $e_1^* = UL$,

(iii) $C_1(e_1^*,e_2^*;z) = a_1^R$ *and* $C_2(e_1^*,e_2^*;z) = a_2$ *if* $e_1^* > UL$.

[4] The first and third assumptions will be relaxed later. For the second one, note that this is a very natural assumption: if the agent becomes more efficient and its costs decrease, she will exert more effort. It is indeed guaranteed for the case of complements as we already know that costs are decreasing on z. For the case of substitutes, one possible scenario is that at some value of z the optimal strategy becomes to increase e_2 and decrease e_1. Its main implication would be that unless other parameters vary, there should be a maximum value for e_1. For ease of exposition, we abstract for this case.

For the cases (i) and (iii), in which the optimal e_1^* is not at the kink, the optima are given by equating marginal costs to marginal benefits, which are given by the corresponding piece rate a_1^R or a_1^L, and a_2. As expected, whenever $e_1^* < UL$ or $e_1^* > UL$, more efficient agents (as indexed by larger values of z), choose larger values of e_1^*: $C_1(e_1^*, e_2^*)$ is decreasing in z so to satisfy (i) or (iii) a larger value of e_1^* is needed for larger values of z. This is shown in Figure 2 by the corresponding increasing solid response functions.

Case (ii) is also represented in Figure 2 by those agents for whom the optimal choice of task 1 effort is UL. These are agents that, at the piece-rate of a_1^L, are efficient enough to produce at least UL, but given the lower piece-rate, a_1^R, do not find optimal to produce more than UL. Note that $a_1^R = C_1(e_1 = UL, e_2^*; \bar{z}) < C_1(e_1 = UL, e_2^*; \tilde{z}) < C_1(e_1 = UL, e_2^*; \underline{z}) = a_1^L$. At $e_1 = UL$ (point B in Figure 2) the agent with productivity $\tilde{z}$ would decrease profits if she increased e_1 because its marginal cost, $C_1(e_1 = UL, e_2^*; \tilde{z})$, is larger than the marginal benefit (a_1^R). Similarly, she would decrease profits if she decreased e_1 because its marginal cost is smaller than the marginal benefit, a_1^L. Hence, the optimal choice for the agent with productivity $\tilde{z}$ is UL. The same argument can be repeated for any agent whose productivity lies in the $[\underline{z}, \bar{z}]$ interval. This positive mass of agents whose efficiency parameter lies within $(\underline{z}, \bar{z})$ are for whom case (ii) in Proposition 1 applies, and they are the ones that constitute the bunching mass, which takes us to Proposition 2. ■

Proposition 2. *Without uncertainty, the presence of a kink at $e_1 = UL$ generates bunching on the distribution of effort on task one, $H(e_1)$.*

Proof: see Appendix A.

Figure 3 extends the previous example in Figure 2 and considers a uniform density $g(z)$ and how it transforms into $h(e_1)$. For $z < \underline{z}$, the kink makes no difference at all: $h(e_1) = h^L(e_1)$. However, for those $z \in [\underline{z}, \bar{z}]$ the presence of the kink becomes binding. Without the kink, such provider would have exerted $e_1 \in [UL, UL + \Delta e]$, between points A and D in the figure, which would have followed the density $h^L(e_1)$. Because of the kink, AD became AC and the entire area b is now collapsed into a unique spike at $e_1 = UL$. Finally, for $z > \bar{z}$ we have that optimal effort is given by $e_1^R(z)$, which is reflected by density $h^R(e_1)$. [5]

[5] Notice that it is required that $1 - H^L(UL) = 1 - H^R(UL) + b$, so the final $H(e_1)$ is a valid

Having established the existence of bunching, we now focus on analysing how task 1 optimal effort, e_1^*, changes when the piece rate of task 2, a_2, changes. The following proposition establishes the key insight from the theoretical model, which will constitute the building block of our identification strategy: that the optimal task 1 effort, e_1^*, of agents who are bunched at $e_1 = UL$ is less sensitive to changes in a_2 than those who are not bunched.

Proposition 3. *If there is no uncertainty, and if effort on task 2 is remunerated linearly at rate a_2:*

(i) if $e_1^* \neq UL$ then $\frac{de_1}{da_2} = -\frac{C_{12}}{C_{11}C_{22} - C_{12}^2}$, and the sign of $\frac{de_1}{da_2}$ is opposite to the sign of C_{12}.

  (a) If the tasks are substitutes $(C_{12} > 0)$, we will have that $\frac{de_1}{da_2} < 0$.

  (b) If the tasks are complements $(C_{12} < 0)$ then $\frac{de_1}{da_2} > 0$;

(ii) if $e_1^* = UL$, then $\frac{de_1}{da_2} = 0$

Proof: see Appendix A.

2.2 Uncertainty and risk aversion

The model above provides the basic intuition for understanding the implications of the two-parts contract on the agent's effort in one task to marginal changes in the financial reward of another task. Here, we discuss the implications of introducing uncertainty in the relation between effort and output (i.e. output, x_i, is a noisy measure of effort: $x_i = e_i + \varepsilon$).[6]

Figure 4 presents a simulation exercise using a Constant Relative Risk Aversion utility function, illustrating the optimal effort choice and the distribution of output under two scenarios: no uncertainty (the benchmark case) and uncertainty with

CDF. This is reflected in the fact that all observations that would have covered $e_1 \in [UL + \Delta e, \infty)$, are now spread into $e_1 \in [UL, \infty)$. For the uniform example in Figure 3, this means that the maximum value of e_1 will fall, but the density at any point will be larger $(H^R(e_1) > H^L(e_1)$ for $e_1 \in [UL, e_1^R(z^{max})])$.

[6] We focus on the implications of uncertainty in a multitask setting across heterogeneous agents. See Zhou and Swan (2003), Oxholm (2016) and Oxholm et al. (2018) for the analysis of two-part contracts for one task.

risk-averse agents. In the benchmark case, there is bunching of e_1^* exactly at UL. Under uncertainty with risk-averse agents, bunching occurs not only at UL but also in its neighbourhood, asymmetrically towards higher values of e_1.

Figure 5 is useful to explain the implication of introducing uncertainty on $\frac{\partial e_1^*}{\partial a_2}$. Focusing on the top left panel (substitute tasks): at $a_2 = 3.1$, the agent with $z = 2$ is bunched at the kink ($e_1^* = UL$) whilst the agent with $z = 1.3$ chooses to exert less effort ($e_1^* < UL$). Mirroring what happens in the case without uncertainty, the optimal effort, e_1^*, of the agent below the kink ($e_1^* < UL$, $z = 1.3$) is more sensitive to changes in a_2 (larger $|\frac{\partial e_1^*}{\partial a_2}|$) than that of the bunched agent ($e_1^* = UL$, $z = 2$). As shown in the bottom left panel, $\frac{\partial e_1^*}{\partial a_2} \simeq -0.02$ at $a_2 = 3.1$ for the agent whose e_1 is below the kink ($e_1^* < UL$, $z = 1.3$) whilst it is $\frac{\partial e_1^*}{\partial a_2} \simeq -0.01$ for the agent at the kink ($e_1^* = UL$,$z = 2$).

From the above discussion, it is clear that proposition 3 does not strictly hold with uncertainty, the derivative $\frac{\partial e_1^*}{\partial a_2}$ is not null, however the qualitative conclusion remains that the absolute value of $\frac{\partial e_1^*}{\partial a_2}$ is smaller for the agent at the kink than that for agents located below the kink. The same happens for the case of complements (right panel of Figure 5). The crux of our empirical strategy is that the effort of agents who are bunched at the kink is less sensitive to changes in a_2 than that of agents who are not bunched.

2.3 Non-linear payment function in x_2

So far, we have only considered a non-linear payment function in task e_1 but not in task e_2. If there is kink at $x_2 = UL_2$, there is an interval $(\underline{\underline{z}}, \bar{\bar{z}})$ in which there is bunching in the distribution of x_2 at UL_2 (or slightly above under uncertainty and risk aversion). The presence of bunching in the distribution task 1 is not affected (propositions 1 and 2). With respect to proposition 3, $\frac{de_1}{da_2}$ would not only be null for those which $e_1 = UL$, but also for those for which $e_2 = UL_2$.

3 Empirical Strategy

In this section, we describe how to use the insights of the theoretical model described above to estimate whether tasks are complement or substitutes when incentive contracts are piece-wise linear.

3.1 Identification Strategy

In this subsection, we describe the main idea behind our identification strategy. In the basic set-up, we assume that there are only two tasks, although this can be generalised (as we do in our application). The contract of one task remains constant, whilst the contract of the other task changes from one period to the next (but changes are the same across all agents). We also assume that there is panel data available containing the output of each task and time period for a sample of agents of the population.

The main idea behind our identification strategy is shown in Figure 6, which shows the piecewise linear contracts for two tasks: 1 and 2. The effort levels for each task, e_1 and e_2, are shown in the horizontal axis and the monetary reward in the vertical one. For the sake of the argument, consider two agents: A and B. In time t, the effort levels are (e_1^A, e_2^A) and (e_1^B, e_2^B) for agent A and B respectively, whilst they are $(e_1^{A'}, e_2^{A'})$ and $(e_1^{B'}, e_2^{B'})$ in time $t+1$.

Without loss of generality, assume that the incentive contract is the same at time t and $t+1$ for task 1, but task 2 becomes less lucrative in time $t+1$ (for task 2, the solid line represents the incentive contract in time t, and the dashed line for time $t+1$). Because task 2 becomes less lucrative, both agents exert less effort in $t+1$ than in t ($e_2^{B'} < e_2^B$) and ($e_2^{A'} < e_2^A$). The consequence of this decrease in effort in task 2 will be different for task 1 depending on whether task 1 and 2 are complements or substitutes. Assume that they are complements ($C_{12} < 0$), hence the decrease in e_2 for agent A leads to increase the marginal cost of e_1, and consequently a decrease in e_1, ($e_1^{A'} < e_1^A$). However, agent B is at the kink for task 1, and hence if the change in the contract of task 2 is small enough, such agent B will continue to be at the same kink in $t+1$ (note $\frac{de_1}{da_2} = 0$ from proposition 3). This is because if e_1^B is at the kink, then $a_1^R < C_1^B(e_1^B, e_2^B) < a_1^L$, which will continue to hold if the change in e_2^B is small enough. This insensitivity of agent B's task 1 effort to small enough changes in the task 2 contract implies that it can be used as if he had been "unexposed" or at least "less-exposed" to the changes in the task 2 contract, and be useful to control for common shocks that happen after the effort choice, and hence be the basis for a difference-in-differences estimator (DiD), which we develop more formally below.

An implicit assumption of our approach is that the sign of C_{12} is the same independently of the value of z. Otherwise, positive and negative values might offset each other when averaging across agents with different values of z.

3.2 Empirical model

Assume that we have available a random sample of N agents, observed consecutively for two time periods $(t = 1, 2)$. For each agent i and time period t, we observe their task 1 output in both periods, that is, $x_{1i,t}$ $(i = 1...N, t = 1, 2)$, which has been produced exerting effort e_1. Assume that the payment function for output x_1 is exactly as (2) and is the same in the two periods. Regarding the contract, assume that the piece rate for task 2 output is different in $t = 1$ than in $t = 2$: $a_{2,t}{=}a_2^{'}$ if $t = 1$; and $a_{2,t}{=}a_2^{''}$ if $t = 2$. Without loss of generality, we assume that $a_2^{''}{<}a_2^{'}$.

Building on the argument explained at the end of subsection 3.1, we propose the following DiD regression:[7]

$$x_{1it} = \alpha_0 + \alpha_1 \mathbb{1}\{x_{1i,t=1} < UL\} + \alpha_2 \mathbb{1}(t = 2) + \alpha_3 \mathbb{1}\{x_{1i,t=1} < UL\} \cdot \mathbb{1}\{t = 2\} + v_{i,t}, \\ \text{for } t = 1, 2 \text{ and } \; i \in \{j \; : \; x_{1j,t=1} \leq UL\} \quad (3)$$

where $v_{i,t}$ is a randomly distributed error term, α_1 absorbs the idiosyncratic difference between the "sensitive group" (those agents i below the kink, $i \in \{j : x_{1j,t=1} < UL\}$) and the "insensitive"/"less-sensitive" group (those agents i at the kink, $i \in \{j : x_{1j,t=1} = UL\}$), α_2 absorbs aggregate shocks that take place after effort levels have been chosen, and α_3 estimates the change in x_{1i2} due to the change in task 2 incentives that took place in $t = 2$, that is, $\frac{de_1}{da_2}$ (which has the opposite sign to that of C_{12}). The above strategy implicitly assumes that x_1 is highly autocorrelated in the absence of changes to the contracts so that, those who are below the kink in $t = 1$ would also be below the kink in $t = 2$ should the contract have remained the same.

[7] As there is no staggered implementation of a policy and as controls do not play a central role in the design, more complex estimation strategies are not needed for this analysis.

If there is an additional period available, the estimation can be made robust to mean reversion as well as reference-point effects (Fehr et al., 2009; Abeler et al., 2011), whereas those who produce below the kink at period $t-1$ try to improve in period t, even if the contracts have not changed.[8] To take this into account, assume that there are three years of data ($t = 1, 2, 3$) and that the incentive contract is the same in the first two years, and it is only in the third when a_2 changes. In this case, the first-difference regression

$$\Delta x_{1i,t} = \alpha_1 \mathbb{1}\{x_{1i,t-1} < UL\} + \alpha_2 \mathbb{1}\{t = 3\} + \alpha_3 \mathbb{1}\{x_{1i,t-1} < UL\} \cdot \mathbb{1}\{t = 3\} + v_{i,t}$$
$$for\ t = 2, 3 \text{ and } \ i \in \{j \ : \ x_{1j,t=2} \leq UL\} \qquad (4)$$

would absorb the reference-point effect and potential mean reversion through α_1, α_3 will capture the change in x_{13} that is due to the change in task 2 piece-wise rate, a_2, net of reference-point and mean reversion effects. Note that, unlike regression (3), the dependent variable in regression (4) is in differences.

Concerning what we mention in subsection 2.3, the regression above can be adjusted to exclude from the sensitive group, $\mathbb{1}\{x_{1i,t-1} < UL\}$, those agents which are bunched in x_2 due to the non-linearity of the payment function of x_2.

3.3 Detection of Bunching

The above discussion assumes that there is bunching of e_1 at UL. As a preliminary step to apply the DiD estimator, we need to test for such bunching. To do this, we fit a parametric model on the observed distribution excluding an interval around UL and compare it with the observed distribution (as developed by Kleven (2016)).

We fit restricted cubic splines on the histogram excluding the interval $[UL_j, UL_j + L]$.[9] This strategy essentially splits the domain into segments defined by K knots (joint

[8] This reference-point effect is outside our model, and might or not exist. This strategy is agnostic and simply allows for it. In our particular application, we tend to find a robust positive effect, indicating that those who just fell short of the kink, increase their output in the following period, even if incentives remain the same between the two time periods.

[9] We use splines instead of polynomials as the latter can produce poor approximations in certain cases (Harrell, 2015, Chap 2.4.2). Spline interpolation is a parametric approach that is as easy to implement as a polynomial, without several of its limitations. For this purpose, we define bins on achievement following McCrary (2008) procedure ($\tilde{x}_h$) and count the number of

points) in order to fit the histogram (n_{hj}) of indicator j with a piece-wise cubic polynomial in the middle segments, and a linear function in the first and last ones. It requires the transformation of the domain variable (the midpoint of the bins, $\tilde{x}_h$) into $K-1$ constructed variables $\left(X_{jh}^{(k)}\right)$ that ensure that the resulting function's first and second derivatives are the same.[10] Such variables are included in the linear expression presented in regression (5) which also considers dummy variables that indicate the presence of an excluded bin $\left(\mathbb{1}\left\{\tilde{x}_h = l\right\}, \forall l \in [UL_j, UL_j + L]\right)$. The error term, u_{jh}, is assumed to be i.i.d. and normally distributed. This error should not be 'too' large, otherwise it would not be possible to detect the bunching even if the kink was present.[11] Thus, the equation to estimate becomes:

$$n_{jh} = \sum_{k=1}^{K} \omega_k X_{jh}^{(k)} + \sum_{l=UL}^{UL+L} \gamma_l \mathbb{1}\left\{\tilde{x}_h = l\right\} + u_{jh} \tag{5}$$

After the vector of parameters $\{\omega, \gamma\}$ is estimated, the counterfactual density is the predicted value of this regression without the dummies for the excluded range's contribution: $\hat{n}_{jh} = \sum_{k=1}^{K} \hat{\omega}_k X_{jh}^{(k)}$. Then, the excess number of observations that bunch above UL relative to the calculated counterfactual is the difference between the observed and counterfactual histograms in the excluded range. This is equivalent to the sum of the omitted dummies γ:

$$\tilde{b}_j = \sum_{l=UL}^{UL+L} \hat{\gamma}_l = \sum_{l=UL}^{UL+L} \left(n_{jh} - \hat{n}_{jh}\right)$$

Following Chetty et al. (2009), we compare the amount of excess bunching with

agents in each bin (n_{hj}). More precisely,

$$n_{jh} = \sum_{i=1}^{N} \mathbb{1}\left\{\frac{\tilde{x}_h - \tilde{x}_{h-1}}{2} \leq x_{ij} < \frac{\tilde{x}_{h+1} - \tilde{x}_h}{2}\right\}, \ \tilde{x}_h \in \{0.5, 1, 1.5, ..., 99.5\}$$

[10] The procedure was implemented in STATA 13 using mkspline command, using 5 to 7 knots determined by percentiles recommended in Harrell (2015, Chap 2.4.6).

[11] Appendix G presents simulations intended to show how estimates from both the bunching and DiD steps are modified with increasing variance. If the variance is large, not only it is not possible to detect bunching, but also the DiD results are unreliable. Theoretically, a large variance would also render the scheme less effective, as it means that the agent has little control over the outcome of his actions.

the average density per 1 pp. in the excluded range:[12]

$$b_j = \frac{\tilde{b}_j}{\frac{1}{L+1}\sum_{l=UL}^{UL+L} \hat{n}_{jh}}$$

To determine whether or not there is bunching, we perform a joint significance test of the omitted dummies from regression (5):

$$H_0 : \sum_{l=UL}^{UL+L} \hat{\gamma}_l = 0 \tag{6}$$

4 An application: The Quality and Outcomes Framework

4.1 Background

The incentive contracts and data that we exploit come from the *Quality and Outcomes Framework* (QOF), a performance reward programme which was introduced in England in 2004 to improve quality of health care. The programme financially rewards performance in a set of administrative and clinical tasks. The QOF was introduced as part of a broader reform of the contract that governed the relationship between the Department of Health and the PHCs (known as GP practices), which gave PHCs more flexibility on treating their patients and organising their staff: they can hire additional staff (even salaried doctors who are not partners of the PHC) and offer additional services in exchange for extra resources (NAO, 2008). The role of complementarities and substitution of tasks were a central concern since the introduction of the programme (Sutton et al., 2010; Doran et al., 2011).

We apply our proposed test to the tasks carried out by PHCs in 2009, 2010, and

[12] In case there is bunching, the estimated b_j overestimates the amount of it because it does not consider that some of the bunched observations in the interval $[UL_j, UL_j + L]$ should be above $UL_j + L$ in the counterfactual distribution, as predicted by the model. Chetty et al. (2009) correct for this using an iterative procedure in which the area above $UL_j + L$ is artificially increased in such a way that the area under both the observed and counterfactual densities is the same. Given that we only want to know if b_j is different from zero, this correction is not necessary for our procedure.

2011. In these years, there were around $8,000$ PHCs covering $7,000$ patients on average. They are staffed by General Practitioners (family doctors), nurses as well as administrative staff. The PHCs are independent businesses contracted by the Department of Health to provide primary care to the patients enrolled with them. The PHCs are funded by capitation transfers from the Department of Health and performance pay income that we describe below.[13]

Possibly to deal with the added workload that the QOF scheme incentivized, the number of nurses working in PHCs grew strongly until 2009 when it stabilised as shown in Figure 7 (HSCIC, 2012). This highlights the advantage of our proposed test, which can estimate longer-term responses after capital and labour have adjusted.

4.2 Incentive contract

We are concerned with the QOF clinical indicators, which are related to the quality in the management and prevention of chronic diseases: coronary heart disease, heart failure, diabetes, and provision of advice for smoking cessation, obesity, etc. (see Appendix C for a detailed description of the indicators).[14] Clinical indicators are all based on proportions, in which the numerator is the number of patients who fulfilled the indicator, and the denominator is the number who should have fulfilled it. For instance, indicators THYROI02 and COPD13 are defined as:

> **Indicator THYROI02:** The percentage of patients with hypothyroidism with a thyroid function tests recorded in the previous 15 months.
>
> **Indicator COPD13:** The percentage of patients with chronic obstructive pulmonary disease (COPD) who have had a review under-

[13] The incentive is paid to the PHC. While some PHCs consist of a single physician, the majority include multiple physicians. In our analysis, we abstract from the team-based nature of the incentive; for a detailed discussion of team-based incentives in healthcare, see Gaynor et al. (2004)

[14] In addition to the clinical indicators, there are administrative indicators (organisational, patient experience and provision of additional services). We focus on the clinical indicators because they are the ones most dependent on clinicians' efforts and because they are the most important in terms of the size of the reward. In 2009 and 2010, GPs could obtain up to 1000 points: 697 for the clinical domain, 167.5 for the organisational domain, 91.5 for patient experience, and 44 for additional services. In 2011, the clinical domain was reduced to 661 points, patient experience to 33, and 262 points were reallocated to organisational indicators.

taken by a healthcare professional, including an assessment of breathlessness using the MRC dyspnea score in the preceding 15 months.[15]

PHCs are given points for each indicator. Crucially for us, the mapping between the value of the indicator (the percentage) and the points is piece-wise linear as in the top panel of Figure 8, where the horizontal axis presents the percentage (achievement) and the vertical axis presents the number of points. The points are then translated into income after adjusting for the number of patients enrolled in the PHC and the prevalence of the health condition in the PHC's population (see Appendix B for further details).[16] All the piece-wise linear contracts of all indicators follow the same pattern: If achievement is below a lower limit (LL_j) zero points are awarded, and if it is above the upper limit (UL_j) the maximum amount of available points for indicator j are awarded.

There are two differences to note between the piece-wise linear contract outlined in Figure 8 and the simpler one in Figure 1 that we used in our theoretical section. In Figure 8, there is a flat line below the lower limit (LL_j) which does not feature in Figure 1. However, this is of little relevance in practice because, as the bottom panel of Figure 8 shows, the mass of PHCs at low values of the horizontal axis is negligible (and this is the case for all QOF indicators).

The second difference between Figures 1 and 8 is that for high values of the horizontal axis, the mapping in Figure 8 is flat whilst it is increasing in Figure 1. To reconcile both, we will appeal to health care providers' altruism, which is a standard assumption in the health economics literature.[17] Hence, the piece-rates (a^R and a^L) include both the monetary component and the intrinsic motivation.[18]

The bottom panel of Figure 8 also shows the presence of bunching around the upper kink (UL), as our Proposition 2 predicted, as it is typical of situations with kinks in budget constraints (Saez, 2010; Kleven, 2016). The bunching is very

[15] Dyspnea refers to a breathing complication.

[16] The adjustment for the prevalence of the health condition in the PHC's population should counteract an incentive to underdiagnose to score higher in the achievement measure.

[17] Kolstad (2013); Chalkley and Malcomson (1998); Ellis and McGuire (1986); Godager and Wiesen (2013); Makris and Siciliani (2013); Olivella and Siciliani (2017).

[18] The assumption of a linear benefit to patients' welfare is relaxed by Kaarboe and Siciliani (2011). In such a scenario, the relevant function is not $C(\cdot)$ but $B(\cdot) - C(\cdot)$, hence our results will signal complementarity or substitutability of this function.

clear for the COPD13 indicator: there is a sudden increase in the density just above the upper kink point. As we will see later on, this happens with the vast majority of QOF indicators, but there are some exceptions as THYROI02. As our model section indicated, substantial uncertainty will smooth out the bunching.[19] Measurement error, which is another typical reason for not detecting bunching (Kleven, 2016) is probably of not much concern in our case as the QOF data are based on administrative records for a large number of PHCs.[20]

4.3 The 2011 changes

We use data from 2009, 2010, and 2011. There main reason is that there were no changes in the contracts between 2009 and 2010, which allow us to take into account reference-point effects as well as possible mean reversion, as we explained in subsection 3.2. From 2010 to 2011, out of the 68 clinical indicators with piecewise contracts, 41 of them (worth 383 points) remained completely unchanged. The remaining were either removed, modified or replaced by new ones.

Table 1 summarises the changes in the clinical indicators that took place in 2011 in three broad categories: reward reduction, ambiguous changes, and increase reward. The top panel of Table 1 summarises the changes that we interpret as a reduction in a_2: (i) eight tasks were no longer rewarded in 2011 (32 points), (ii) two tasks were still rewarded, but Primary Health Centres could get a maximum of 22 points for them in 2011 instead of 26 in 2009/10, (iii) the UL was increased on two tasks responsible for a total of 22 points,(iv) the description associated to 4

[19] For instance, the staff of the PHC might have complete control in keeping records of standard tests or lab results. However, patients attending a specialised assessment, as required by some indicators, might depend on patients' willingness to attend the assessments as well as the availability of suitable time slots in nearby clinics. Indeed, Fichera et al. (2014) present a game in which physicians and doctors interact using their available tools, prescriptions and lifestyle, in response to QOF incentives.

[20] Physicians could artificially inflate the denominator by classifying certain patients as exceptions to QOF guidelines, deeming that they should not be treated according to QOF recommendations due to specific health conditions (Doran et al. 2006; Gravelle et al. 2010). During our sample period, an NHS-commissioned study found no evidence of gaming (Dixon et al., 2011). Consistent with this, using a regression discontinuity estimator, we estimate that the exception rate is statistically higher to the right than to the left of UL in only three of the indicators out of the 40 that we analyze. The NHS has in place a procedure to verify that the QOF related information is correctly reported through visits to the PHCs, which could be random or when inaccuracies are suspected (NHS Employers and BMA 2011, pg. 7).

tasks became stricter or had to be fulfilled in a shorter time-frame (18 points), and (v) for two tasks not only there was a decrease in the maximum number of points that could be obtained, but also the updated task description became stricter. A more detailed explanation of these changes is presented in Tables C1 and C3 of Appendix C.

The second category (ambiguous change) covers several amendments that are not straightforward to classify as an increase or decrease of a_2. In these cases, typically, the task description became stricter but it is accompanied by additional points in compensation. There are four indicators in this category for which the total number of points increased from 51 to 59, but the difficulty in accomplishing them also increased.

The third category (increase in a_2) includes one indicator in which the number of points remained the same (17 points) but the 2011 task description became more lenient than the 2009/10 one as well as three new indicators, covering 12 points, which refer to three tasks that were neither financially rewarded in 2009 nor in 2010.

As summarised in Table 1, the total amount of points (as well as indicators) associated with a decrease in a_2 are far more than those associated with an increase, even if we consider all ambiguous changes as increases. Hence, we interpret the overall changes in 2011 as an overall reduction of a_2.[21]

5 Results

Figure 9 reports a heatmap of the achievement values of the clinical indicators (relative to UL) in 2009 and 2010 when there were no changes in the incentive contracts. As was the case in Figure 8, most of the mass is to the right of UL,

[21] Apart from clinical indicators, administrative (non-clinincal) ones also changed in 2011. Two-thirds of the *patient experience* domain were removed in favour of the new *quality and productivity* indicators. PHCs had to agree with the local commissioning body on a plan with three main goals: prescribing (28 points), outpatient referrals (21 points) and emergency admissions (47.5 points). The exact indicator definition and its upper threshold were defined at the local level. The objective of these changes was to reduce the costs of the local commissioning body by improving the cost-efficiency of prescribing and by treating more patients at the primary care level, reducing both referrals and emergency admission rates. They were offset by the withdrawal of four non-clinical indicators totalling 60.5 points.

which corresponds to 0 in Figure 9. It also shows that there is far less mass in the left tail of achievement each year. A very important point to note is the high degree of autocorrelation, as most observations are in the 45-degree diagonal or very close to it. Appendix D presents a more detailed analysis of the high autocorrelation of the indicators.

5.1 Basic specification

To implement the empirical strategy outlined in section 3.1, we divide the QOF clinical indicators with piece-wise linear rewards between those 41 for which there was no change in the contract between 2010 and 2011, x_1, and those 26 for which there was a change (including three completely new ones), x_2.

The results are presented in two steps. First, we assess which of the x_1 indicators feature bunching at the upper kink point (UL), as we discussed in subsection 3.3. Second, we estimate the response of each x_1 indicator to the changes that occurred in the x_2 indicators. Note that because 26 indicators changed simultaneously, we can only estimate how each x_1 reacted to the aggregate changes to the x_2, rather than pairwise, and moreover the value of x_2 is not observed when the indicator changes definition or is removed from the incentive scheme. As concluded at the end of subsection 4.3, we interpret the changes in the x_2 contracts as an overall reduction in a_2.

For each x_1 indicator, we assess the existence of bunching by pooling data from 2009 and 2010 and set a 10 pp estimation window below and above UL. We also discard the bins corresponding to 100%, where achievement is naturally truncated whenever the window covers this value. Figure 10 reports the amount of excess bunching for two example indicators (THYROI02 and COPD13). The figure presents the fitted model, including dummies γ covering $[UL, UL + 3\,\text{pp.}]$ (orange line) and excluding them from the prediction (black line). For COPD13, the difference between the histogram and the counterfactual difference is 28.4% of the average density in the interval $[UL - 10\text{pp.}, UL + 10\,\text{pp.})$; and for THYROI02 it is 2.7% only. Whilst we reject that the bunching estimate of COPD13 is zero at the 95% level, we do not for THYROI02. Therefore, our empirical strategy will be informative about the former indicator but not the latter. In total, we find evidence of bunching for 25 out of the 41 x_1 indicators, see Appendix E for a detailed description of the bunching results.

Table 2 presents the second part of the analysis in which we estimate regression (4) within the sample range $UL - 10 \leq x_{1,t=2} \leq UL + 3$, where x_1 refers to one of the 25 indicators whose contract remained unchanged throughout the three years that we consider (2009-2011) and for which we found evidence of bunching. [22] For each indicator (rows), Columns (1) and (2) report the number of observations used to estimate regression (4): Column (1) refers to the number of observations in the $[UL - 10, UL)$ interval and Column (2) in the $[UL, UL + 3]$ interval. Columns (3) to (5) of the Table present the estimates for α_1, α_2, and α_3 of regression (4).

The estimate of α_3 corresponds to the change in the x_1 indicator, dx_1, which is due to the change in the incentive contracts of the x_2 indicators, which overall was a net reduction in the marginal revenue, $da_2 < 0$, as explained in subsection 4.3. Hence, a negative sign of $\hat{\alpha}_3$, $dx_1 < 0$, would indicate that the positive cross-derivative is positive $\left(\frac{dx_1}{da_2} > 0\right)$, consistent with the x_1 indicator being a complement to the x_2 indicators. This would not mean that the corresponding x_1 indicator is a complement to all the x_2 indicators, but that overall, the net response is equivalent to complements. [23]

Table 2 reports negative and statistically significant estimates of α_3 associated to x_1 indicators AF03, AF04, ASTHMA06, CKD06, and DM13, consistent with them being overall complements of the x_2 indicators: effort exerted on these x_1 indicators was reduced in response to the overall reduction in the piece-rate of the x_2 indicators.

Table 3 shows the robustness of the results to different specifications of the estimation window. The first four columns report the estimates of α_3 for the sample included in $[UL{-}l, UL{+}k]$, for different values of l and k. In the last four columns, a

[22] Regression (4) was define for $x_{1,t=2} \leq UL$, in which $x_{1,t=2} = UL$ corresponds to the insensitive group. In practice, the PHCs with exactly $x_{1,t=2} = UL$ is very small so we expand the definition of the insensitive group to $[UL, UL{+}3]$. We also restrict the sample $UL{-}10 \leq x_{1,t=2}$ to keep some homogeneity amongst the PHCs compared, and we discard observations above $UL + 3$ because such observations have a value close to 100, where bunching also tends to occur. In Tables 4 and 5, we show that our results are robust to alternative thresholds.

[23] Another possibility is that the task is a substitute only of those tasks for which the marginal reward was increased instead of reduced. This seems unlikely as only four indicators had their reward increased whilst 18 had their reward decreased.

doughnut specification is considered, in which PHCs with achievement within 1 pp. of UL are removed. This table is restricted to those cases in which the hypothesis $\alpha_3 = 0$ is rejected with a 10% significance in at least one of the specified estimation windows: this is the case for the five indicators from the benchmark specification as well as for ASTHMA08, COPD13, DEM02, SMOKE04 and STROKE10. Note that the first column, $(l = 10, k = 3)$, of Table 3 corresponds to the benchmark results (column (5) of Table 2). Estimates for AF03, AF04, CKD06, and DM13 are generally stable across the different estimation windows used.

Our empirical strategy can also be implemented by estimating jointly the regressions of the indicators of the same disease group using seemingly unrelated regressions (SUR), and adjusting the p-values for multiple hypothesis testing at the disease group level using (Romano and Wolf, 2016, 2005a,b), see Appendix F for details.[24] The results are qualitatively very similar to those of Table 2, except that DM13 and ASTHMA06 marginally lose statistical significance with adjusted p-values of 0.11 and 0.14 respectively.

5.2 Bunching in several indicators

As we saw in subsection 2.3, even if agent i is not at the kink for indicator j, (i.e. $x_{ij} < UL_j$), x_{ij} might still be insensitive to the 2011 contract changes because agent i is at the kink in an indicator other than j. To test this prediction of our model, Table 4 reports how the effect of the changes in the reward schemes in 2011 varied with the number of indicators that PHCs had at the kink, $[UL, UL + 3]$, in 2010 (Figure 11 shows that there is significant variation across PHCs in the number of indicators that they had at the kink). Although the estimates of the last two columns are less precise, in general the results are in line with our prediction: the estimates are larger (in absolute value) for those PHCs which have fewer indicators at the bunching window (column 4), which are the PHCs that our model predicts should be more sensitive. Indeed, for these PHCs, we find the size of the effect can be quite large, with reductions of up to 14.7 percentage points.

[24] P-values adjusted for multiple hypothesis testing were computed using 1000 bootstrap repetitions, using a routine modified from Clarke (2016) STATA rwolf module.

5.3 Rationale of the findings

Why was the effort on the indicators AF03, AF04, CKD06, and DM13 reduced if the incentive contract for these indicators did not change in 2011? Why did a reduction in the marginal benefit of the x_2 indicators lead to a reduction in the effort exerted on these four x_1 indicators? Below, we explain why these indicators are complements of other indicators whose marginal benefit was reduced in 2011.[25]

The reduction in DM13 (percentage of patients with diabetes who have a record of micro-albuminuria testing in the previous 15 months) might be explained by the reduction in the marginal benefit of other indicators related to diabetes mellitus.[26] In particular, the financial rewards for keeping records of plasma glucose concentration, blood pressure, and cholesterol (DM5, DM11, DM16) for diabetic patients were removed. This probably led to a decrease in the levels of these tasks, increasing the marginal cost to measure microalbuminuria and fulfil DM13: the marginal cost of measuring micro-albuminuria (DM13) is smaller if plasma glucose concentration, blood pressure, and cholesterol are also being measured (they are complements in the cost function).[27]

We have also documented a decrease in indicators related to patients with atrial fibrillation (a rapid and irregular heartbeat), in particular in the percentage of patients with atrial fibrillation who are being treated with anticoagulant drug therapy (AF03), and in the percentage of atrial fibrillation patients who had their diagnosis confirmed by a specialist or with a specialised test (AF04). Atrial fibrillation is more common amongst diabetic patients,[28] and hence the marginal cost of fulfilling AF03 and AF04 is higher if diabetic patients visit the PHC less often because their plasma glucose concentration, blood pressure, cholesterol, and micro-albuminuria are being measured less often. This is particularly so because such costs include the ones related to identifying and contacting patients, and ar-

[25] We would like to thank Dr Alberto Vera González for his guidance on this subsection.

[26] Micro-albuminuria is a small increase in the level of albumin in the urine compared to normal. It can be an early sign of kidney disease, which often occurs as a complication of diabetes, high blood pressure, and heart failure.

[27] The data does not contain information on indicators that were removed or changed, so we cannot fully ascertain that there was a decrease in DM5, DM11, DM16.

[28] See, for instance, Ahmadi et al. (2020) and https://www.nhs.uk/conditions/atrial-fibrillation/causes/.

ranging for them to visit the PHC or hospital for consultations and tests. The marginal costs would be smaller if the patient is attending the consultations for other reasons, such as measuring of plasma glucose concentration, blood pressure, cholesterol, and micro-albuminuria.

A similar reason can explain the decrease in the CKD06 indicator: the percentage of patients on the chronic kidney disease register whose notes have a record of albumin-creatinine ratio value in the previous 15 months. Diabetes is a leading cause of chronic kidney disease, so a high percentage of patients in the chronic kidney disease register will also be diabetic. The marginal cost of measuring the value of the albumin-creatinine ratio for these patients would be higher if they are not attending the PHC to have their plasma glucose concentration, blood pressure, and cholesterol measured.

6 Discussion

As a case study, we have analysed the Quality and Outcomes Framework, a pay-for-performance scheme for family doctors in the UK, and amongst the largest primary care pay-for-performance schemes worldwide. We do find evidence of complementarities among the tasks that we consider, and no evidence of substitution, which might be due to increases in PHC staff that have taken place since the introduction of the pay-for-performance scheme in 2004.

Programmes aiming to align incentives between insurers and health care providers, like QOF, are becoming common around the world. In the US, the Affordable Care Act (ACA) includes numerous measures to incentivise the quality (or value) of care, rather than the volume of care, including Accountable Care Organisations, as well as Merit-Based Incentive Payment Systems, bundled payments, and payment jumps for long-term care hospitals implemented under the Medicare, and it is expected that their importance will only grow (Bhattacharya, 2018).[29]

[29] See for instance Cheng et al. (2020); Colla and Fisher (2014); Doran et al. (2017); Greene et al. (2015); Hussey et al. (2017); Lin et al. (2017); Song (2014); Einav et al. (2021, 2020, 2018). Other public initiatives such as the Primary Care Information Project of the New York Department of Health, as well as commercial insurers, have implemented payment schemes that reward quality of care (Bardach et al., 2013; Forward, 2016). The ACA also established initiatives that linked remuneration to performance in a hospital setting, such as the Hospital Readmission Reduction

Pay for performance programmes with a very similar structure to that of QOF are also found elsewhere. Examples include the *rémunération sur objectifs de santé publique* (Rosp), which was introduced in France in 2011, with similar concerns about the potential issues associated with task substitution (Dormont, 2013), as well as the 'Meta Asistencial' in Uruguay. More recently, the Medicare Hospital Value-based Purchasing Program also uses piece-wise linear contract across several indicators to link financial incentives to hospital performance on quality and costs (Norton et al., 2018).

More generally, piece-wise linear incentive contracts for healthcare providers are widespread, possibly to concentrate the financial rewards around the presumed "optimal" level. Health Maintenance Organizations award monetary bonuses to group of physicians if their expenditure fall below a target (Gaynor et al., 2004). Gruber et al. (2023) studies a scheme that penalized emergency departments in England if they took longer than four hours to treat patients, while Gupta (2021) studies the Hospital Readmission Reduction Program, which penalizes hospitals with above average readmission rates, with no gains for those below average.

The success of these pay-for-performance initiatives depends crucially on whether the different activities that providers undertake are complements or substitutes. Despite the importance of incentive schemes in health care, little is known on whether tasks are complements or substitutes. Recent cluster randomised trials from high-income countries have failed to examine this question (Asch et al., 2015; Bardach et al., 2013). Dumont et al. (2008) found Canadian physicians who voluntary signed up to a contract which decreased the marginal revenue of a consultation, decreased the number of consultations and increased the average time per consultation (an indicator of quality) as well as other activities unremunerated at the margin (i.e. teaching).[30] Assuming common trends between incentivised

Program, the Hospital-Acquired Condition Reduction Program, and the Hospital Value-Based Purchasing Program. Since 2008, hospital quality in the UK has been incentivised through the Advancing Quality programme, which was modelled after the Hospital Quality Incentive Demonstration of the US (Sutton et al., 2012).

[30] Mullen et al. (2010) examine performance reports of physician medical groups contracting with a large network HMO and compare clinical quality before and after the implementation of a pay-for-performance scheme, relative to a control group. Although they do not find much evidence that quality on unrewarded tasks deteriorated, they do not find much improvement in rewarded tasks either, probably because the size/salience of the reward was too small.

and unincentivised tasks, Sutton et al. (2010) exploits the introduction of QOF to find evidence that the tasks under consideration are complements.[31]

Building on Hölmstrom and Milgrom (1991)'s original insight that jobs should group together tasks that are complements, our finding that tasks involved in managing chronic diseases are complements suggests that health systems relying more heavily on primary care, where the same health center initially handles a patient's various illnesses, might be more efficient. This efficiency arises because effort exerted in one task reduces the marginal cost of other tasks.[32]

Our findings are very relevant given that the number of co-morbidities per individual increases with population ageing (MacMahon, 2018), and the number of chronic conditions per individual is already very sizeable in developed economies and it is increasing rapidly in lower income countries.[33] Common clusters of diseases involve conditions associated with cardiovascular disease (diabetes, hypertension, coronary artery disease), mental health conditions (mainly depression), or osteoarthritis (Violan et al., 2014). Therefore, increasing resources to serve patients with certain conditions might lead to the detection and/or proper control of other conditions that the same patient suffers from.

The presence of complementarities in primary care supports the use of financial incentives and the pooling of health care activities within family doctors. Our results provide evidence in favour of health care systems in which primary care centers play an important role, where the family doctor is, in the first instance, responsible for managing the patient's care across all health conditions, potentially exploiting the complementarities that we have found. Such complementarities could also partially explain the massive consolidation that has taken place among primary care physicians in the US during the last two decades, as well as the reductions on healthcare spending associated with such consolidation (Muhlestein

[31] There are a number of important differences with our paper: (1) our result does not rely on the assumption of common trends across tasks, (2) we can also study complementary/substitution patterns amongst rewarded tasks, and not only compare rewarded and unrewarded, (3) our results are after six years of programme implementation, which allows capital and labour to adjust.

[32] In line with our findings, using records from the Veterans Health Administration, (Currie and Zhang, 2021) find that healthcare providers who are more effective at treating one condition are also more effective at treating other conditions, consistent with the complementarities we identified.

[33] At least 60% of adult Americans had at least one chronic condition, and 42% suffer from multimorbidity (Buttorff et al., 2017).

and Smith, 2016; Zhang et al., 2021).

7 Conclusion

Whether tasks are complements or substitutes is crucial to understanding how best to split tasks amongst agents (job design), as well as the optimal design and ultimate success of any pay-for-performance scheme (Hölmstrom and Milgrom, 1991). We propose an empirical strategy to identify whether tasks are complements or substitutes in a setting when there is a two-part linear contract. The test, which requires variation of the pay-for-performance contract over time, works by considering as an "insensitive/less-sensitive" group those agents who are bunched at the "kink" of the reward scheme.

Our test has three advantages: (1) it does not require variation of contracts across agents, and hence it can be used if all agents in the population face the same contracts as long as there is some variation across time in the contract of at least one of the tasks, (2) it is not restricted to assess substitution patters between rewarded and unrewarded tasks, but can also estimate such patters amongst rewarded tasks, (3) it does not require data on agents' performance before the introduction of the incentive contracts, which allows obtaining the estimates after the incentives contracts have been implemented for several years, and hence to recover steady-state responses after capital and labor have adjusted. We hope that our proposed empirical strategy will contribute to closing the gap between the assumptions made in principal-agent models and the empirical relevance of such assumptions.

We apply our empirical strategy to the *Quality Outcomes Framework* (QOF), a nationwide pay for performance programme which was rolled out simultaneously across England in 2004. We find that several tasks are complements and none are substitutes. In interpreting the results, it is important to note that we are measuring the effects six years after QOF was implemented, and hence capital and labour have already adjusted, leading to the relaxation of short-term capacity constraints. The finding that tasks are complements might help explain the widespread use of pay for performance programmes in health care.

References

Abeler, J., A. Falk, L. Goette, and D. Huffman (2011). Reference points and effort provision. *American Economic Review 101*(2), 470–92.

National Audit Office (NAO) (2008). *NHS pay modernisation: new contracts for general practice services in England*, Volume 307. The Stationery Office.

Ahmadi, S. S., A.-M. Svensson, A. Pivodic, A. Rosengren, and M. Lind (2020). Risk of atrial fibrillation in persons with type 2 diabetes and the excess risk in relation to glycaemic control and renal function: a swedish cohort study. *Cardiovascular diabetology 19*(1), 1–12.

Al-Ubaydli, O., S. Andersen, U. Gneezy, and J. A. List (2015, January). Carrots That Look Like Sticks: Toward an Understanding of Multitasking Incentive Schemes. *Southern Economic Journal 81*(3), 538–561.

Asch, D. A., A. B. Troxel, W. F. Stewart, T. D. Sequist, J. B. Jones, A. G. Hirsch, K. Hoffer, J. Zhu, W. Wang, A. Hodlofski, et al. (2015). Effect of financial incentives to physicians, patients, or both on lipid levels: A randomized clinical trial. *JAMA 314*(18), 1926–1935.

Atkinson, A., S. Burgess, B. Croxson, P. Gregg, C. Propper, H. Slater, and D. Wilson (2009). Evaluating the impact of performance-related pay for teachers in england. *Labour Economics 16*(3), 251–261.

Bardach, N., J. Wang, S. De Leon, and et al (2013). Effect of pay-for-performance incentives on quality of care in small practices with electronic health records: A randomized trial. *JAMA 310*(10), 1051–1059.

Bénabou, R. and J. Tirole (2003, July). Intrinsic and Extrinsic Motivation. *The Review of Economic Studies 70*(3), 489–520. ArticleType: research-article / Full publication date: Jul., 2003 / Copyright 2003 The Review of Economic Studies, Ltd.

Bénabou, R. and J. Tirole (2006, December). Incentives and Prosocial Behavior. *The American Economic Review 96*(5), 1652–1678. ArticleType: research-article / Full publication date: Dec., 2006 / Copyright 2006 American Economic Association.

Bhattacharya, J. (2018). MACRA and the Future of Medicine.

Bradler, C., R. Dur, S. Neckermann, and A. Non (2016, February). Employee Recognition and Performance: A Field Experiment. *Management Science 62*(11), 3085–3099.

Brickley, J. A. and J. L. Zimmerman (2001, December). Changing incentives in a multitask environment: evidence from a top-tier business school. *Journal of Corporate Finance 7*(4), 367–396.

Buttorff, C., T. Ruder, and M. Bauman (2017). *Multiple chronic conditions in the United States.* RAND Santa Monica, CA.

Carr, D. and M. E. Pebesma (2019). Package âĂŸhexbinâĂŹ.

Celhay, P. A., P. J. Gertler, P. Giovagnoli, and C. Vermeersch (2019). Long-Run Effects of Temporary Incentives on Medical Care Productivity. *American Economic Journal: Applied Economics.*

Chalkley, M. and J. M. Malcomson (1998). Contracting for health services when patient demand does not reflect quality. *Journal of health economics 17*(1), 1–19.

Chalkley, M. J., A. Mirelman, L. Siciliani, and M. E. Suhrcke (2016, December). Paying for performance for health care in low- and middle-income countries: : an economic perspective. Discussion paper, ARRAY(0x7fd3580747b0), York, UK.

Cheng, J., J. Kim, S. D. Bieber, and E. Lin (2020). Four years into macra: What has changed? In *Seminars in Dialysis*, Volume 33, pp. 26–34. Wiley Online Library.

Chetty, R., J. N. Friedman, T. Olsen, and L. Pistaferri (2009). Adjustment costs, firm responses, and micro vs. macro labor supply elasticities: Evidence from danish tax records. Technical report, National Bureau of Economic Research.

Chiappori, P.-A. (2000). Econometric models of insurance under asymmetric information. In G. Dionne (Ed.), *Handbook of Insurance*, pp. 365–393. Dordrecht: Springer Netherlands.

Chiappori, P.-A. and B. Salanié (1997). Empirical contract theory: The case of insurance data. *European Economic Review 41*(3-5), 943–950.

Chiappori, P.-A. and B. Salanié (2003). Testing contract theory: A survey of some recent work. In M. Dewatripont, L. Hansen, and S. Turnovsky (Eds.), *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress. Econometric Society Monographs*, pp. 115–149. Cambridge University Press.

Clarke, D. (2016). Rwolf: Stata module to calculate romano-wolf stepdown p-values for multiple hypothesis testing.

Colla, C. H. and E. S. Fisher (2014). Beyond pcmhs and accountable care organizations: payment reform that encourages customized care.

Currie, J. and J. Zhang (2021, June). Doing more with less: Predicting primary care provider effectiveness. Working Paper 28929, National Bureau of Economic Research.

De Philippis, M. (2015, November). Multitask Agents and Incentives: The Case of Teaching and Research for University Professors.

Deci, E. L. and R. M. Ryan (1985). *Intrinsic motivation and self-determination in human behavior*. Perspectives in social psychology. New York: Plenum.

Dixon, A., A. Khachatryan, A. Wallace, S. Peckham, T. Boyce, and S. Gillam (2011). Impact of quality and outcomes framework on health inequalities. *London: The King's Fund*.

Doran, T., C. Fullwood, H. Gravelle, D. Reeves, E. Kontopantelis, U. Hiroeh, and M. Roland (2006). Pay-for-performance programs in family practices in the united kingdom. *New England journal of medicine 355*(4), 375–384.

Doran, T., E. Kontopantelis, J. M. Valderas, S. Campbell, M. Roland, C. Salisbury, and D. Reeves (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*, d3590.

Doran, T., K. A. Maurer, and A. M. Ryan (2017, March). Impact of Provider Incentives on Quality and Value of Health Care. *Annual Review of Public Health 38*(1), 449–465.

Dormont, B. (2013). Le paiement à la performance: contraire à lâĂŹéthique ou au service de la santé publique? *Sève* (3), 53–61.

Dumont, E., B. Fortin, N. Jacquemet, and B. Shearer (2008). Physicians' multitasking and incentives: Empirical evidence from a natural experiment. *Journal of Health Economics 27*(6), 1436 – 1450.

Eijkenaar, F. (2012). Pay for performance in health care: An international overview of initiatives. *Medical Care Research and Review 69*(3), 251–276. PMID: 22311954.

Einav, L., A. Finkelstein, Y. Ji, and N. Mahoney (2020). Randomized trial shows healthcare payment reform has equal-sized spillover effects on patients not targeted by reform. *Proceedings of the National Academy of Sciences 117*(32), 18939–18947.

Einav, L., A. Finkelstein, Y. Ji, and N. Mahoney (2021). Voluntary regulation: Evidence from medicare payment reform. *Quarterly Journal of Economics, forthcoming*.

Einav, L., A. Finkelstein, and N. Mahoney (2018). Provider incentives and healthcare costs: Evidence from long-term care hospitals. *Econometrica 86*(6), 2161–2219.

Ellis, R. P. and T. G. McGuire (1986). Provider behavior under prospective reimbursement: Cost sharing and supply. *Journal of health economics 5*(2), 129–151.

Englmaier, F., A. Roider, and U. Sunde (2016, October). The Role of Communication of Performance Schemes: Evidence from a Field Experiment. *Management Science 63*(12), 4061–4080.

Fehr, E. and A. Falk (2002, May). Psychological foundations of incentives. *European Economic Review 46*(4), 687–724.

Fehr, E., C. Zehnder, and O. Hart (2009). Contracts, reference points, and competition - behavioral effects of the fundamental transformation. *Journal of the European Economic Association 7*(2-3), 561–572.

Fichera, E., J. Banks, and M. Sutton (2014). Health behaviours and the patient-doctor interaction: The double moral hazard problem. Technical report, Economics, The University of Manchester.

Finan, F., B. Olken, and R. Pande (2017). The Personnel Economics of the Developing State. In *Handbook of Economic Field Experiments*, Volume 2, pp. 467–514. Elsevier.

Forward, H. G. (2016, dec). More doctors to retire as macra and value-based pay hit.

Gaynor, M., J. Rebitzer, and L. Taylor (2004, August). Physician Incentives in Health Maintenance Organizations. *Journal of Political Economy 112*(4), 915–931.

Glewwe, P., N. Illias, and M. Kremer (2010). Teacher incentives. *American Economic Journal: Applied Economics 2*(3), 205–227.

Gneezy, U. and A. Rustichini (2000, August). Pay Enough or Don't Pay at All. *The Quarterly Journal of Economics 115*(3), 791–810. ArticleType: research-article / Full publication date: Aug., 2000 / Copyright Âľ 2000 Oxford University Press.

Godager, G. and D. Wiesen (2013). Profit or patientsâĂŹ health benefit? exploring the heterogeneity in physician altruism. *Journal of health economics 32*(6), 1105–1116.

Gravelle, H., M. Sutton, and A. Ma (2010). Doctor behaviour under a pay for performance contract: Treating, cheating and case finding?*. *The Economic Journal 120*(542), F129–F156.

Greene, J., J. H. Hibbard, and V. Overton (2015, April). Large Performance Incentives Had The Greatest Impact On Providers Whose Quality Metrics Were Lowest At Baseline. *Health Affairs 34*(4), 673–680.

Gruber, J., T. P. Hoe, and G. Stoye (2023, January). Saving Lives by Tying Hands: The Unexpected Effects of Constraining Health Care Providers. *The Review of Economics and Statistics 105*(1), 1–19.

Gupta, A. (2021, April). Impacts of Performance Pay for Hospitals: The Readmissions Reduction Program. *American Economic Review 111*(4), 1241–1283.

Handel, B. R., J. Kolstad, A. Root, M. Whinston, and R. Huckman (2020). Outcomes-based payments and physician productivity: Evidence from diabetes care in hawaii. In *9th Annual Conference of the American Society of Health Economists*. ASHECON.

Harrell, F. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer.

Harris, M. and A. Raviv (1979, April). Optimal incentive contracts with imperfect information. *Journal of Economic Theory 20*(2), 231–259.

Hölmstrom, B. (1979). Moral hazard and observability. *The Bell Journal of Economics 10*(1), 74–91.

Hölmstrom, B. and P. Milgrom (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization 7*, 24–52.

Hong, F., T. Hossain, J. A. List, and M. Tanaka (2018). Testing the Theory of Multitasking: Evidence from a Natural Field Experiment in Chinese Factories. *International Economic Review 59*(2), 511–536.

HSCIC (2012). NHS workforce: Summary of staff in the NHS: Results from september 2011 census. `https://files.digital.nhs.uk/publicationimport/pub05xxx/pub05221/nhs-staf-2001-2011-medi-dent-work-rep.pdf`. Accesed: 2019-07-11.

Hussey, P. S., J. L. Liu, and C. White (2017). The medicare access and chip reauthorization act: Effects on medicare payment policy and spending. *Health Affairs 36*(4), 697–705.

Jacob, B. A. (2005, June). Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics 89*(5-6), 761–796.

Johnson, R. M., D. H. Reiley, and J. C. MuÃśoz (2015). âĂIJThe War for the FareâĂİ: How Driver Compensation Affects Bus System Performance. *Economic Inquiry 53*(3), 1401–1419.

Kaarboe, O. and L. Siciliani (2011). Multi-tasking, quality and pay for performance. *Health Economics 20*(2), 225–238.

Kleven, H. J. (2016). Bunching. *Annual Review of Economics 8*(1).

Kolstad, J. T. (2013). Information and quality when motivation is intrinsic: Evidence from surgeon report cards. *American Economic Review 103*(7), 2875–2910.

Kosfeld, M. and S. Neckermann (2011). Getting more work for nothing? symbolic awards and worker performance. *American Economic Journal: Micoeconomics 3*, 86–99.

Lazear (2000). Performance pay and productivity. *American Economic Review 90*(5), 1346–61.

Lin, E., T. MaCurdy, and J. Bhattacharya (2017). The medicare access and chip reauthorization act: Implications for nephrology. *Journal of the American Society of Nephrology 28*(9), 2590–2596.

MacMahon, S. (2018). Multimorbidity: A priority for global health research. *The Academy of Medical Sciences: London, UK*.

Makris, M. and L. Siciliani (2013). Optimal incentive schemes for altruistic providers. *Journal of Public Economic Theory 15*(5), 675–699.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics 142*(2), 698 – 714. The regression discontinuity design: Theory and applications.

Mohanan, M., K. Donato, G. Miller, Y. Truskinovsky, and M. Vera-Hernández (2021). Different strokes for different folks? experimental evidence on the effectiveness of input and output incentive contracts for health care providers with varying skills. *American Economic Journal: Applied Economic , (Forthcoming)*.

Muhlestein, D. B. and N. J. Smith (2016). Physician consolidation: rapid movement from small to large group practices, 2013–15. *Health Affairs 35*(9), 1638–1642.

Mullen, K. J., R. G. Frank, and M. B. Rosenthal (2010, March). Can you get what you pay for? Pay-for-performance and the quality of healthcare providers. *The RAND Journal of Economics 41*(1), 64–91.

Neal, D. (2011). The design of performance pay in education. In E. Hanushek, S. Machin, and L. Woessmann (Eds.), *Handbook of Economics of Education*, Volume 4, pp. 495–550. Oxford: Elsevier.

NHS Employers and BMA (2011). Quality and outcomes framework guidance for gms contract 2011/12. The NHS Confederation (Employers) Company. Ref: EGUI09201.

Norton, E. C., J. Li, A. Das, and L. M. Chen (2018, September). Moneyball in Medicare. *Journal of Health Economics 61*, 259–273.

Olivella, P. and L. Siciliani (2017). Reputational concerns with altruistic providers. *Journal of health economics 55*, 1–13.

Olken, B. A., J. Onishi, and S. Wong (2014, October). Should aid reward performance? evidence from a field experiment on health and education in indonesia. *American Economic Journal: Applied Economics 6*(4), 1–34.

Oxholm, A. S. (2016). Physician response to target-based performance payment. *COHERE Discussion Papers, University of Southern Denmark 9*.

Oxholm, A. S., S. R. Kristensen, and M. Sutton (2018). Uncertainty about the effort–performance relationship in threshold-based payment schemes. *Journal of health economics 62*, 69–83.

Paarsch, H. J. and B. Shearer (2000). Piece Rates, Fixed Wages, and Incentive Effects: Statistical Evidence from Payroll Records. *International Economic Review 41*(1), 59–92.

Prendergast, C. (2002). The Tenuous TradeâĂŘoff between Risk and Incentives. *Journal of Political Economy 110*(5), 1071–1102.

Roland, M. and F. Olesen (2016). Can pay for performance improve the quality of primary care? *BMJ 354*, i4058.

Romano, J. P. and M. Wolf (2005a). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association 100*(469), 94–108.

Romano, J. P. and M. Wolf (2005b). Stepwise multiple testing as formalized data snooping. *Econometrica 73*(4), 1237–1282.

Romano, J. P. and M. Wolf (2016). Efficient computation of adjusted p-values for resampling-based stepdown multiple testing. *Statistics & Probability Letters 113*, 38–40.

Saez, E. (2010). Do taxpayers bunch at kink points? *American Economic Journal: Economic Policy 2*(3), 180–212.

Shearer, B. (2004). Piece rates, fixed wages and incentives: Evidence from a field experiment. *Review of Economic Studies 71*(2), 513–34.

Song, Z. (2014). Accountable care organizations in the us health care system. *Journal of clinical outcomes management: JCOM 21*(8), 364.

Stiglitz, J. E. (1974). Incentives and risk sharing in sharecropping12. *The Review of Economic Studies 41*(2), 219–255.

Sutton, M., R. Elder, B. Guthrie, and G. Watt (2010). Record rewards: the effects of targeted quality incentives on the recording of risk factors by primary care providers. *Health economics 19*(1), 1–13.

Sutton, M., S. Nikolova, R. Boaden, H. Lester, R. McDonald, and M. Roland (2012, November). Reduced Mortality with Hospital Pay for Performance in England. *New England Journal of Medicine 367*(19), 1821–1828.

Sylvia, S., R. Luo, L. Zhang, Y. Shi, A. Medina, and S. Rozelle (2013). Do you get what you pay for with school-based health programs? evidence from a child nutrition experiment in rural china. *Economics of Education Review 37*, 1 – 12.

Violan, C., Q. Foguet-Boreu, G. Flores-Mateo, C. Salisbury, J. Blom, M. Freitag, L. Glynn, C. Muth, and J. M. Valderas (2014). Prevalence, determinants and patterns of multimorbidity in primary care: a systematic review of observational studies. *PloS one 9*(7), e102149.

Zhang, J., Y. Chen, L. Einav, J. Levin, and J. Bhattacharya (2021). Consolidation of primary care physicians and its impact on healthcare utilization. *Health Economics*.

Zhou, X. and P. L. Swan (2003). Performance thresholds in managerial incentive contracts. *The Journal of Business 76*(4), 665–696.

Table 1: Changes in QOF 2011 with respect to 2009-2010

Change in rewards (points)	**Status**	**Description**	**Total Points**	**Number Indicators**
Reduction (143 to 87)	Withdrawn I	No longer rewarded tasks	32	8
	Points reduced	Number of assigned points per indicator was reduced.	26 to 22	2
	Upper Limit Increased	Increase on UL	22	2
	Replacement I	New wording with a more strict definition of a goal or a reduced time-frame for accomplishing it	18	4
	Replacement II	The decrease in points and new wording is more detailed	45 to 25	2
Ambiguous (51 to 59)	Replacement III	Harder to accomplish or more detailed goals but compensated with extra points	51 to 59	4
Increase (29)	Replacement IV	More lenient task description	17	1
	New	New tasks to be rewarded	12	3
NA (486)	Replacement V	Similar or same wording, but expressed in new units or highlight recent changes on diagnostic procedures.	29	4
	Replacement V	As above, but these are fixed-payment indicators	3	1
	Unchanged	No change on points, thresholds or wording	383	41
	Unchanged	As above, but these are fixed-payment indicators	71	17

Note: Authors' interpretation based on NHS Employers public documents. There are 68 clinical indicators with a non-linear payment scheme in 2009 and 2010. In 2011, 3 indicators were added, 8 were removed, 19 were modified and 41 were unchanged.

Table 2: Estimates of regression (4) on the QOF dataset

		(1)	(2)	(3)	(4)	(5)	
		N Observations		Estim. Regression Coefficients			Classif.
Indicator	UL	$[UL-10,UL)$	$[UL,UL+3]$	BELOW	AFTER	INTER	
		Below	Above	α_1	α_2	α_3	
AF 3 - AF patients on anticoagulants/antiplatelets	90%	544	2455	0.026^{***}	0.001	-0.007^{**}	Comp
				(0.002)	(0.001)	(0.003)	
AF 4 - AF diagnosis confirmed by ECG/specialist	90%	284	799	0.033^{***}	-0.002	-0.016^{***}	Comp
				(0.004)	(0.002)	(0.006)	
ASTHMA 3 - Smoking status in young asthma patients	80%	248	737	0.025^{***}	-0.004	-0.007	
				(0.007)	(0.004)	(0.010)	
ASTHMA 6 - Asthma review in last 15 months	70%	421	1116	0.032^{***}	-0.011^{***}	-0.012^{*}	Comp
				(0.004)	(0.002)	(0.007)	
ASTHMA 8 - Asthma diagnosis with variability test	80%	325	1067	0.030^{***}	-0.007^{**}	-0.011	
				(0.005)	(0.003)	(0.008)	
CANCER 3 - Cancer review 6 months post-diagnosis	90%	897	970	0.022^{***}	-0.001	-0.009	
				(0.005)	(0.003)	(0.007)	
CHD 8 - Cholesterol ≤ 5mmol/L in CHD patients	90%	528	2275	0.011^{***}	-0.000	0.002	
				(0.003)	(0.001)	(0.003)	
CHD 10 - CHD patients on beta-blockers	60%	175	542	0.015^{*}	0.005	-0.006	
				(0.009)	(0.004)	(0.011)	
CHD 12 - Flu vaccine in CHD patients	90%	1240	2795	0.020^{***}	-0.002^{*}	-0.000	
				(0.002)	(0.001)	(0.003)	
COPD 13 - COPD review with breathlessness assessment	90%	1019	2516	0.023^{***}	0.000	-0.007	
				(0.003)	(0.001)	(0.004)	
CKD 3 - BP ≤ 140/85 in CKD patients	70%	1498	1428	0.017^{***}	0.011^{***}	-0.001	
				(0.002)	(0.002)	(0.003)	
CKD 6 - Albumin/creatinine ratio in CKD patients	80%	1167	1064	0.031^{***}	-0.017^{***}	-0.018^{***}	Comp
				(0.004)	(0.003)	(0.005)	
DEM 2 - Dementia patient care review	60%	102	300	0.050^{**}	0.021^{**}	-0.049	
				(0.024)	(0.010)	(0.035)	
DM 13 - Microalbuminuria test in diabetes patients	90%	2151	2327	0.015^{***}	-0.001	-0.006^{**}	Comp
				(0.002)	(0.001)	(0.002)	
DM 15 - ACE/ARB therapy in diabetes with proteinuria	80%	326	585	0.019^{***}	0.001	-0.007	
				(0.007)	(0.005)	(0.009)	
DM 21 - Retinal screening in diabetes patients	90%	1585	2397	0.014^{***}	-0.002^{*}	0.002	
				(0.002)	(0.001)	(0.003)	
EPILEPSY 8 - Seizure-free epilepsy patients recorded	70%	1139	976	0.012^{***}	0.007^{*}	-0.002	
				(0.004)	(0.004)	(0.006)	
HF 2 - Heart failure diagnosis confirmed by echo/specialist	90%	463	966	0.020^{***}	-0.003	0.004	
				(0.004)	(0.002)	(0.006)	

Continued on next page

Table 2: Estimates of regression (4) on the QoF dataset (Continued)

		(1)	(2)	(3)	(4)	(5)	
		N Observations		Estim. Regression Coefficients			Classif.
Indicator	UL	$[UL-10, UL)$ Below	$[UL, UL+3]$ Above	BELOW α_1	AFTER α_2	INTER α_3	
HF 3 - HF patients on ACE/ARB therapy	80%	186	606	0.019**	0.002	0.007	
				(0.009)	(0.004)	(0.011)	
SMOKING 4 - Smoking cessation advice for smokers	90%	950	2722	0.022***	0.000	−0.003	
				(0.002)	(0.001)	(0.003)	
STROKE 7 - Cholesterol recorded in stroke/TIA patients	90%	1550	2322	0.012***	−0.003**	0.002	
				(0.002)	(0.001)	(0.003)	
STROKE 8 - Cholesterol âĽď5mmol/L in stroke/TIA patients	60%	210	203	0.019*	−0.006	0.006	
				(0.010)	(0.009)	(0.015)	
STROKE 10 - Flu vaccine in stroke/TIA patients	85%	904	1497	0.015***	0.001	0.007	
				(0.003)	(0.002)	(0.005)	
STROKE 12 - Antiplatelet/anticoagulant in stroke/TIA patients	90%	580	1927	0.020***	0.000	−0.003	
				(0.003)	(0.001)	(0.004)	
STROKE 13 - New stroke/TIA patients referred for tests	80%	191	569	0.022**	−0.004	−0.010	
				(0.010)	(0.005)	(0.013)	

Notes: Own calculations based on QOF data. Sample defined over the interval: [UL - 10 pp.,UL + 3 pp.]. **BELOW:** To have attained below the respective upper threshold UL in the first year of the variation: 2009 for 2009-2010 and 2010 for 2010-2011 ($\mathbb{1}\{x_{1i,t-1} < UL\}$). **AFTER:** 2010 to 2011 variation ($\mathbb{1}\{t = 3\}$). **INTER:** Interaction between BELOW and AFTER. All regressions include the number of items in which the PHC was in the bunching area $[UL, UL+3]$ in the previous year as a control variable. Clustered at PHC-level standard errors in parenthesis. † PHCs' descriptive statistics are presented according to 2010 achievement, within the 3 points window around UL . Significance: * 1%, ** 5%, *** 1%.

Table 3: Estimates of the interaction term in regression (4) on the QOF dataset. Multiple windows

Estimate of α_3 under the sample in $[UL - l, UL + k]$
Presents only indicators for which $\alpha_3 = 0$ is rejected in at least one specification.

	Entire interval				Removing $[UL - 1, UL + 1]$			
	k=3 pp. above UL		k=2 pp. above UL		k=3 pp. above UL		k=2 pp. above UL	
Indicator	l=10	l=5	l=10	l=5	l=10	l=5	l=10	l=5
AF 3 - AF patients on anticoagulants/antiplatelets (UL=90)	−0.007**	−0.001	−0.007**	−0.002	−0.009*	−0.000	−0.009**	−0.001
	(0.003)	(0.003)	(0.004)	(0.003)	(0.005)	(0.005)	(0.005)	(0.005)
AF 4 - AF diagnosis confirmed by ECG/specialist (UL=90)	−0.016***	−0.015***	−0.016**	−0.015***	−0.023**	−0.027***	−0.021**	−0.025***
	(0.006)	(0.005)	(0.006)	(0.005)	(0.009)	(0.009)	(0.010)	(0.009)
ASTHMA 6 - Asthma review in last 15 months (UL=70)	−0.012*	−0.004	−0.009	−0.001	−0.016**	−0.005	−0.013	−0.002
	(0.007)	(0.007)	(0.007)	(0.007)	(0.008)	(0.009)	(0.009)	(0.010)
ASTHMA 8 - Asthma diagnosis with variability test (UL=80)	−0.011	−0.003	−0.011	−0.002	−0.016*	−0.006	−0.016*	−0.007
	(0.008)	(0.008)	(0.008)	(0.009)	(0.009)	(0.010)	(0.010)	(0.011)
COPD 13 - COPD review with breathlessness assessment (UL=90)	−0.007	−0.009*	−0.006	−0.009*	−0.006	−0.008	−0.005	−0.008
	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)	(0.007)	(0.006)	(0.007)
CKD 6 - Albumin/creatinine ratio in CKD patients (UL=80)	−0.018***	−0.015**	−0.013**	−0.010	−0.017***	−0.014**	−0.008	−0.006
	(0.005)	(0.006)	(0.006)	(0.007)	(0.006)	(0.007)	(0.007)	(0.008)
DEM 2 - Dementia patient care review (UL=60)	−0.049	−0.005	−0.056	−0.012	−0.058	−0.011	−0.074*	−0.026
	(0.035)	(0.036)	(0.036)	(0.037)	(0.038)	(0.043)	(0.040)	(0.045)
DM 13 - Microalbuminuria test in diabetes patients (UL=90)	−0.006**	−0.005*	−0.004*	−0.004	−0.007***	−0.005*	−0.007**	−0.005
	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
SMOKING 4 - Smoking cessation advice for smokers (UL=90)	−0.003	−0.004	−0.003	−0.005	−0.004	−0.007*	−0.005	−0.008*
	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)
STROKE 10 - Flu vaccine in stroke/TIA patients (UL=85)	0.007	0.004	0.008	0.005	0.008	0.003	0.011*	0.007
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)	(0.006)

Notes: Own calculations based on QOF data. Indicators presented in the table are those which were significant at the 90% level in at least one specification. All regressions include the number of items in which the PHC was in the bunching area $[UL, UL + 3]$ in the previous year as a control. Clustered at PHC-level standard errors in parenthesis. Significance: * 10%, ** 5%, *** 1%.

Table 4: Estimates of α_3. Heterogeneity according to the number of indicators that PHCs have at the bunching window $[UL, UL+3]$.

Indicator	(1) Main	(2)	(3)	(4)
		At the bunching region $[UL, UL+3]$ in at most		
		20 indicators	14 indicators	10 indicators
AF 3 - AF patients on anticoagulants/antiplatelets	-0.007^{**}	-0.010^{**}	-0.013^{*}	-0.013
	(0.003)	(0.004)	(0.007)	(0.012)
AF 4 - AF diagnosis confirmed by ECG/specialist	-0.016^{***}	-0.019^{***}	-0.023^{**}	0.005
	(0.006)	(0.007)	(0.010)	(0.023)
ASTHMA 6 - Asthma review in last 15 months	-0.012^{*}	-0.014^{*}	-0.013	-0.012
	(0.007)	(0.008)	(0.013)	(0.027)
ASTHMA 8 - Asthma diagnosis with variability test	-0.011	-0.017^{*}	-0.009	-0.054^{*}
	(0.008)	(0.009)	(0.015)	(0.030)
COPD 13 - COPD review with breathlessness assessment	-0.007	-0.009^{*}	-0.002	0.009
	(0.004)	(0.005)	(0.009)	(0.017)
CKD 6 - Albumin/creatinine ratio in CKD patients	-0.018^{***}	-0.019^{***}	-0.020^{*}	-0.042^{*}
	(0.005)	(0.006)	(0.011)	(0.023)
DEM 2 - Dementia patient care review	-0.049	-0.078^{**}	-0.103^{*}	-0.147^{*}
	(0.035)	(0.038)	(0.054)	(0.085)
DM 13 - Microalbuminuria test in diabetes patients	-0.006^{**}	-0.006^{**}	-0.011^{**}	-0.014
	(0.002)	(0.003)	(0.005)	(0.009)
SMOKING 4 - Smoking cessation advice for smokers	-0.003	-0.001	-0.004	0.011
	(0.003)	(0.004)	(0.006)	(0.012)
STROKE 10 - Flu vaccine in stroke/TIA patients	0.007	0.001	-0.014	-0.021
	(0.005)	(0.006)	(0.010)	(0.019)

Notes: Own calculations based on QOF data. Estimates obtained using regression (4), but with interactions between the covariates and the dummy variable $\mathbb{1}\{PB_i \geq s\}$ that indicates whether PHC i has at least s indicators in the bunching region, $[UL, UL+3]$, in 2010. Columns (2), (3), (4) report the estimates of α_3, without the interaction with $\mathbb{1}\{PB_i \geq s\}$, hence the coefficient corresponding to the PHCs with fewer than s indicators in the bunching window. All regressions include the number of indicators in which the PHC was in the bunching area $[UL, UL+3]$ in $t-1$ as a control. Standard errors, reported in parenthesis, are clustered at PHC level. Significance: * 10%, ** 5%, *** 1%.

Figure 1: Two-part contract for task 1

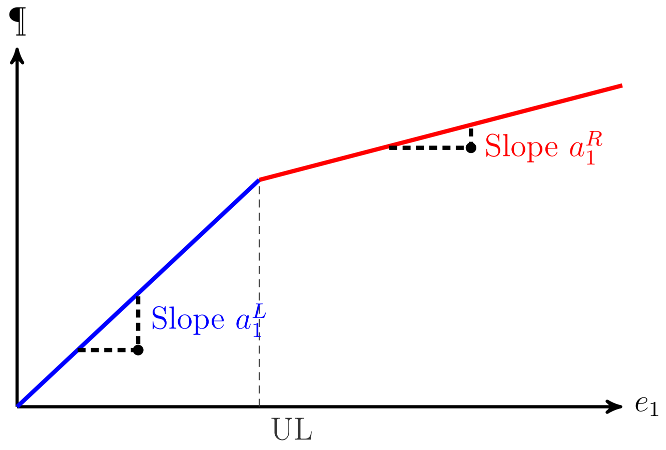
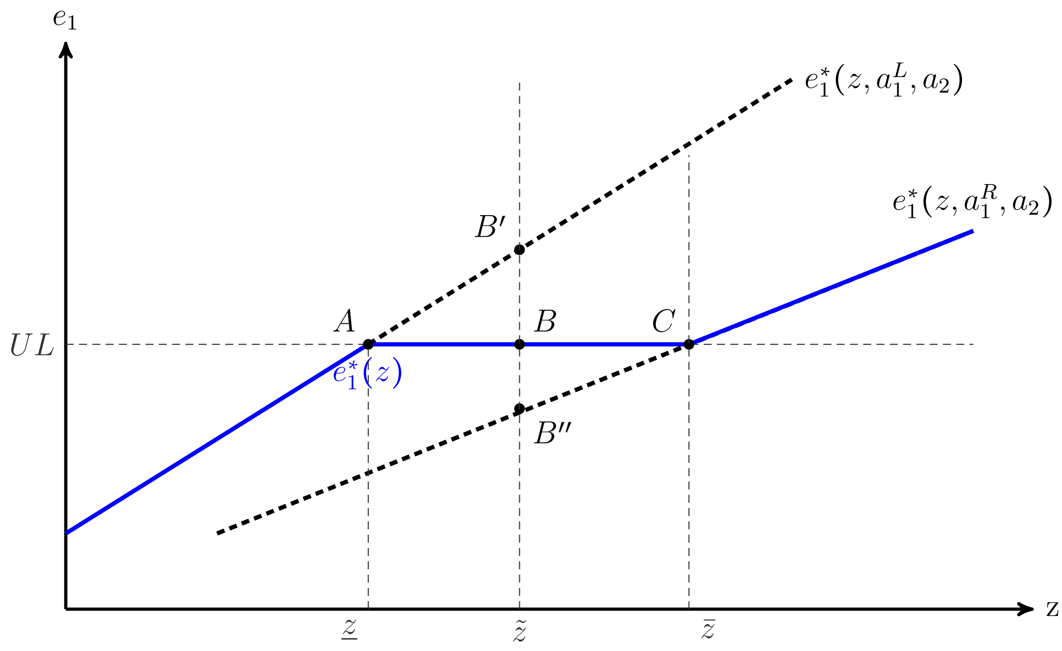
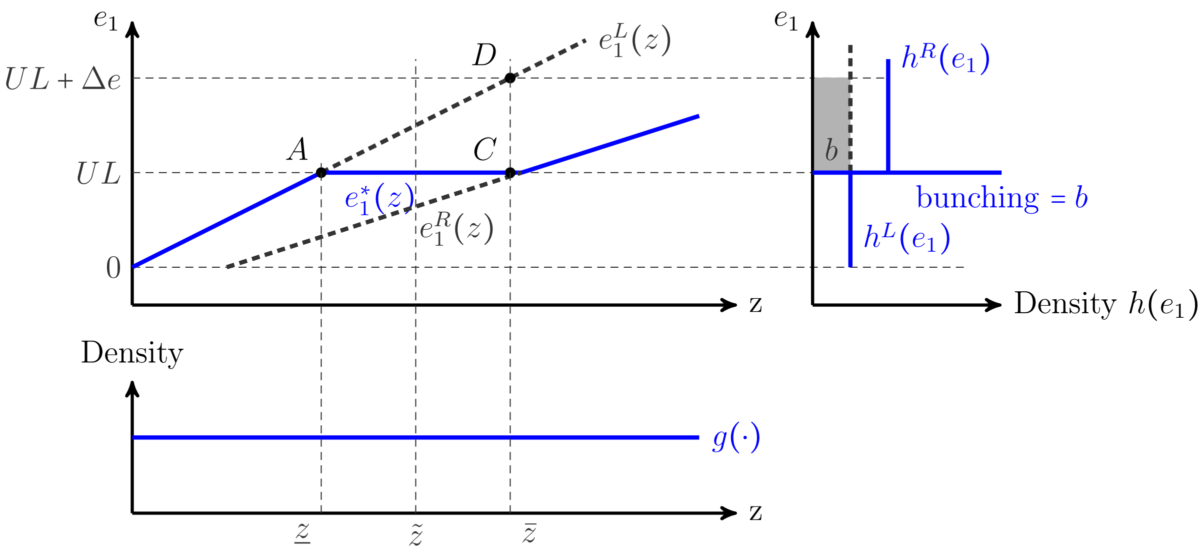


Figure 2: Agents' efficiency and task 1 optimal effort choice



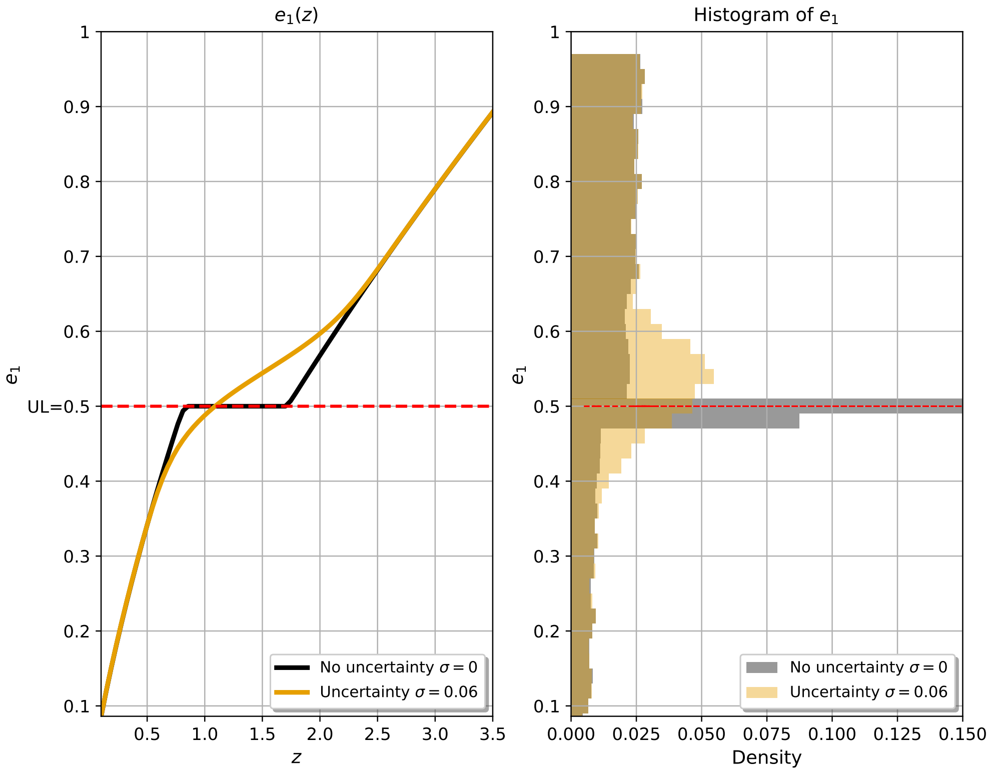
Note: Piece rate for e_1 is a_1^L for $e_1 \leq UL$, and it is a_1^L for $e_1 > UL$, $a_1^R < a_1^L$. The piecewise optimal effort function is $e_1^*(z) = e_1^L(z) \times \mathbb{1}(z < \underline{z}) + UL \times \mathbb{1}(\underline{z} \leq z \leq \bar{z}) + e_1^R(z) \times \mathbb{1}(z > \bar{z})$, where $e_1^R(z) = e_1^*(z, a_1^R, a_2)$ and $e_1^L(z) = e_1^*(z, a_1^L, a_2)$. This diagram assumes constant second derivatives of the function $C(e_1, e_2)$. It is also assumed that both tasks are substitutes, so the slope above UL is smaller than below it. In this graph, B' and B" would be the optimal level of $e_1^*(\tilde{z})$ if there were no kink, at the piece rates a_1^L and a_1^R respectively.

Figure 3: The effect on the density of e_1 of a kink on the payment function at $e_1 = UL$



Note: Piece rate for e_1 is a_1^L for $e_1 \leq UL$, and it is a_1^L for $e_1 > UL$, $a_1^R < a_1^L$. The piecewise optimal effort function $e_1^*(z) = e_1^L(z) \times \mathbb{1}(z < \underline{z}) + UL \times \mathbb{1}(\underline{z} \leq z \leq \bar{z}) + e_1^R(z) \times \mathbb{1}(z > \bar{z})$, where $e_1^R(z) = e_1^*(z, a_1^R, a_2)$ and $e_1^L(z) = e_1^*(z, a_1^L, a_2)$. This diagram assumes constant second derivatives of the function $C(e_1, e_2)$. It is also assumed that both tasks are substitutes, so the slope above UL is smaller than below it.

Figure 4: Implications of Uncertainty and Risk Aversion



Notes: The simulations above use a CRRA utility function (risk aversion given by η) for the provider using a quadratic cost function defined by parameters c_1, c_2 and δ. In the first scenario (black), there is no uncertainty ($\sigma = 0$). In the second (gold), uncertainty is allowed for achievement in task 1 where $\varepsilon_1 \sim N(0, \sigma)$. The agent's problem (see below) was solved numerically using a Gauss-Legendre approximation restricted to $\varepsilon_1 \in [-10\sigma, 10\sigma]$. For the purpose of the graph, 10.000 values for z were drawn from a uniform distribution between 0 and 2 and the resulting response functions $e_1^*(z)$ are plotted on the left, and its distribution on the right. The histogram was made over $e_1 \in [0, 0.98]$ and its x-axis censored at 0.07 for exposition purposes only. The threshold $UL = 0.5$ is used for the non-linearity of the problem:

$$\max_{e_1, e_2 \in [0,1]} U = E_{\varepsilon_1}\left[u\left(P(x_1, x_2; a_1^R, a_1^L, a_2, UL)\right) - C(e_1, e_2; z)\right]$$

$$s.t. \quad u(c) = \frac{(c+1)^{1-\eta}}{1-\eta}$$

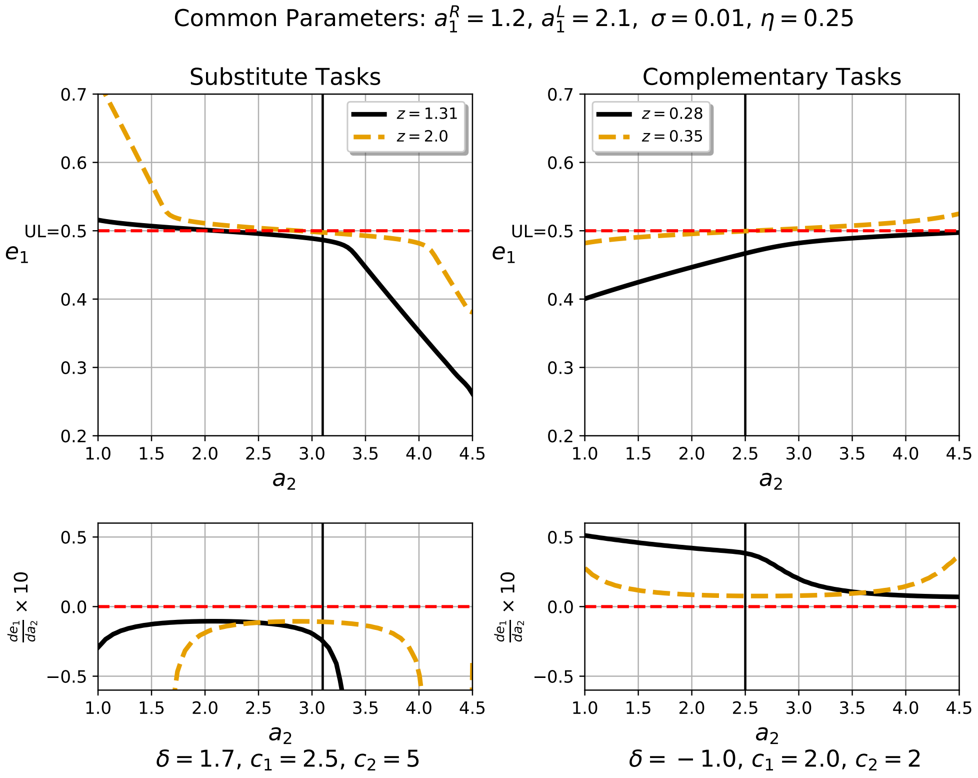
$$x_1 = e_1 + \varepsilon_1$$

$$x_2 = e_2$$

$$C(e_1, e_2; z) = \frac{1}{z} \cdot \left(\frac{1}{2} \cdot \left(c_1 \cdot e_1^2 + c_2 \cdot e_2^2\right) + \delta \cdot e_1 \cdot e_2\right)$$

Parameters in the simulation: $\delta = 1$, $c_1 = 2$, $c_2 = 7$, $a_1^R = 1$, $a_1^L = 1.9$, $a_2 = 1$, $\eta = 0.65$

Figure 5: Simulation exercise: e_1^* as a function of a_2

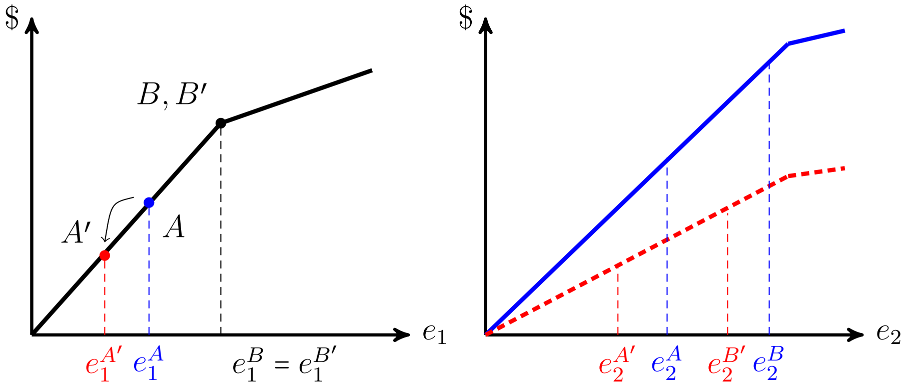


Notes: The simulations above use a CRRA utility function (risk aversion given by η) for the provider using a quadratic cost function defined by parameters c_1, c_2 and δ. Uncertainty is allowed for achievement in task 1 where $\varepsilon_1 \sim N(0, \sigma)$. Two scenarios are considered: substitute tasks on the left, and complementary tasks on the right. In both cases, we allowed for two levels of z. The upper panels present the response function, and the lower ones have its corresponding derivative. The vertical black lines (fixed value of a_2) are given as a reference for discussion.

The figure shows that for the case of substitute tasks, agents with $z = 1.31$ will be insensitive (bunched around UL) if $a_2 \in [1.4, 2.6]$, and those agents with $z = 2.0$ will be insensitive if $a_2 \in [2.1, 3.6]$. For the case of complements, agents with $z = 0.28$ would be bunched above 4.0 (we do not observe the upper limit in this simulation), while for $z = 0.35$ the bunching domain is $a_2 \in [1.5, 3.5]$.

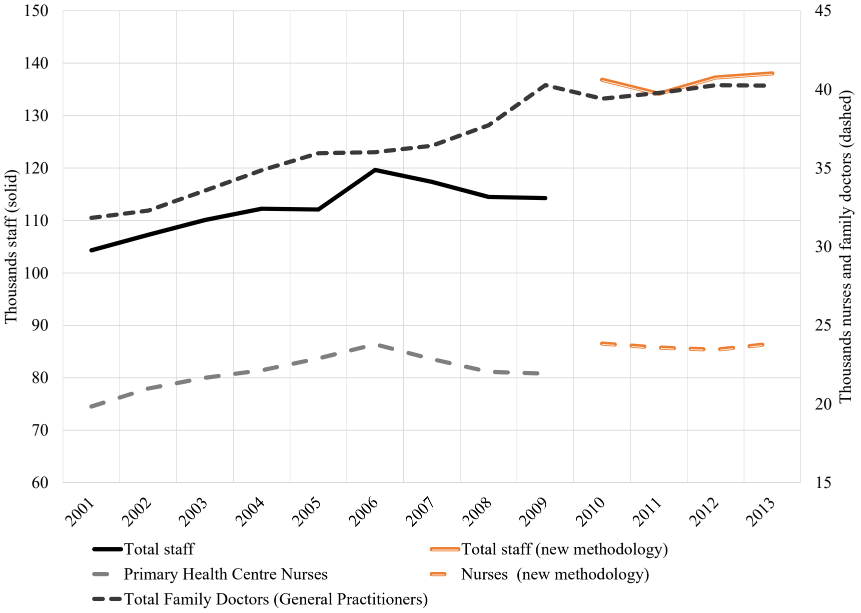
$$\begin{aligned}
\max_{e_1, e_2 \in [0,1]} \; & U = E_{\varepsilon_1}\left[u\left(P(x_1, x_2; a_1^R, a_1^L, a_2, UL)\right) - C(e_1, e_2; z)\right] \\
s.t. \quad & u(c) = \frac{(c+1)^{1-\eta}}{1-\eta} \\
& x_1 = e_1 + \varepsilon_1 \\
& x_2 = e_2 \\
& C(e_1, e_2; z) = \frac{1}{z} \cdot \left(\frac{1}{2} \cdot \left(c_1 \cdot e_1^2 + c_2 \cdot e_2^2\right) + \delta \cdot e_1 \cdot e_2\right)
\end{aligned}$$

Figure 6: The identification strategy: the case of complements


$
B, B'
A'
A
e_1
$e_1^{A'}$ e_1^A $e_1^B = e_1^{B'}$
$
e_2
$e_2^{A'}$ e_2^A $e_2^{B'}$ e_2^B

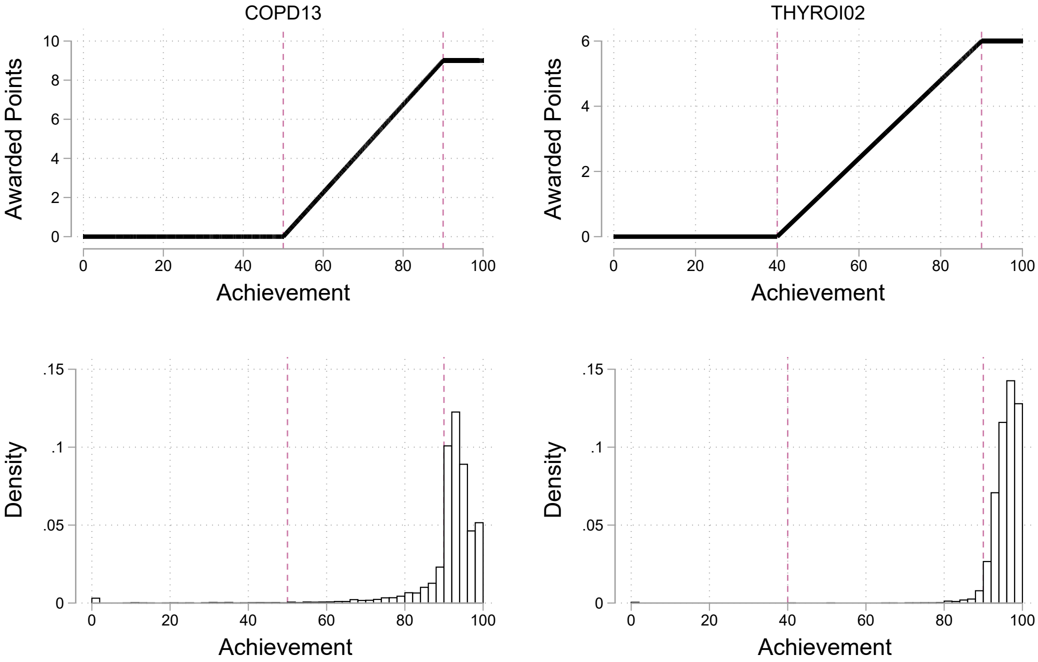
Notes: There are two tasks: 1 and 2. e_1 and e_2 represent the effort levels for tasks 1 and 2 respectively. There are two agents: A and B. In time t, the effort levels are (e_1^A, e_2^A) and (e_1^B, e_2^B) for agent A and B respectively, whilst they are $(e_1^{A'}, e_2^{A'})$ and $(e_1^{B'}, e_2^{B'})$ in time $t+1$. For task 1, the incentive contract is the same at time t and $t+1$. For task 2, the solid line (dashed line) represents the incentive contract in time t $(t+1)$.

Figure 7: Primary health centres workforce



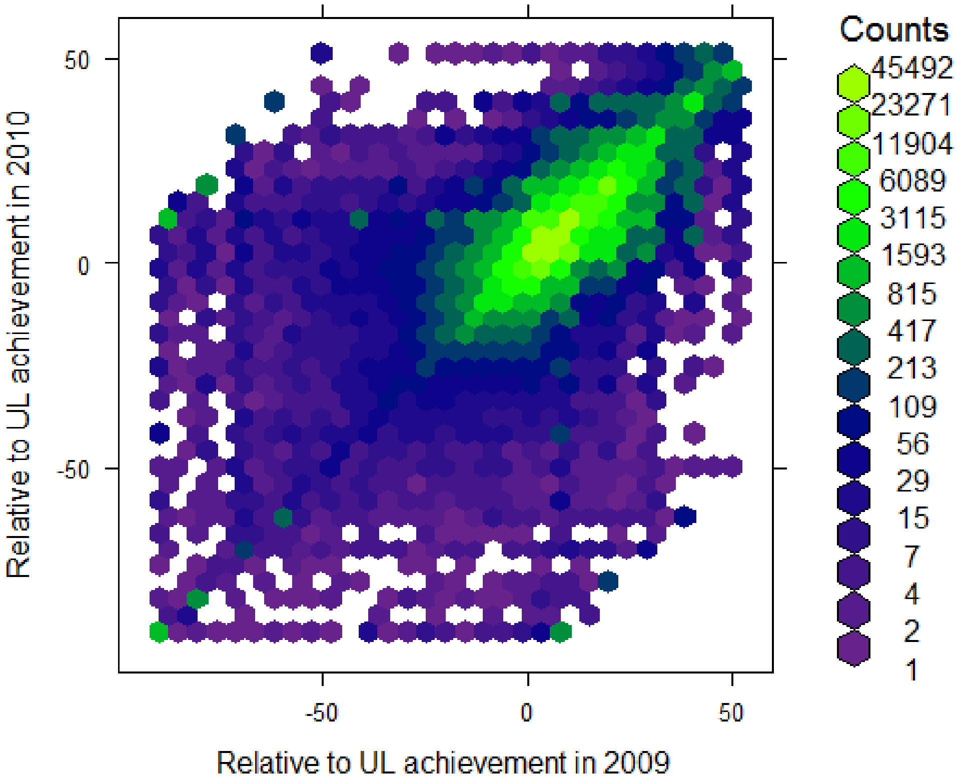
Notes: Data for Primary Health Centres (PHCs) from Table 1a of the General and Personal Medical Services, England - 2001-2011 and 2003-2013. Staff figures from 2010 onward follow a different methodology than prior years which affects comparison.

Figure 8: Points reward function and achievement density for THYROI02 and COPD13 for the 2010/11 financial year ($t = 2$)



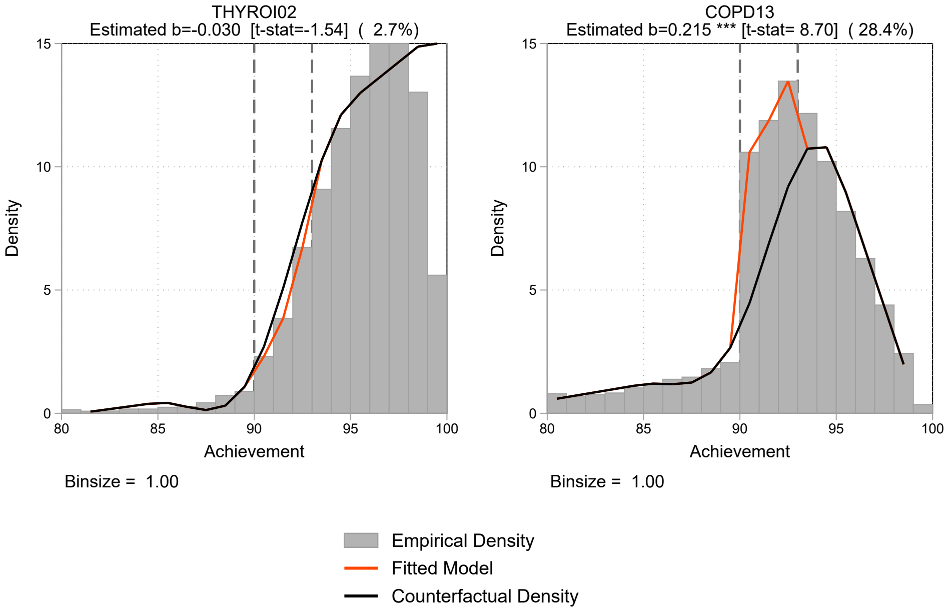
Notes: Own calculations based on QOF data archived at NHS Digital.

Figure 9: Heat map at indicator - Primary Health Centre level



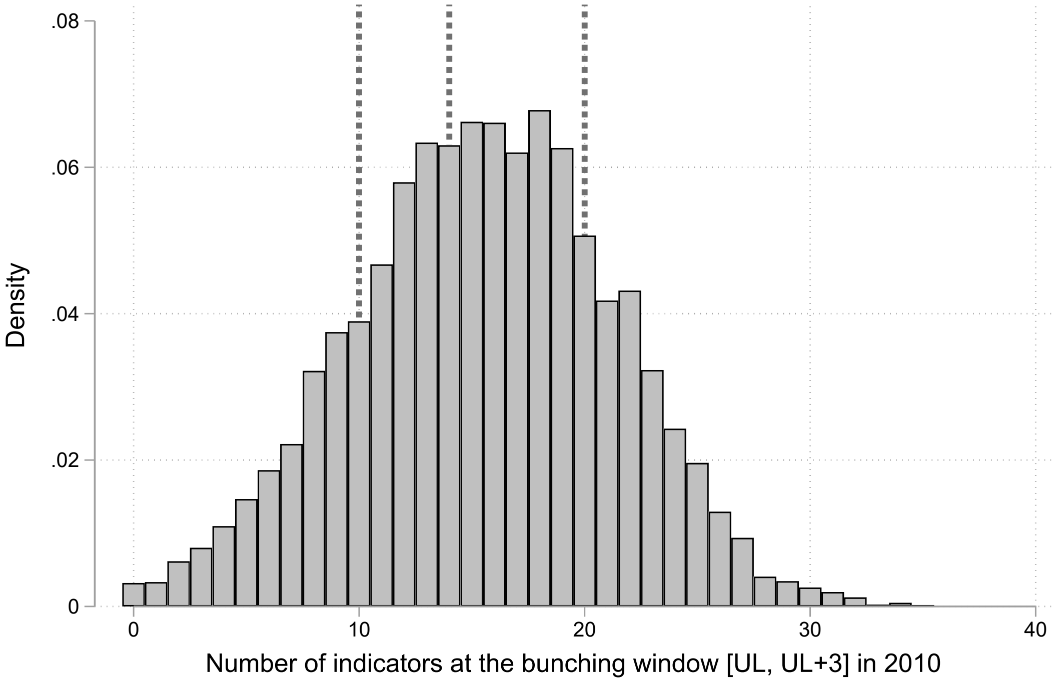
Notes: Own calculations based on QOF data archived at NHS Digital and computed using hexbin package for R (Carr and Pebesma, 2019). This two-dimensional histogram classifies the transitions of all clinical indicators per Primary Health Centre according to their relative performance against the upper-limit threshold of the payment function (UL). The x-axis presents achievement in 2009, and the y-axis the 2010 result. Each cell counts how many indicator-PHCs follow the specified transition, and the result is classified following the categories on the right of the graph, where the lighter colour represents a higher frequency.

Figure 10: Testing for bunching: THYROI02 and COPD13



Note: The empirical density is fitted with a restricted cubic spline based on 5 knots. Domain was restricted to a 10 pp. window around UL, and the excluded range is $[UL_j, UL_j + 3pp.]$ Equivalent figures for all other indicators are shown in Table E3 of Appendix E.

Figure 11: Distribution of PHCs according to the number of indicators at the bunching window [UL,UL+3]



Multitasking, Two-part Contracts, and Bunching: an Application to Doctors' Tasks.

Appendix - For Online Publication

February 19, 2025

A Proofs

Proof Proposition 2:

Following similar steps to those of Saez's (2010) proof in a taxation context,[1] we define the Cumulative Distribution Function (CDF) $H(\tilde{e}_1) = \Pr\left[e_1^*(z, a_1, a_2) \leq \tilde{e}_1\right] = \Pr\left[z \leq e_1^{*-1}\left(\tilde{e}_1; a_1, a_2\right)\right] = G\left[e_1^{*-1}\left(\tilde{e}_1; a_1, a_2\right)\right]$, where $e_1^{*-1}(\cdot)$ is the inverse function of $e_1^*(z)$. This CDF, $H(\tilde{e}_1)$, has a corresponding probability density function $h(\tilde{e}_1)$.

As explained above, the optimal task 1 effort, $e_1^*(z)$, is piecewise-defined, and consequently $H(\tilde{e}_1)$ also is. On the segment below UL, we have $H^L(\tilde{e}_1) = G\left[e_1^{*-1}\left(\tilde{e}_1; a_1^L, a_2\right)\right]$, and above it, the relevant function is $H^R(\tilde{e}_1) = G\left[e_1^{*-1}\left(\tilde{e}_1; a_1^R, a_2\right)\right]$. Given that all providers with a $\tilde{z} \in [\underline{z}, \bar{z}]$ choose $e_1^* = UL$, an entire mass that would have exerted an effort $e_1^L(\tilde{z}) > UL$ if there where no kink, is now collapsed at that single point and has a value of $b = h(UL) = H^L(e_1^L(\bar{z})) - H^L(UL)$. Above $\bar{z}$, the distribution follows $h^R(e_1)$.

Proof Proposition 3:

To prove (i), differentiate the First Order Conditions (Proposition 1, case (i) or case (iii)), while imposing $da_1 = 0$. This leads to $C_{11}de_1 + C_{12}de_2 = 0$ and $C_{12}de_1 + C_{22}de_2 - da_2 = 0$. Using both equations we find that

$$\frac{de_1}{da_2} = -\frac{C_{12}}{C_{11}C_{22} - C_{12}^2}, \tag{1}$$

where the sign of $\frac{de_1}{da_2}$ is the contrary to that of C_{12}, because the denominator is positive (for the Second Order Conditions of the maximization problem (1) to hold).

To prove (ii), the proof of case (ii) of Proposition 1 noted that for agents with productivity $\tilde{z} \in [\underline{z}, \bar{z}]$, the optimal choice is $e_1^* = UL$ because $a_1^R < C_1(e_1^*, e_2^*) < a_1^L$. A small enough change in a_2, will also lead to a small change in $C_1(e_1^*, e_2^*)$ but $C_1(e_1^*, e_2^*)$ will still be within the (a_1^R, a_1^L) interval (provided the change in a_2 is small enough), and hence the optimal choice of task 1 effort

[1] See Kleven (2016) for an intuitive explanation of why bunching arises at income distribution as a result of the presence of kinks on income tax schedules.

will still be $e_1^* = UL$.

B QOF Payment

This Appendix describes how the number of points that a PHC obtains in each QOF indicator translated into income for the PHC.

The formulae below describes how the income for indicator j is computed for PHC i. The achievement of PHC i on indicator j is measured by the ratio x_{ij}, whose numerator is the number of enrollees in PHC i whose clinical management fulfils the definition of indicator j, and the denominator is the number of enrollees who should fullfil it. The achievement ratio, x_{ij}, is translated into points, $Points_{ij}$, using the non-linear scheme (2), which we also described in section 4.2 of the main text: zero points are awarded if x_{ij} is below the lower limit (LL_j), the maximum number of available points is awarded if x_{ij} goes above the upper limit (UL_j), and a linear scheme is used between the lower and the upper limit.

The number of points that the PHC i obtains from indicator j, $Points_{ij}$, is then adjusted by two scaling factors: $Size_i$ and $RelPrev_{ij}$. The former takes into account the relative size of the PHC and it is computed as the number of enrollees in the PHC divided by 5891, which was the average number of enrollees per PHC in 2003.[2] The scaling factor $RelPrev_{ij}$ measures the relative prevalence in PHC i with respect to the national prevalence of the illness to which indicator j refers to. The income that PHC i obtains from indicator j, $Payment_{ij}$, is then computed multiplying the number of points, $Points_{ij}$ (scaled by $Size_i$ and $RelPrev_{ij}$) by a monetary value per point, which is constant across indicators and PHCs.

[2] Since 2013 this figure has been updated annually. More details are available from BMA (2013).

$$x_{ij} = \frac{\text{Enrollees in PHC } i \text{ whose clinical management fulfils the definition of indicator } j}{\text{Enrollees in PHC } i \text{ whose clinical management should fulfil the definition of indicator } j}$$

$$Points_{ij} = \begin{cases} 0 & \text{if } x_{ij} \leq LL_j \\ (x_{ij} - LL_j) \cdot \frac{\text{Max. Avail. Points}_j}{UL_j - LL_j} & \text{if } LL_j < x_{ij} < UL_j \\ \text{Max. Avail. Points}_j & \text{if } x_{ij} \geq UL_j \end{cases} \tag{2}$$

$$Size_i = \frac{\text{Enrollees in PHC } i}{5891}$$

$$RelPrev_{ij} = \frac{\text{Prevalence in PHC } i \text{ of the illness to which indicator } j \text{ refers to}}{\text{National Prevalence of the illness to which indicator } j \text{ refers to}}$$

$$Payment_{ij} = (\text{Value per point in £}) \cdot Points_{ij} \cdot Size_i \cdot RelPrev_{ij} \tag{3}$$

C Definition of all indicators and its changes

Table C1 of this Appendix describes the summary of the changes to the 2011 QOF clinical indicators, with respect to the 2010 and 2009 ones. Table C2 provides the detailed definition of the QOF indicators that did not change, and Table C3 provides the definition of the indicators that changed and a detailed description of how they changed.

Table C1: Changes in QOF 2011 clinical indicators with respect to 2009-2010

Change	**Description**	**Indicators**	**Equivalent Effect on the Piece Rate**	**Maximum Available Points**
Withdrawn	These tasks are not rewarded anymore. Clinical indicators are about having a recent record of certain physical measures, or reviews.	CHD5, CHD7, DM5, DM11, DM16, EPILEPSY7, MH7, STROKE5	Reduction	32
Points reduced	The number of assigned points per indicator was reduced.†	BP4, DEP1	Reduction	26 to 22
Upper Limit Increased	Small increase from 70% to 71%. ♠	CHD6, STROKE6	Reduction	22
Replacement I	For indicators PP01, MH04, MH05, the time for accomplishing a given goal was reduced. For CHD2, the optional specialist referral was made compulsory.	PP01, MH04, MH05, CHD2	Reduction	18
Replacement II	Decrease in points and new wording is more precise and requires actions at the moment of diagnosis instead of treatment starting point.	DEP2, DEP3	Reduction	45 to 25
Replacement III	Most of these indicators were replaced by versions which are harder to accomplish. In a few of them this was compensated with extra points, but in some others there was a reduction as well: • For CHD11/CHD14 there is an increase from 7 to 10 points in exchange for prescribing aspirin and statins on top of an ACE inhibitor or alternative blood pressure treatments. • Requirements for DM9 were increased from checking peripheral pulses to a more comprehensive foot examination. It was also increased from 3 to 4 points. • Indicator DM12 was split into DM30 and DM31, keeping the same number of points. It asked for a percentage of patients below a given blood pressure target (145/85). It was replaced by two targets, one slightly below the original (140/80), and one notoriously above (150/90). • Indicator MH09 was split into MH11, MH12, MH13, MH14, MH15 and MH16. It moved from 23 to 27 points. The original indicator was general and imprecise ("routine health promotion and prevention advice appropriate to their age and health status"), while the replacements ask for specific measurements depending on age and gender.	CHD11/CHD14, DM9, DM12 (DM30,DM31), MH09 (MH11, MH12, MH13, MH14, MH15 and MH16)	Ambiguous	51 to 59

Continued on next page

Table C1: Detailed Changes in QOF 2011 clinical indicators with respect to 2009-2010 (Continued)

Change	**Description**	**Indicators**	**Equivalent Effect on the Piece Rate**	**Maximum Available Points**
Replacement IV	The cutoff was relaxed from last HbA1C to be 7% or less, to HbA1C to be 7.5% or less	DM23/DM26	Increase	17
Replacement V	Similar or the same wording, but the recoding was done in order to highlight recent changes in diagnostic procedures. For diabetes indicators the wording is explicit about new measurement standards.	COPD1/COPD14, COPD12/COPD15, MH6/MH10, DM24/DM27, DM25/DM28	-	32
New	These are tasks that were not considered before. Three new clinical indicators, on dementia, epilepsy and learning disabilities.	DEM3, EPILEPSY 9, LD2	Increase	12
Unchanged	No change on points, thresholds or wording		-	454

Note: This corresponds to our interpretation based on NHS Employers public documents.

Table C2 below provides the exact definition of all QOF indicators that did not change between 2009 and 2011. The indicators with a valid entry in the LL and UL columns are those which follow the non-linear reward scheme which we explain in subsection 4.2 of the text, as well as Appendix A. The LL and UL entries refer to the lower and upper limit (in points) of the non-linear reward scheme respectively. The indicators without valid entries in the LL and UL columns are *bonus* type indicators for which either the full amount of points is awarded or none.

Table C2: Indicators without changes from 2009 to 2011 (1/5)

Indicator Domain	**Description**	**Max. Avail. Points**	**LL**	**UL**
CLINICAL	**AF 1:** The PHC can produce a register of patients with Atrial Fibrillation	5	-	-
CLINICAL	**AF 3:** The percentage of patients with atrial fibrillation who are currently treated with anti-coagulant drug therapy or an anti-platelet drug therapy	12	40	90
CLINICAL	**AF 4:** The percentage of patients with atrial fibrillation diagnosed after 1st April 2008 with ECG or specialist confirmed diagnosis	10	40	90
CLINICAL	**ASTHMA 1:** The PHC can produce a register of patients with asthma, excluding patients with asthma who have been prescribed no asthma-related drugs in the previous twelve months	4	-	-
CLINICAL	**ASTHMA 3:** The percentage of patients with asthma between the ages of 14 and 19 in whom there is a record of smoking status in the previous 15 months	6	40	80
CLINICAL	**ASTHMA 6:** The percentage of patients with asthma who have had an asthma review in the previous 15 months	20	40	70
CLINICAL	**ASTHMA 8:** The percentage of patients aged eight and over diagnosed as having asthma from 1st April 2007 with measures of variability or reversibility	15	40	80
CLINICAL	**BP 1:** The PHC can produce a register of patients with established hypertension	6	-	-
CLINICAL	**BP 5:** The percentage of patients with hypertension in whom the last blood pressure (measured in the previous 9 months) is 150/90 or less	57	40	70
CLINICAL	**CANCER 1:** The PHC can produce a register of all cancer patients defined as a 'register of patients with a diagnosis of cancer excluding non-melanotic skin cancers from 1 April 2003'	5	-	-
CLINICAL	**CANCER 3:** The percentage of patients with cancer, diagnosed within the last 18 months, who have a patient review recorded as occurring at 6 months after the PHC has received confirmation of the diagnosis	6	40	90
CLINICAL	**CHD 1:** The PHC can produce a register of patients with coronary heart disease	4	-	-
CLINICAL	**CHD 8:** The percentage of patients with coronary heart disease whose last measured total cholesterol (measured in the previous 15 months) is 5 mmol/l or less	17	40	70
CLINICAL	**CHD 9:** The percentage of patients with coronary heart disease with a record in the previous 15 months that aspirin, an alternative anti-platelet therapy, or an anti-coagulant is being taken (unless a contraindication or side-effects are recorded)	7	40	90
CLINICAL	**CHD 10:** The percentage of patients with coronary heart disease who are currently treated with a beta blocker (unless a contraindication or side-effects are recorded)	7	40	60
CLINICAL	**CHD 12:** The percentage of patients with coronary heart disease who have a record of influenza immunisation in the preceding 1 September to 31 March	7	40	90
CLINICAL	**CKD 1:** The PHC can produce a register of patients aged 18 years and over with ChKD. (US National Kidney Foundation: Stage 3-5 CKD)	6	-	-
CLINICAL	**CKD 2:** The percentage of patients on the CKD register whose notes have a record of blood pressure in the previous 15 months	6	40	90
CLINICAL	**CKD 3:** The percentage of patients on the CKD register in whom the last blood pressure reading, measured in the previous 15 months, is 140/85 or less	11	40	70
CLINICAL	**CKD 5:** The percentage of patients on the CKD register with hypertension and proteinuraa who are treated with an angiotensin converting enzyme inhibitor (ACE-I) or angiotensin receptor blocker (ARB) (unless a contraindication or side effects are recorded)	9	40	80
CLINICAL	**CKD 6:** The percentage of patients on the CKD register whose notes have a record of an albumin:creatinine ratio (or protein:creatinine ratio) value in the previous 15 months	6	40	80
CLINICAL	**CVD 2:** The percentage of people diagnosed with hypertension diagnosed after 1 April 2009 who are given lifestyle advice in the last 15 months for: increasing physical activity, smoking cessation, safe alcohol consumption and healthy diet	5	40	70
CLINICAL	**COPD 8:** The percentage of patients with COPD who have had influenza immunisation in the preceding 1 September to 31 March	6	40	85

Continued on next page

Table C2: Indicators without changes from 2009 to 2011 (2/5)

Indicator Domain	Description	Max. Avail. Points	LL	UL
CLINICAL	**COPD 10:** The percentage of patients with COPD with a record of FeV1 in the previous 15 months	7	40	70
CLINICAL	**COPD 13:** The percentage of patients with COPD who have had a review, undertaken by a healthcare professional, including an assessment of breathlessness using the MRC dyspnoea score in the preceding 15 months	9	50	90
CLINICAL	**DEM 1:** The PHC can produce a register of patients diagnosed with dementia	5	-	-
CLINICAL	**DEM 2:** The percentage of patients diagnosed with dementia whose care has been reviewed in the previous 15 months	15	25	60
CLINICAL	**DM 2:** The percentage of patients with diabetes whose notes record BMI in the previous 15 months	3	40	90
CLINICAL	**DM 10:** The percentage of patients with diabetes with a record of neuropathy testing in the previous 15 months	3	40	90
CLINICAL	**DM 13:** The percentage of patients with diabetes who have a record of micro-albuminuria testing in the previous 15 months (exception reporting for patients with proteinuria)	3	40	90
CLINICAL	**DM 15:** The percentage of patients with diabetes with proteinuria or micro-albuminuria who are treated with ACE inhibitors (or A2 antagonists)	3	40	80
CLINICAL	**DM 17:** The percentage of patients with diabetes whose last measured total cholesterol within the previous 15 months is 5mmol/l or less	6	40	70
CLINICAL	**DM 18:** The percentage of patients with diabetes who have had influenza immunisation in the preceding 1 September to 31 March	3	40	85
CLINICAL	**DM 19:** The PHC 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	6	-	-
CLINICAL	**DM 21:** The percentage of patients with diabetes who have a record of retinal screening in the previous 15 months	5	40	90
CLINICAL	**DM 22:** The percentage of patients with diabetes who have a record of estimated glomerular filtration rate (eGFR) or serum creatinine testing in the previous 15 months	3	40	90
CLINICAL	**EPILEPSY 5:** The PHC can produce a register of patients aged 18 years and over receiving drug treatment for epilepsy	1	-	-
CLINICAL	**EPILEPSY 6:** The percentage of patients aged 18 years and over on drug treatment for epilepsy who have a record of seizure frequency in the previous 15 months	4	40	90
CLINICAL	**EPILEPSY 8:** The percentage of patients aged 18 years and over on drug treatment for epilepsy who have been seizure free for the last 12 months recorded in the previous 15 months	6	40	70
CLINICAL	**HF 1:** The PHC can produce a register of patients with heart failure	4	-	-
CLINICAL	**HF 2:** The percentage of patients with a diagnosis of heart failure (diagnosed after the 1st April 2006) which has been confirmed by an echocardiogram or by specialist assessment	6	40	90
CLINICAL	**HF 3:** The percentage of patients with a current diagnosis of heart failure due to LVD who are currently treated with an ACE inhibitor or Angiotensin Receptor Blocker (unless a contraindication or side effects are recorded)	10	40	80
CLINICAL	**HF 4:** The percentage of patients with a current diagnosis of heart failure due to LVD who are currently treated with an ACE inhibitor or Angiotensin Receptor Blocker, who are additionally treated with a beta-blocker licensed for heart failure, or recorded as intolerant to or having a contraindication to beta-blockers	9	40	60
CLINICAL	**LD 1:** The PHC can produce a register of patients with learning disabilities	4	-	-
CLINICAL	**MH 8:** The PHC can produce a register of people with schizophrenia, bipolar affective disorder and other psychoses	4	-	-
CLINICAL	**OB 1:** The PHC can produce a register of patients aged 16 years and over with a BMI greater than or equal to 30 in the last 15 months	8	-	-
CLINICAL	**PC 2:** The PHC has regular (at least 3 monthly) multidisciplinary case review meetings where all patients on the palliative care register are discussed	3	-	-
CLINICAL	**PC 3:** The PHC has a complete register of all patients in need of palliative/supportive irrespective of age	3	-	-

Continued on next page

Table C2: Indicators without changes from 2009 to 2011 (3/5)

Indicator Domain	**Description**	**Max. Avail. Points**	**LL**	**UL**
CLINICAL	**SMOKING 3:** The percentage of patients with any or any combination of the following conditions: coronary heart disease, stroke or TIA, hypertension, diabetes, COPD, CKD, asthma, schizophrenia, bipolar affective disorder or other psychoses whose notes record smoking status in the previous 15 months (except those who have never smoked where smoking status need only be recorded once since diagnosis)	30	40	90
CLINICAL	**SMOKING 4:** The percentage of patients with any or any combination of the following conditions: coronary heart disease, stroke or TIA, hypertension, diabetes, COPD, CKD, asthma, schizophrenia, bipolar affective disorder or other psychoses who smoke whose notes contain a record that smoking cessation advice or referral to a specialist service, where available, has been offered within the previous 15 months	30	40	90
CLINICAL	**STROKE 1:** The PHC can produce a register of patients with Stroke or TIA	2	-	-
CLINICAL	**STROKE 7:** The percentage of patients with TIA or stroke who have a record of total cholesterol in the previous 15 months	2	40	90
CLINICAL	**STROKE 8:** The percentage of patients with TIA or stroke whose last measured total cholesterol (measured in the previous 15 months) is 5 mmol/l or less	5	40	60
CLINICAL	**STROKE 10:** The percentage of patients with TIA or stroke who have had influenza immunisation in the preceding 1 September to 31 March	2	40	85
CLINICAL	**STROKE 12:** The percentage of patients with a stroke shown to be non-haemorrhagic, or a history of TIA, who have a record that an anti-platelet agent (aspirin, clopidogrel, dipyridamole or a combination), or an anti-coagulant is being taken (unless a contraindication or side-effects are recorded)	4	40	90
CLINICAL	**STROKE 13:** The percentage of new patients with a stroke or TIA who have been referred for further investigation	2	40	80
CLINICAL	**THYROID 1:** The PHC can produce a register of patients with hypothyroidism	1	-	-
CLINICAL	**THYROID 2:** The percentage of patients with hypothyroidism with thyroid function tests recorded in the previous 15 months	6	40	90
SERVICES	**CHS 1:** Child development checks are offered at intervals that are consistent with national guidelines and policy	6	-	-
SERVICES	**CS 1:** The percentage of patients aged from 25 to 64 (in Scotland from 21 to 60) whose notes record that a cervical smear has been performed in the last five years Standard 40 - 80%	11	40	80
SERVICES	**CS 5:** The PHC has a system for informing all women of the results of cervical smears	2	-	-
SERVICES	**CS 6:** The PHC has a policy for auditing its cervical screening service, and performs an audit of inadequate cervical smears in relation to individual smear takers at least every two years	2	-	-
SERVICES	**CS 7:** The PHC has a protocol that is in line with national guidance and PHC for the management of cervical screening, which includes staff training, management of patient call/recall, exception reporting and the regular monitoring of inadequate smear rates	7	-	-
SERVICES	**MAT 1:** Ante-natal care and screening are offered according to current local guidelines	6	-	-
SERVICES	**SH 1:** The PHC can produce a register of women who have been prescribed any method of contraception at least once in the last year	4	-	-
SERVICES	**SH 2:** The percentage of women prescribed an oral or patch contraceptive method in the last year who have received information from the PHC about long acting reversible methods of contraception in the previous 15 months	3	40	90
SERVICES	**SH 3:** The percentage of women prescribed emergency hormonal contraception at least once in the year by the PHC who have received information from the PHC about long acting reversible methods of contraception at the time of, or within one month or, the prescription	3	40	90

Continued on next page

Table C2: Indicators without changes from 2009 to 2011 (4/5)

Indicator Domain	**Description**	**Max. Avail. Points**	**LL**	**UL**
ORGANISATIONAL	**INFORMATION 5:** The PHC supports smokers in stopping smoking by a strategy, which includes providing literature and offering appropriate therapy	2	-	-
ORGANISATIONAL	**EDUCATION 1:** There is a record of all PHC-employed clinical staff having attended training/ updating in basic life-support skills in the preceding 18 months	4	-	-
ORGANISATIONAL	**EDUCATION 5:** There is a record of all PHC-employed staff having attended training/ updating in basic life support skills in the preceding 36 months	3	-	-
ORGANISATIONAL	**EDUCATION 6:** The PHC conducts an annual review of patient complaints and suggestions to ascertain general learning points which are shared with the team	3	-	-
ORGANISATIONAL	**EDUCATION 7:** The PHC has undertaken a minimum of twelve significant event reviews in the past 3 years which could include: Any death occurring on the PHC premises; New cancer diagnoses; Deaths where terminal care has taken place at home; Any suicides; Sections under the Mental Health Act; Child protection cases; Medication errors; A significant event occuring when a patient may have been subjected to harm, had the circumstance/outcome been different (near miss)	4	-	-
ORGANISATIONAL	**EDUCATION 8:** All PHC-employed nurses have personal learning plans which have been reviewed at annual appraisal	5	-	-
ORGANISATIONAL	**EDUCATION 9:** All PHC-employed non-clinical team members have an annual appraisal	3	-	-
ORGANISATIONAL	**EDUCATION 10:** The PHC has undertaken a minimum of three significant event reviews within the last year	6	-	-
ORGANISATIONAL	**MANAGEMENT 1:** Individual healthcare professionals have access to information on local procedures relating to child protection	1	-	-
ORGANISATIONAL	**MANAGEMENT 2:** There are clearly defined arrangements for backing up computer data, back-up verification, safe storage of back-up tapes and authorisation for loading programmes where a computer is used	1	-	-
ORGANISATIONAL	**MANAGEMENT 3:** The Hepatitis B status of all doctors and relevant PHC employed staff is recorded and immunisation recommended if required in accordance with national guidance	0.5	-	-
ORGANISATIONAL	**MANAGEMENT 5:** The PHC offers a range of appointment times to patients which as a minimum should include morning and afternoon appointments five mornings and four afternoons per week except where agreed with the PCO	3	-	-
ORGANISATIONAL	**MANAGEMENT 7:** The PHC has systems in place to ensure regular and appropriate inspection, calibration, maintenance and replacement of equipment including: a defined responsible person; clear recording; systematic pre-planned schedules; reporting of faults	3	-	-
ORGANISATIONAL	**MANAGEMENT 9:** The PHC has a protocol for the identification of carers and a mechanism for the referral of carers for social services assessment	3	-	-
ORGANISATIONAL	**MANAGEMENT 10:** There is a written procedure manual that includes staff employment policies including equal opportunities, bullying and harassment and sickness absence (including illegal drugs, alcohol and stress) to which staff have access	2	-	-
ORGANISATIONAL	**MEDICINES 2:** The PHC possesses the equipment and up-to-date emergency drugs to treat anaphylaxis	2	-	-
ORGANISATIONAL	**MEDICINES 3:** There is a system for checking expiry dates of emergency drugs at least on an annual basis	2	-	-
ORGANISATIONAL	**MEDICINES 4:** The number of hours from requesting a prescription to availability for collection by the patient is 72 hours or less (excluding weekends and bank/local holidays)	3	-	-
ORGANISATIONAL	**MEDICINES 6:** The PHC meets with the PCO prescribing adviser at least annually and agrees up to three actions related to prescribing	4	-	-
ORGANISATIONAL	**MEDICINES 8:** The number of hours from requesting a prescription to availability for collection by the patient is 48 hours or less (excluding weekends and bank/local holidays)	6	-	-
ORGANISATIONAL	**MEDICINES 10:** The PHC meets with the PCO prescribing adviser at least annually, has agreed up to three actions related to prescribing and subsequently provided evidence of change	4	-	-

Continued on next page

Table C2: Indicators without changes from 2009 to 2011 (5/5)

Indicator Domain	**Description**	**Max. Avail. Points**	**LL**	**UL**
ORGANISATIONAL	**MEDICINES 11:** A medication review is recorded in the notes in the preceding 15 months for all patients being prescribed four or more repeat medicines Standard 80%	7	-	-
ORGANISATIONAL	**MEDICINES 12:** A medication review is recorded in the notes in the preceding 15 months for all patients being prescribed repeat medicines Standard 80%	8	-	-
ORGANISATIONAL	**RECORDS 3:** The PHC has a system for transferring and acting on information about patients seen by other doctors out of hours	1	-	-
ORGANISATIONAL	**RECORDS 8:** There is a designated place for the recording of drug allergies and adverse reactions in the notes and these are clearly recorded	1	-	-
ORGANISATIONAL	**RECORDS 9:** For repeat medicines, an indication for the drug can be identified in the records (for drugs added to repeat prescription with effect from 1st April 2004). Minimum standard 80 per cent	4	-	-
ORGANISATIONAL	**RECORDS 11:** The blood pressure of patients aged 45 and over is recorded in the preceding 5 years for at least 65% of patients	10	-	-
ORGANISATIONAL	**RECORDS 13:** There is a system to alert the out-of-hours service or duty doctor to patients dying at home	2	-	-
ORGANISATIONAL	**RECORDS 15:** The PHC has up-to-date clinical summaries in at least 60 per cent of patient records	25	-	-
ORGANISATIONAL	**RECORDS 17:** The blood pressure of patients aged 45 and over is recorded in the preceding 5 years for at least 80% of patients	5	-	-
ORGANISATIONAL	**RECORDS 18:** The PHC has up-to-date clinical summaries in at least 80 per cent of patient records	8	-	-
ORGANISATIONAL	**RECORDS 19:** 80 per cent of newly registered patients have had their notes summarised within eight weeks of receipt by the PHC	7	-	-
ORGANISATIONAL	**RECORDS 20:** The PHC has up-to-date clinical summaries in at least 70% of patient records	12	-	-
ORGANISATIONAL	**RECORDS 23:** The percentage of patients aged over 15 whose notes record smoking status in the past 27 months	11	40	90
PATIENT EXPERIENCE	**PE 1:** The length of routine booked appointments with the doctors in the PHC is not less than 10 minutes. [If the PHC routinely sees extras during booked surgeries, then the average booked consultation length should allow for the average number of extras seen in a surgery session. If the extras are seen at the end, then it is not necessary to make this adjustment.] For PHCs with only an open surgery system, the average face to face time spent by the GP with the patient is at least 8 minutes. For PHCs that routinely operate a mixed economy of booked and open surgeries should report on both criteria.	33	-	-

Table C3 below provides the definition of the QOF indicators which changed between 2011 and 2010-2009, as well as a detailed description of the change indicator per indicator. For those indicators which were replaced, we categorise them in Replacement I, Replacement II, etc. as we did in Table C1 of this Appendix, which summarised the changes. As in Table C2 of this Appendix, the indicators with a valid entry in the LL and UL columns are those which follow the non-linear reward scheme which we explain in subsection 4.2 of the text, as well as Appendix A. The LL and UL entries refer to the lower and upper limit (in points) of the non-linear reward scheme respectively. The indicators without valid entries in the LL and UL columns are *bonus* type indicators for which either the full amount of points is awarded or none.

Table C3: Indicators with changes from 2010 to 2011 (1/5)

Change	**Indicator Domain**	**Description**	**Max. Avail. Points**	**LL**	**UL**
Withdrawn	CLINICAL	**CHD 5:** The percentage of patients with coronary heart disease whose notes have a record of blood pressure in the previous 15 months	7	40	90
Withdrawn	CLINICAL	**CHD 7:** The percentage of patients with coronary heart disease whose notes have a record of total cholesterol in the previous 15 months	7	40	90
Withdrawn	CLINICAL	**DM 5:** The percentage of patients with diabetes who have a record of HbA1c or equivalent in the previous 15 months	3	40	90
Withdrawn	CLINICAL	**DM 11:** The percentage of patients with diabetes who have a record of the blood pressure in the previous 15 months	3	40	90
Withdrawn	CLINICAL	**DM 16:** The percentage of patients with diabetes who have a record of total cholesterol in the previous 15 months	3	40	90
Withdrawn	CLINICAL	**EPILEPSY 7:** The percentage of patients aged 18 years and over on drug treatment for epilepsy who have a record of medication review involving the patient or carer in the previous 15 months	4	40	90
Withdrawn	CLINICAL	**MH 7:** The percentage of patients with schizophrenia, bipolar affective disorder and other psychoses who do not attend the PHC for their annual review who are identified and followed up by PHC team within 14 days of non attendance	3	40	90
Withdrawn	CLINICAL	**STROKE 5:** The percentage of patients with TIA or stroke who have a record of blood pressure in the notes in the preceding 15 months	2	40	90
Reduced Points	CLINICAL	**BP 4:** The percentage of patients with hypertension in whom there is a record of the blood pressure in the previous 9 months	18 → 16	40	90
Reduced Points	CLINICAL	**DEP 1:** The percentage of patients with diabetes and/or heart disease for whom case finding for depression has been undertaken on one occasion during the previous 15 months using the two standard screening questions	8 → 6	40	90
Increased UL	CLINICAL	**CHD 6:** The percentage of patients with coronary heart disease in whom the last blood pressure reading (measured in the previous 15 months) is 150/90 or less	17	40	70 → 71
Increased UL	CLINICAL	**STROKE 6:** The percentage of patients with a history of TIA or stroke in whom the last blood pressure reading (measured in the previous 15 months) is 150/90 or less	5	40	70 → 71
New	CLINICAL	**DEM 3:** The percentage of patient with a new diagnosis of dementia from April 2011 to have FBC, calcium, glucose, renal and liver function, thyroid function tests, serum vitamin B12 and folate levels recorded 6 months before or after entering on to the register	6	40	80
New	CLINICAL	**EPILEPSY 9:** The percentage of women under the age of 55 years who are taking antiepileptic drugs who have a record of information and counselling about contraception, conception and pregnancy in the preceding 15 months	3	40	90
New	CLINICAL	**LD 2:** Percentage of patients on the Learning Disability register with DownâĂŹs Syndrome aged 18 years and over who have a record of blood TSH in the preceding 15 months (excluding those who are on the thyroid disease register)	3	40	70

Continued on next page

Table C3: Indicators with changes from 2010 to 2011 (2/5)

Change	**Indicator Domain**	**Description**	**Max. Avail. Points**	**LL**	**UL**
Replacement I	CLINICAL	**CVD 1:** OLD: In those patients with a new diagnosis of hypertension (excluding those with pre-existing CHD, diabetes, stroke and/or TIA) recorded between the preceding 1 April to 31 March: the percentage of patients who have had a face to face cardiovascular risk assessment at the outset of diagnosis using an agreed risk assessment treatment tool NEW:In those patients with a new diagnosis of hypertension (excluding those with pre-existing CHD, diabetes, stroke and/or TIA) recorded between the preceding 1 April to 31 March: the percentage of patients aged 30 to 74 years who have had a face to face cardiovascular risk assessment at the outset of diagnosis (within 3 months of the initial diagnosis) using an agreed risk assessment tool	8	40	70
Replacement I	CLINICAL	**CHD 2 → CHD 13:** OLD: The percentage of patients with newly diagnosed angina (diagnosed after 1 April 2003) who are referred for exercise testing and/or specialist assessment NEW:For patients with newly diagnosed angina (diagnosed after 1 April 2011), the percentage who are referred for specialist assessment	7	40	90
Replacement I	CLINICAL	**MH 4 → MH 17:** OLD: The percentage of patients on lithium therapy with a record of serum creatinine and TSH in the previous 15 months NEW:The percentage of patients on lithium therapy with a record of serum creatinine and TSH in the preceding 9 months	1	40	90
Replacement I	CLINICAL	**MH 5 → MH 18:** OLD: The percentage of patients on lithium therapy with a record of lithium levels in a therapeutic range within the previous 6 months NEW:The percentage of patients on lithium therapy with a record of lithium levels in the therapeutic range within the preceding 4 months	2	40	90
Replacement II	CLINICAL	**DEP 2 → DEP 4:** OLD: In those patients with a new diagnosis of depression, recorded between the preceeding 1 April and 31st March, the percentage of patients who have had an assessment of severity at the outset of treatment using an assessment tool validated for use in primary care NEW:In those patients with a new diagnosis of depression, recorded between the preceding 1 April to 31 March, the percentage of patients who have had an assessment of severity at the time of diagnosis using an assessment tool validated for use in primary care	25 → 17	40	90
Replacement II	CLINICAL	**DEP 3 → DEP 5:** OLD: In those patients with a new diagnosis of depression and assessment of severity recorded between the preceding 1 April to 31 March, the percentage of patients who have had a further assessment of severity 5-12 weeks (inclusive) after the initial recording of the assessment of severity. Both assessments should be completed using an assessment tool validated for use in primary care NEW:In those patients with a new diagnosis of depression and assessment of severity recorded between the preceding 1 April to 31 March, the percentage of patients who have had a further assessment of severity 4-12 weeks (inclusive) after the initial recording of the assessment of severity. Both assessments should be completed using an assessment tool validated for use in primary care	20 → 8	40	90 → 80

Continued on next page

Table C3: Indicators with changes from 2010 to 2011 (3/5)

Change	Indicator Domain	Description	Max. Avail. Points	LL	UL
Replacement III	CLINICAL	**CHD 11 → CHD 14:** OLD: The percentage of patients with a history of myocardial infarction (diagnosed after 1 April 2003) who are currently treated with an ACE inhibitor or angiotensin II antagonist NEW:The percentage of patients with a history of myocardial infarction (from 1 April 2011) currently treated with an ACE inhibitor (or ARB if ACE intolerant), aspirin or an alternative anti-platelet therapy, beta blocker and statin (unless a contraindication or side effects are recorded)	7 → 10	40	80
Replacement III	CLINICAL	**DM 9 → DM 29:** OLD: The percentage of patients with diabetes with a record of the presence or absence of peripheral pulses in the previous 15 months NEW:The percentage of patients with diabetes with a record of a foot examination and risk classification: 1) low risk (normal sensation, palpable pulses), 2) increased risk (neuropathy or absent pulses), 3) high risk (neuropathy or absent pulses plus deformity or skin changes or previous ulcer) or 4) ulcerated foot within the preceding 15 months	3 → 4	40	90
Replacement III	CLINICAL	**DM 12 (withdrawn):** The percentage of patients with diabetes in whom the last blood pressure is 145/85 or less	18	40	60
Replacement III	CLINICAL	**DM 12 → DM 30:** The percentage of patients with diabetes in whom the last blood pressure is 150/90 or less in the preceding 15 months	8	40	71
Replacement III	CLINICAL	**DM 12 → DM 31:** The percentage of patients with diabetes in whom the last blood pressure is 140/80 or less in the preceding 15 months	10	40	60
Replacement III	CLINICAL	**MH 9 (withdrawn):** The percentage of patients with schizophrenia and bipolar affective disorder and other psychoses with a review recorded in the previous 15 months. In the review there is evidence that the patient has participated in routine health promotion and prevention advice appropriate to their age and health status	23	40	90
Replacement III	CLINICAL	**MH 9 → MH 11:** The percentage of patients with schizophrenia, bipolar affective disorder and other psychoses who have a record of alcohol consumption in the preceding 15 months	4	40	90
Replacement III	CLINICAL	**MH 9 → MH 12:** The percentage of patients with schizophrenia, bipolar affective disorder and other psychoses who have a record of BMI in the preceding 15 months	4	40	90
Replacement III	CLINICAL	**MH 9 → MH 13:** The percentage of patients with schizophrenia, bipolar affective disorder and other psychoses who have a record of blood pressure in the preceding 15 months	4	40	90
Replacement III	CLINICAL	**MH 9 → MH 14:** The percentage of patients aged 40 years and over with schizophrenia, bipolar affective disorder and other psychoses who have a record of total cholesterol:hdl ratio in the preceding 15 months	5	40	80
Replacement III	CLINICAL	**MH 9 → MH 15:** The percentage of patients aged 40 years and over with schizophrenia, bipolar affective disorder and other psychoses who have a record of blood glucose level in the preceding 15 months	5	40	80
Replacement III	CLINICAL	**MH 9 → MH 16:** The percentage of patients (aged from 25 to 64 in England and Northern Ireland, from 20 to 60 in Scotland and from 20 to 64 in Wales) with schizophrenia, bipolar affective disorder and other psychoses whose notes record that a cervical screening test has been performed in the preceding 5 years	5	40	80

Continued on next page

Table C3: Indicators with changes from 2010 to 2011 (4/5)

Change	**Indicator Domain**	**Description**	**Max. Avail. Points**	**LL**	**UL**
Replacement IV	CLINICAL	**DM 23 → DM 26:** OLD: The percentage of patients with diabetes in whom the last HbA1C is 7 or less (or equivalent test/reference range depending on local laboratory) in the previous 15 months NEW:The percentage of patients with diabetes in whom the last IFCC-HbA1c is 59 mmol/mol (equivalent to HbA1c of 7.5% in DCCT values) or less (or equivalent test/reference range depending on local laboratory) in the preceding 15 months	17	40	50
"Replacement V	CLINICAL	**COPD 1 → COPD 14:** OLD: The PHC can produce a register of patients with COPD NEW:The PHC can produce a register of patients with COPD. Note: OPD1 was been renumbered to COPD14 following a change in the diagnostic threshold as per the updated NICE guideline.	3	-	-
" "Replacement V	CLINICAL	**COPD 12 → COPD 15:** OLD: The percentage of all patients with COPD diagnosed after 1st April 2008 in whom the diagnosis has been confirmed by post bronchodilator spirometry NEW:The percentage of all patients with COPD diagnosed after 1st April 2011 in whom the diagnosis has been confirmed by post bronchodilator spirometry Note: COPD12 was been renumbered to COPD15 in recognition of a coding change to include new codes for post bronchodilator spirometry and the removal of the reversibility testing codes.	5	40	80
" "Replacement V	CLINICAL	**MH 6 → MH 10:** OLD: The percentage of patients on the register who have a comprehensive care plan documented in the records agreed between individuals, their family and/or carers as appropriate NEW:The percentage of patients on the register who have a comprehensive care plan documented in the records agreed between individuals, their family and/or carers as appropriate Note: The MH6 business rules logic was updated in recognition of the 'unbundled' care review indicator (previously MH9) and the inclusion of the 'remission exclusion' codes, to ensure that the care plan is still reviewed annually and updated following a patient's relapse from remission.	6	25	50
" Replacement V	CLINICAL	**DM 24 → DM 27:** OLD: The percentage of patients with diabetes in whom the last HbA1C is 8 or less (or equivalent test/reference range depending on local laboratory) in the previous 15 months NEW:The percentage of patients with diabetes in whom the last IFCC-HbA1c is 64 mmol/mol (equivalent to HbA1c of 8% in DCCT values) or less (or equivalent test/reference range depending on local laboratory) in the preceding 15 months	8	40	70
Replacement V	CLINICAL	**DM 25 → DM 28:** OLD: The percentage of patients with diabetes in whom the last HbA1C is 9 or less (or equivalent test/reference range depending on local laboratory) in the previous 15 months NEW:The percentage of patients with diabetes in whom the last IFCC-HbA1c is 75 mmol/mol (equivalent to HbA1c of 9% in DCCT values) or less (or equivalent test/reference range depending on local laboratory) in the preceding 15 months	10	40	90

Continued on next page

Table C3: Indicators with changes from 2010 to 2011 (5/5)

Change	Indicator Domain	Description	Max. Avail. Points	LL	UL
Withdrawn	ORGANISATIONAL	**INFORMATION 4:** If a patient is removed from a PHC's list, the PHC provides an explanation of the reasons in writing to the patient and information on how to find a new PHC, unless it is perceived such an action would result in a violent response by the patient	1	-	-
Withdrawn	ORGANISATIONAL	**RECORDS 21:** Ethnic origin is recorded for 100% of new registrations from 1st April 2006	1	-	-
Withdrawn	PATIENT EXPERIENCE	**PE 7:** The percentage of patients who, using an approved survey, indicate that they were able to obtain a consultation with an appropriate health care professional within 2 working days	23.5	70	90
Withdrawn	PATIENT EXPERIENCE	**PE 8:** The percentage of patients who, using an approved survey, indicate that they were able to book an appointment with a GP more than 2 days ahead	35	60	90
New	ORGANISATIONAL	**QP 1:** The PHC conducts an internal review of their prescribing to assess whether it is clinically appropriate and cost effective, agrees with the PCO three areas for improvement and produces a draft plan for each area no later than 30 June 2011	6	-	-
New	ORGANISATIONAL	**QP 2:** The PHC participates in an external peer review of prescribing with a group of PHCs and agrees plans for three prescribing areas for improvement firstly with the group and then with the PCO no later than 30 September 2011	7	-	-
New	ORGANISATIONAL	**QP 3:** The percentage of prescriptions complying with the agreed plan for the first improvement area as a percentage of all prescriptions in that improvement area during the period 1 January 2012 to 31 March 2012 - (Payment stages to be determined locally according to the method set out in the indicator guidance with 20 percentage points between upper and lower thresholds)	5	-	-
New	ORGANISATIONAL	**QP 4:** The percentage of prescriptions complying with the agreed plan for the second improvement area as a percentage of all prescriptions in that improvement area during the period 1 January 2012 to 31 March 2012.- (Payment stages to be determined locally according to the method set out in the indicator guidance with 20 percentage points between upper and lower thresholds)	5	-	-
New	ORGANISATIONAL	**QP 5:** The percentage of prescriptions complying with the agreed plan for the third improvement area as a percentage of all prescriptions in that improvement area during the period 1 January 2012 to 31 March 2012 - (Payment stages to be determined locally according to the method set out in the indicator guidance with 20 percentage points between upper and lower thresholds)	5	-	-
New	ORGANISATIONAL	**QP 6:** The PHC meets internally to review the data on secondary care outpatient referrals provided by the PCO	5	-	-
New	ORGANISATIONAL	**QP 7:** The PHC participates in an external peer review with a group of PHCs to compare its secondary care outpatient referral data either with PHCs in the group of PHCs or with PHCs in the PCO area and proposes areas for commissioning or service design improvements to the PCO	5	-	-
New	ORGANISATIONAL	**QP 8:** The PHC engages with the development of and follows three agreed care pathways for improving the management of patients in the primary care setting (unless in individual cases they justify clinical reasons for not doing this) to avoid inappropriate outpatient referrals and produces a report of the action taken to the PCO no later than 31 March 2012	11	-	-
New	ORGANISATIONAL	**QP 9:** The PHC meets internally to review the data on emergency admissions provided by the PCO	5	-	-
New	ORGANISATIONAL	**QP 10:** The PHC participates in an external peer review with a group of PHCs to compare its data on emergency admissions either with PHCs in the group of PHCs or PHCs in the PCO area and proposes areas for commissioning or service design improvements to the PCO	15	-	-
New	ORGANISATIONAL	**QP 11:** The PHC engages with the development of and follows three agreed care pathways (unless in individual cases they justify clinical reasons for not doing this) in the management and treatment of patients in aiming to avoid emergency admissions and produces a report of the action taken to the PCO no later than 31 March 2012	27.5	-	-

D Autocorrelation of QOF Achievement Measures

As we explained in subsection 3.2, our empirical strategy relies on a high degree of autocorrelation on the performance measures (QOF indicators) in the absence of changes to the reward schemes. Table D1 below provides descriptive statistics on the QOF achievement indicators, with emphasis on showing the high degree of autocorrelation, thus complementing Figure 9. As in Figure 9, we use the QOF data of 2009 and 2010 for which there were no changes neither in the definition of the indicators nor in the reward schemes.

Column 4 of Table D1 of this Appendix shows that, for most indicators, the autocorrelation coefficients is above 0.4. Moreover, we can observe high persistence along the 45 degrees diagonal (column 5): for most indicators the percentage of PHCs whose achievement was within 5 points of the previous year is above 70%. In average, PHCs which are below the upper limit, UL, increase their performance in the following year (column 6), while those which are above this threshold tend to hardly change their performance (column 7).

Table D1: QOF indicators descriptives for t =2010/11 financial year ($t = 2$)

		(1)	(2)	(3)	(4)	(5)	(6)	(7)
Indicator	UL	$E[x_t]$	$SD[x_t]$	$\text{P}[x_t < UL]$	$\rho(x_t)$	$E[\mathbb{1}\{\lvert x_t - x_{t-1}\rvert$	$E[x_t - x_{t-1}$	$E[x_t - x_{t-1}$
						$\leq 5\} * 100]$	$\lvert x_{t-1} < UL]$	$\lvert x_{t-1} > UL]$
AF03	90%	93.82	7.92	7.14	0.50	83.10	9.88	-0.55
AF04	90%	95.28	12.11	6.43	0.51	74.64	26.20	-1.39
ASTHMA03	80%	90.00	10.77	4.69	0.41	52.00	18.64	-0.76
ASTHMA06	70%	79.58	8.31	5.29	0.54	62.43	11.03	-0.19
ASTHMA08	80%	87.89	9.06	6.37	0.46	53.61	15.63	-0.98
BP5	70%	79.68	6.57	5.17	0.64	78.87	5.87	-0.09
CANCER03	90%	92.75	15.77	17.84	0.34	61.84	18.18	-2.69
CHD08	70%	81.90	7.64	3.51	0.56	74.25	11.30	-0.41
CHD09	90%	93.58	5.67	7.56	0.46	92.31	4.30	-0.66
CHD10	60%	74.91	10.95	2.60	0.67	65.06	13.30	-0.70
CHD12	90%	92.73	6.78	16.53	0.48	78.59	5.17	-0.31
COPD08	85%	93.52	7.38	4.12	0.47	73.14	13.30	-0.10
COPD10	70%	88.48	10.99	3.60	0.51	57.78	25.17	-0.75
COPD13	90%	91.17	10.77	17.95	0.47	69.90	13.54	-1.31

Continued on next page

Table D1: (Continued)

		(1)	(2)	(3)	(4)	(5)	(6)	(7)
Indicator	UL	$E[x_t]$	$SD[x_t]$	P$[x_t < UL]$	$\rho(x_t)$	$E[\mathbb{1}\{(x_t - x_{t-1})$	$E[x_t - x_{t-1}$	$E[x_t - x_{t-1}$
						$\leq 5\}]$	$\vert x_{t-1} < UL]$	$\vert x_{t-1} > UL]$
CKD02	90%	97.26	6.97	1.29	0.41	95.33	26.83	-0.37
CKD03	70%	74.86	10.41	21.73	0.52	60.55	5.58	-1.72
CKD05	80%	90.78	17.96	6.03	0.46	61.61	40.20	-2.70
CKD06	80%	82.35	13.05	24.29	0.53	48.00	14.80	-1.33
CVD02	70%	82.61	14.05	7.94	0.37	32.02	34.13	-5.68
DEM02	60%	80.54	16.22	3.04	0.42	36.60	36.15	-0.92
DM2	90%	94.87	4.49	7.00	0.54	90.67	4.47	-0.36
DM10	90%	91.39	7.57	22.84	0.58	77.87	4.93	-0.69
DM13	90%	88.80	9.05	37.48	0.65	74.51	3.38	-1.38
DM15	80%	89.28	13.53	8.07	0.53	64.75	20.31	-1.66
DM17	70%	82.73	6.52	2.43	0.60	78.13	8.70	-0.55
DM18	85%	91.19	6.30	9.76	0.47	75.85	5.99	-0.17
DM21	90%	91.08	7.54	24.33	0.52	74.24	5.46	-0.97
DM22	90%	96.95	3.75	2.44	0.44	94.54	7.78	-0.02
EPILEP06	90%	95.62	7.40	6.95	0.27	73.54	13.07	-0.64
EPILEP08	70%	73.96	15.26	26.14	0.56	43.38	7.72	-3.09
HF02	90%	95.46	11.25	8.02	0.51	74.46	17.25	-1.03
HF03	80%	90.26	12.20	4.24	0.46	61.26	27.42	-1.10
HF04	60%	83.15	17.97	3.26	0.47	44.45	41.65	-1.39
SMOKE03	90%	95.61	2.85	2.66	0.52	95.25	5.69	0.00
SMOKE04	90%	93.07	5.11	12.48	0.44	79.02	4.98	-0.72
STROKE07	90%	91.49	7.81	23.91	0.43	73.96	5.01	-1.08
STROKE08	60%	77.18	10.02	3.07	0.50	58.70	20.72	-0.57
STROKE10	85%	90.09	8.63	13.45	0.40	64.69	8.97	-0.49
STROKE12	90%	93.79	8.41	8.98	0.45	82.43	9.97	-0.93
STROKE13	80%	88.90	15.87	7.51	0.58	55.50	25.92	-1.87
THYROI02	90%	95.81	4.33	3.24	0.41	91.02	10.46	-0.11

Notes: Own calculations based on QOF data of 2009 and 2010, when there was no change in neither the definition of the indicators nor the reward schemes. Number of PHCS is 8245. $E[x_t]$ **:** Average achievement per indicator. $\mathbf{P}[x_t < UL]$ **:** Proportion of PHCs with an achievement below the upper limit of the non-linear reward scheme, UL. $\rho(x_t)$ **:** Correlation between 2010 and 2009 achievement. $E[\mathbb{1}\{(x_t - x_{t-1}) \leq 5\}]x100$ **:** Percentage of PHCs with achievement difference between 2009 and 2010 of 5 percentage points. $E[x_t - x_{t-1}|x_{t-1} < UL]$ **:** Difference on achievement between 2010 and 2009, conditional on achievement below UL in 2009. $E[x_t - x_{t-1}|x_{t-1} > UL]$ **:** Difference on achievement between 2010 and 2009, conditional on achievement above UL in 2009.

E Bunching Results

Table E1 of this Appendix reports the estimates of the excess bunching at UL for the indicators that did not change between 2009-2010 and 2011 (we refer to these indicators as x_1 in the main text), assuming a bunching region of 2 (second column) or 3 (third column). We keep constant the estimation window w and the bin-size $b = 1$. The coefficients in the second and third columns corresponds to the estimate of excess bunching (B) at [UL,UL+h] relative to the total number of PHCs included in [UL-w,UL+w]. Column 4 of the Table classifies indicators as bunched only if the estimate of the bunching is statistically significant at the 95% level under both $h = 2$ and $h = 3$. In total, 25 out of 41 indicators present evidence of bunching. Below, Table E2 reports similar results under different values of the estimation window and bunching region. Table E3, also below, reports the bunching graphs, equivalent to those of Figure 10 of the main text, but for all the QOF clinical indicators whose reward scheme remained unchanged between 2009 and 2011.

Table E1: Results of QOF indicators bunching test

	(1)	(2)	(3)	(4)		(5)	(6)	(7)	(8)
		w=10,	w=10,				w=10,	w=10,	
Indicator	UL	h=2,	h=3,	Bunched	Indicator	UL	h=2,	h=3,	Bunched
		b=1	b=1				b=1	b=1	
AF03	90	13.4 **	23.5 ***	Yes	DM13	90	8.7 **	17.2 ***	Yes
		[2.93]	[9.40]				[2.26]	[6.40]	
AF04	90	11.1 ***	14.8 ***	Yes	DM15	80	13.0 ***	9.8 ***	Yes
		[4.35]	[5.01]				[7.37]	[3.17]	
ASTHMA03	80	9.4 ***	7.3 ***	Yes	DM17	70	0.3	2.1 ***	
		[7.28]	[3.13]				[0.31]	[3.57]	
ASTHMA06	70	3.6 **	6.9 ***	Yes	DM18	85	1.7 **	1.0	
		[2.20]	[7.57]				[2.93]	[0.91]	
ASTHMA08	80	12.1 ***	15.4 ***	Yes	DM21	90	12.3 ***	20.7 ***	Yes
		[6.51]	[9.86]				[3.44]	[11.49]	
BP5	70	0.8	1.9 ***		DM22	90	0.9	2.6 **	
		[1.26]	[5.10]				[0.97]	[2.39]	
CANCER03	90	15.3 **	27.0 ***	Yes	EPILEP06	90	3.9 **	4.2	
		[2.30]	[4.55]				[2.79]	[1.39]	
CHD08	70	1.1 **	2.0		EPILEP08	70	10.5 ***	15.6 ***	Yes
		[2.93]	[1.68]				[4.52]	[3.90]	

Continued on next page

Table E1: (Continued)

	(1)	(2)	(3)	(4)		(5)	(6)	(7)	(8)
		w=10,	w=10,				w=10,	w=10,	
Indicator	UL	h=2,	h=3,	Bunched	Indicator	UL	h=2,	h=3,	Bunched
		b=1	b=1				b=1	b=1	
CHD09	90	7.6 ***	11.9 ***	Yes	HF02	90	17.0 ***	24.5 ***	Yes
		[3.23]	[6.97]				[4.35]	[7.56]	
CHD10	60	4.7 ***	6.2 ***	Yes	HF03	80	14.9 ***	10.9 ***	Yes
		[4.39]	[5.93]				[7.61]	[3.86]	
CHD12	90	13.0 ***	21.2 ***	Yes	HF04	60	15.4 ***	18.0 *	
		[3.56]	[10.56]				[5.97]	[1.99]	
COPD08	85	1.5 **	1.5		SMOKE03	90	0.5	5.4	
		[2.91]	[1.14]				[-0.21]	[-1.41]	
COPD10	70	3.2 ***	3.2		SMOKE04	90	15.2 **	28.5 ***	Yes
		[5.94]	[1.19]				[2.79]	[6.94]	
COPD13	90	16.9 ***	28.4 ***	Yes	STROKE07	90	8.4 **	15.9 ***	Yes
		[3.41]	[8.70]				[2.44]	[7.95]	
CKD02	90	1.3	3.7 **		STROKE08	60	3.5 ***	5.3 ***	Yes
		[1.05]	[2.46]				[3.28]	[4.83]	
CKD03	70	7.4 ***	10.8 ***	Yes	STROKE10	85	4.7 ***	6.6 ***	Yes
		[5.24]	[9.84]				[5.26]	[5.74]	
CKD05	80	21.0 ***	11.1		STROKE12	90	8.0 **	13.5 ***	Yes
		[4.53]	[1.67]				[2.68]	[5.54]	
CKD06	80	4.8 ***	6.1 ***	Yes	STROKE13	80	18.0 ***	14.2 ***	Yes
		[4.16]	[4.57]				[8.45]	[3.52]	
CVD02	70	4.6 ***	5.3		THYROI02	90	1.1	2.7	
		[5.23]	[1.74]				[-1.07]	[-1.54]	
DEM02	60	12.7 ***	12.7 **	Yes					
		[9.36]	[2.74]						
DM2	90	0.2	2.6						
		[-0.11]	[-1.04]						
DM10	90	4.3 *	8.5 ***						
		[1.95]	[4.95]						

Notes: Own calculations based on QOF data. In each column, a spline with 5 knots is estimated over the histogram of the indicator in the interval [UL-w,UL+w]. The regressor includes dummies γ_k for each bin of size b located in [UL,UL+h]. Each coefficient in the table corresponds to the estimate of excess density (B) at [UL,UL+h] relative to the total number of PHCs included in [UL-w,UL+w]. The z-stats in brackets corresponds to H0: $\sum_k \gamma_k = 0$. Significance: ** 5%, *** 1%.

Table E2: Results of QOF indicators bunching test: additional specifications

	(1)	(2)	(3)	(4)	(5)	(6)
		w=10,	w=10,	w=12,	w=12,	w=10,
Indicator	UL	h=2,	h=3,	h=2,	h=3,	h=4,
		b=1	b=1	b=1	b=1	b=1
AF03	90	13.4 **	23.5 ***	19.4 *	41.5 ***	27.2 ***
		[2.93]	[9.40]	[1.96]	[5.10]	[10.13]
AF04	90	11.1 ***	14.8 ***	15.4 ***	25.3 ***	11.4 *
		[4.35]	[5.01]	[3.21]	[7.41]	[2.01]
ASTHMA03	80	9.4 ***	7.3 ***	7.1 ***	6.8 ***	8.8
		[7.28]	[3.13]	[5.55]	[3.18]	[1.74]
ASTHMA06	70	3.6 **	6.9 ***	3.0	7.1 ***	9.1 ***
		[2.20]	[7.57]	[1.35]	[3.76]	[30.53]
ASTHMA08	80	12.1 ***	15.4 ***	9.8 ***	13.8 ***	15.6 ***
		[6.51]	[9.86]	[4.59]	[7.22]	[6.66]
BP5	70	0.8	1.9 ***	0.8 *	1.4 ***	1.8 *
		[1.26]	[5.10]	[1.96]	[3.01]	[1.85]
CANCER03	90	15.3 **	27.0 ***	21.6 **	37.6 ***	28.0 ***
		[2.30]	[4.55]	[2.66]	[5.07]	[3.64]
CHD08	70	1.1 **	2.0	0.8 **	1.2 *	1.7
		[2.93]	[1.68]	[2.66]	[1.79]	[0.45]
CHD09	90	7.6 ***	11.9 ***	9.5	20.5 ***	14.6 ***
		[3.23]	[6.97]	[1.56]	[3.23]	[10.05]
CHD10	60	4.7 ***	6.2 ***	3.8 ***	5.5 ***	6.2 ***
		[4.39]	[5.93]	[3.48]	[5.77]	[4.18]
CHD12	90	13.0 ***	21.2 ***	14.5 ***	24.8 ***	25.3 ***
		[3.56]	[10.56]	[3.17]	[7.14]	[16.39]
COPD08	85	1.5 **	1.5	1.2	0.6	1.1
		[2.91]	[1.14]	[1.65]	[0.44]	[0.31]
COPD10	70	3.2 ***	3.2	1.7 ***	2.9 **	1.6
		[5.94]	[1.19]	[3.48]	[2.16]	[0.20]
COPD13	90	16.9 ***	28.4 ***	21.0 ***	36.8 ***	36.3 ***
		[3.41]	[8.70]	[3.04]	[6.35]	[18.56]
CKD02	90	1.3	3.7 **	0.5 ***	0.5 **	7.2 ***
		[1.05]	[2.46]	[3.95]	[2.78]	[3.30]
CKD03	70	7.4 ***	10.8 ***	6.6 ***	10.5 ***	12.2 ***
		[5.24]	[9.84]	[3.90]	[7.99]	[6.08]
CKD05	80	21.0 ***	11.1	14.6 ***	9.6	23.8 *
		[4.53]	[1.67]	[4.14]	[1.67]	[1.93]

Continued on next page

Table E2: (Continued)

	(1)	(2)	(3)	(4)	(5)	(6)
		w=10,	w=10,	w=12,	w=12,	w=10,
Indicator	UL	h=2,	h=3,	h=2,	h=3,	h=4,
		b=1	b=1	b=1	b=1	b=1
CKD06	80	4.8 ***	6.1 ***	4.0 ***	6.0 ***	6.8 **
		[4.16]	[4.57]	[3.03]	[4.39]	[2.89]
CVD02	70	4.6 ***	5.3	3.6 ***	5.2 ***	3.2
		[5.23]	[1.74]	[3.73]	[3.13]	[0.36]
DEM02	60	12.7 ***	12.7 **	10.5 ***	12.5 ***	9.1
		[9.36]	[2.74]	[6.86]	[6.28]	[0.60]
DM2	90	0.2	2.6	0.4	0.2	7.7 *
		[-0.11]	[-1.04]	[0.60]	[-0.23]	[-1.90]
DM10	90	4.3 *	8.5 ***	4.9	13.9 **	11.8 ***
		[1.95]	[4.95]	[0.95]	[2.61]	[11.23]
DM13	90	8.7 **	17.2 ***	10.5 *	21.6 ***	23.9 ***
		[2.26]	[6.40]	[2.08]	[5.01]	[22.44]
DM15	80	13.0 ***	9.8 ***	9.5 ***	8.4 ***	12.4 *
		[7.37]	[3.17]	[6.73]	[3.31]	[1.81]
DM17	70	0.3	2.1 ***	0.2	1.0 *	1.8
		[0.31]	[3.57]	[0.44]	[1.99]	[1.41]
DM18	85	1.7 **	1.0	1.5 ***	1.2	1.0
		[2.93]	[0.91]	[3.03]	[1.45]	[-0.58]
DM21	90	12.3 ***	20.7 ***	14.5 **	27.4 ***	24.7 ***
		[3.44]	[11.49]	[2.38]	[5.80]	[14.95]
DM22	90	0.9	2.6 **	0.4 **	0.5 *	5.0 **
		[0.97]	[2.39]	[2.40]	[1.98]	[2.88]
EPILEP06	90	3.9 **	4.2	5.2 **	8.3 ***	0.9
		[2.79]	[1.39]	[2.84]	[4.57]	[-0.15]
EPILEP08	70	10.5 ***	15.6 ***	9.2 ***	15.0 ***	18.2
		[4.52]	[3.90]	[3.39]	[4.63]	[1.62]
HF02	90	17.0 ***	24.5 ***	33.3 ***	58.9 ***	22.3 ***
		[4.35]	[7.56]	[3.11]	[7.53]	[3.81]
HF03	80	14.9 ***	10.9 ***	10.7 ***	9.1 ***	12.2 *
		[7.61]	[3.86]	[7.66]	[3.89]	[1.80]
HF04	60	15.4 ***	18.0 *	13.9 ***	18.2 ***	7.7
		[5.97]	[1.99]	[4.41]	[4.45]	[0.28]
SMOKE03	90	0.5	5.4	0.9	2.0	12.4
		[-0.21]	[-1.41]	[0.40]	[-0.73]	[-1.66]
SMOKE04	90	15.2 **	28.5 ***	16.1 *	33.9 ***	40.5 ***

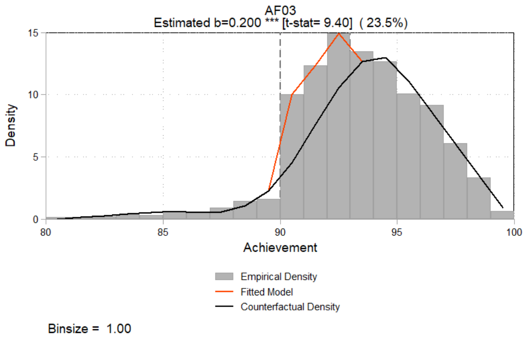
Continued on next page

Table E2: (Continued)

	(1)	(2)	(3)	(4)	(5)	(6)
		w=10,	w=10,	w=12,	w=12,	w=10,
Indicator	UL	h=2,	h=3,	h=2,	h=3,	h=4,
		b=1	b=1	b=1	b=1	b=1
		[2.79]	[6.94]	[1.94]	[4.41]	[15.62]
STROKE07	90	8.4 **	15.9 ***	10.9	25.4 ***	20.3 ***
		[2.44]	[7.95]	[1.60]	[4.38]	[18.12]
STROKE08	60	3.5 ***	5.3 ***	3.0 ***	3.9 ***	5.3 **
		[3.28]	[4.83]	[4.30]	[6.41]	[2.32]
STROKE10	85	4.7 ***	6.6 ***	4.0 ***	6.7 ***	7.1 **
		[5.26]	[5.74]	[2.94]	[5.43]	[2.24]
STROKE12	90	8.0 **	13.5 ***	12.2	26.6 ***	14.8 ***
		[2.68]	[5.54]	[1.66]	[4.03]	[3.94]
STROKE13	80	18.0 ***	14.2 ***	13.1 ***	11.9 ***	20.2 **
		[8.45]	[3.52]	[6.74]	[3.37]	[2.48]
THYROI02	90	1.1	2.7	0.4	0.8	5.9
		[-1.07]	[-1.54]	[-0.91]	[-0.99]	[-1.80]

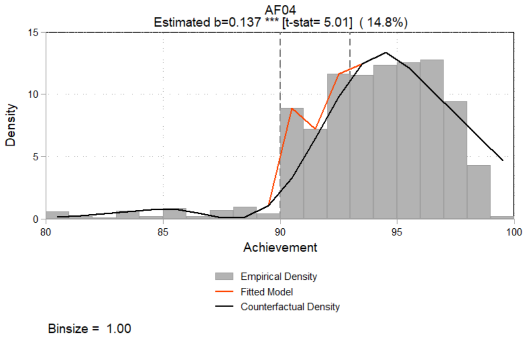
Notes: Own calculations based on QOF data. In each column, a spline with 5 knots is estimated over the histogram of the indicator in the interval [UL-w,UL+w]. The regressor includes dummies γ_k for each bin of size b located in [UL,UL+h]. Each coefficient in the table corresponds to the estimate of excess density (B) at [UL,UL+h] relative to the total number of PHCs included in [UL-w,UL+w]. The z-stats in brackets corresponds to H0: $\sum_k \gamma_k = 0$. Significance: ** 5%, *** 1%.

Table E3: Bunching graphs of QOF clinical indicators which did not change between 2009-2010 and 2011



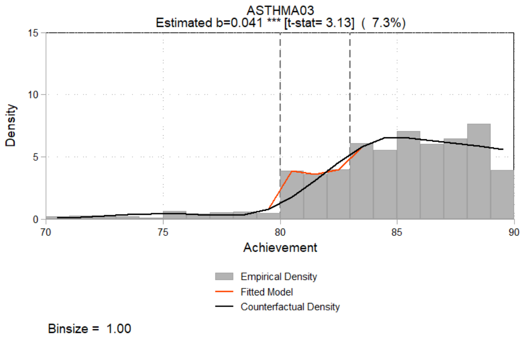
AF03: The percentage of patients with atrial fibrillation who are currently treated with anti-coagulant drug therapy or an anti-platelet drug therapy

12 points. LL=40, UL=90.



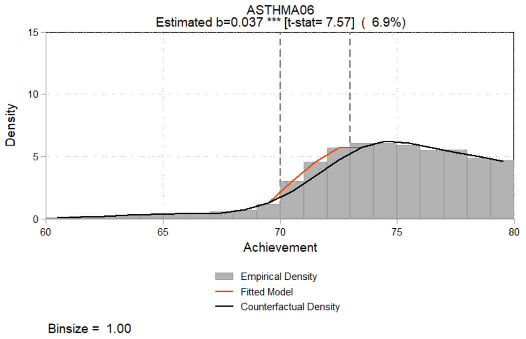
AF04: The percentage of patients with atrial fibrillation diagnosed after 1st April 2008 with ECG or specialist confirmed diagnosis

10 points. LL=40, UL=90.



ASTHMA03: The percentage of patients with asthma between the ages of 14 and 19 in whom there is a record of smoking status in the previous 15 months

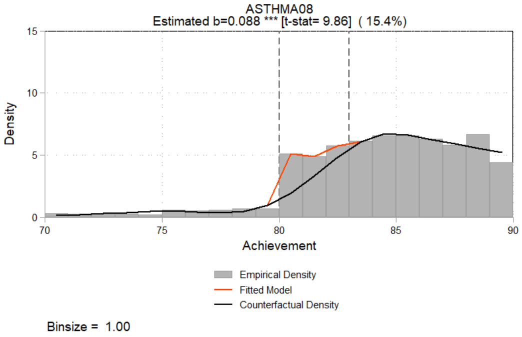
6 points. LL=40 UL=80.



ASTHMA06: The percentage of patients with asthma who have had an asthma review in the previous 15 months

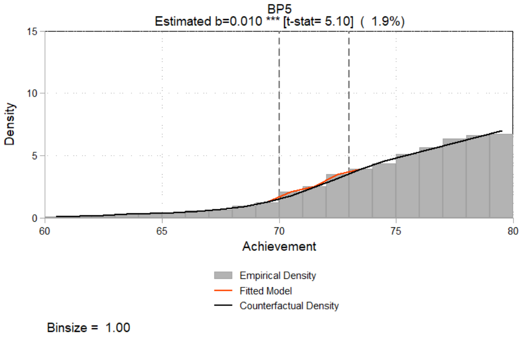
20 points. LL=40 UL=70.

Table E3: Bunching graphs of QOF clinical indicators which did not change between 2009-2010 and 2011 (Continued)



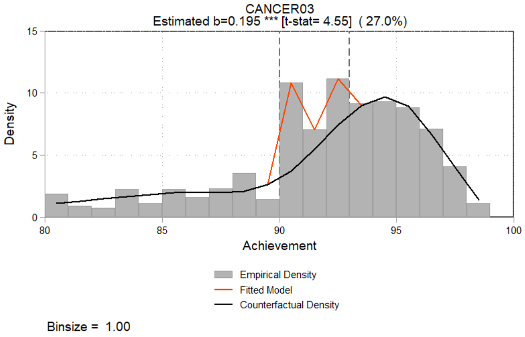
ASTHMA08: The percentage of patients aged eight and over diagnosed as having asthma from 1st April 2007 with measures of variability or reversibility

15 points. LL=40 UL=80.



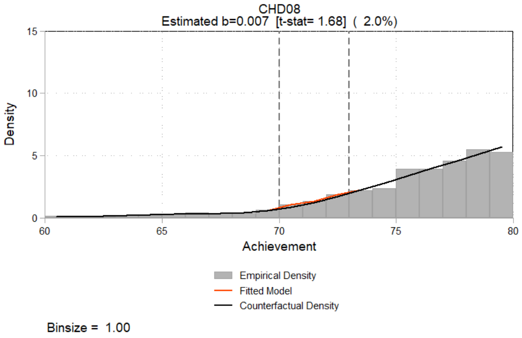
BP05: The percentage of patients with hypertension in whom the last blood pressure (measured in the previous 9 months) is 150/90 or less

57 points. LL=40 UL=70.



CANCER03: The percentage of patients with cancer, diagnosed within the last 18 months, who have a patient review recorded as occurring at 6 months after the practice (PHC) has received confirmation of the diagnosis

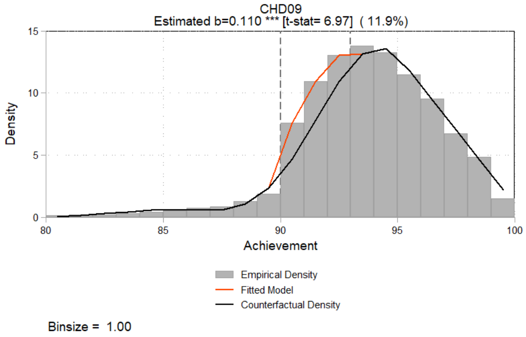
6 points. LL=40 UL=90.



CHD08: The percentage of patients with coronary heart disease whose last measured total cholesterol (measured in the previous 15 months) is 5 mmol/l or less

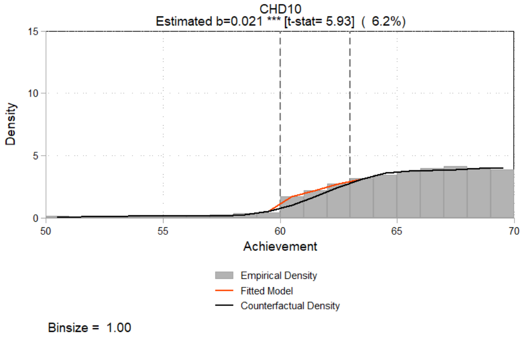
17 points. LL=40 UL=70.

Table E3: Bunching graphs of QOF clinical indicators which did not change between 2009-2010 and 2011 (Continued)



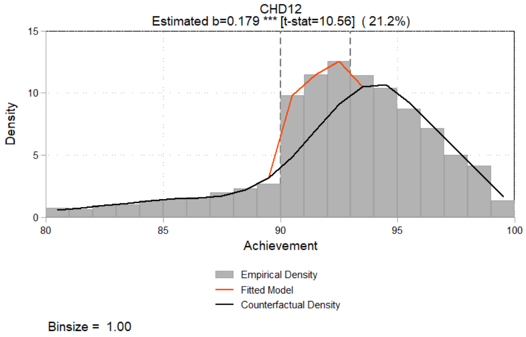
CHD09: The percentage of patients with coronary heart disease with a record in the previous 15 months that aspirin, an alternative anti-platelet therapy, or an anti-coagulant is being taken (unless a contraindication or side-effects are recorded)

7 points. LL=40 UL=90.



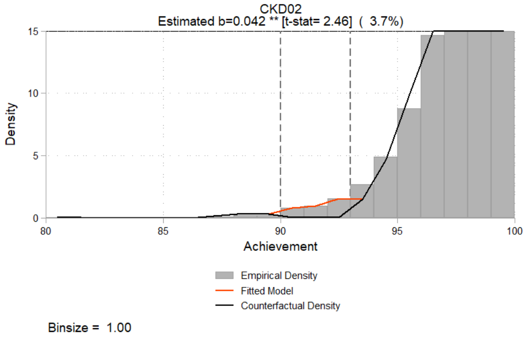
CHD10: The percentage of patients with coronary heart disease who are currently treated with a beta blocker (unless a contraindication or side-effects are recorded)

7 points. LL=40 UL=60.



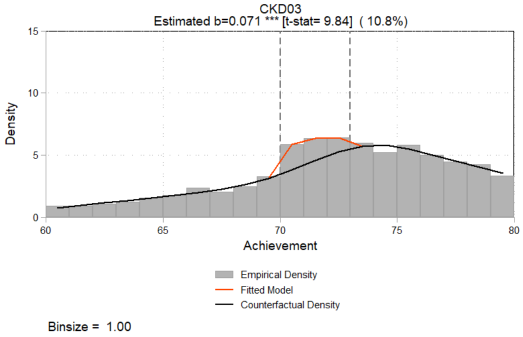
CHD12: The percentage of patients with coronary heart disease who have a record of influenza immunisation in the preceding 1 September to 31 March

7 points. LL=40 UL=90.



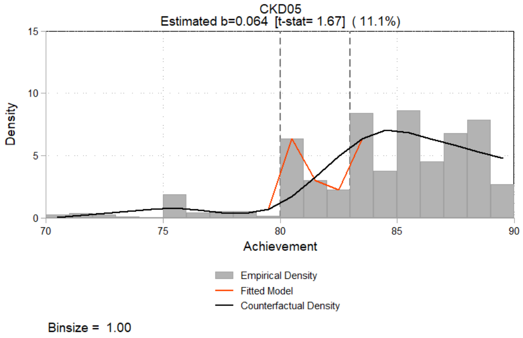
CKD02: The percentage of patients on the CKD register whose notes have a record of blood pressure in the previous 15 months *6 points. LL=40 UL=90.*

Table E3: Bunching graphs of QOF clinical indicators which did not change between 2009-2010 and 2011 (Continued)



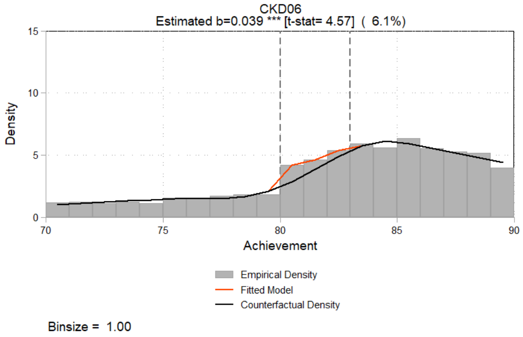
CKD03: The percentage of patients on the CKD register in whom the last blood pressure reading, measured in the previous 15 months, is 140/85 or less

11 points. LL=40 UL=70.



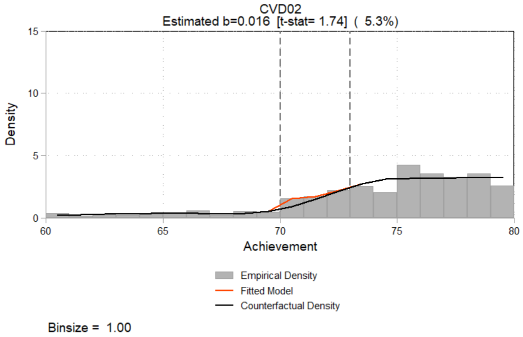
CKD05: The percentage of patients on the CKD register with hypertension and proteinuraa who are treated with an angiotensin converting enzyme inhibitor (ACE-I) or angiotensin receptor blocker (ARB) (unless a contraindication or side effects are recorded)

9 points. LL=40 UL=80.



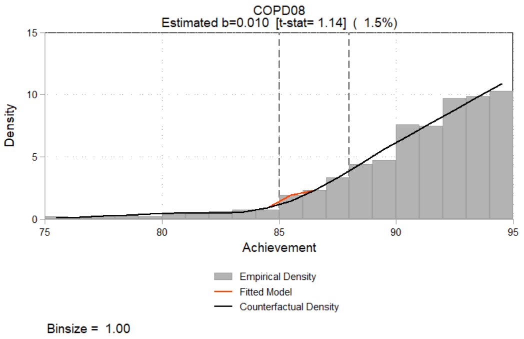
CKD06: The percentage of patients on the CKD register whose notes have a record of an albumin:creatinine ratio (or protein:creatinine ratio) value in the previous 15 months

6 points. LL=40 UL=80.



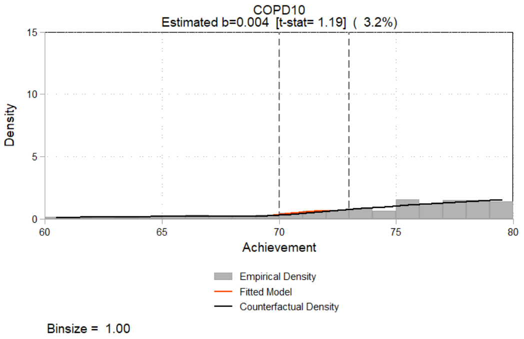
PP02: The percentage of people diagnosed with hypertension diagnosed after 1 April 2009 who are given lifestyle advice in the last 15 months for: increasing physical activity, smoking cessation, safe alcohol consumption and healthy diet *5 points. LL=40 UL=70.*

Table E3: Bunching graphs of QOF clinical indicators which did not change between 2009-2010 and 2011 (Continued)



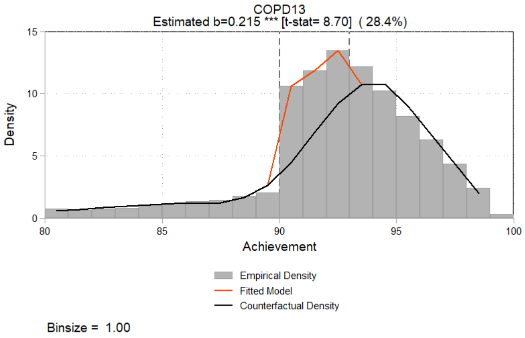
COPD08: The percentage of patients with COPD who have had influenza immunisation in the preceding 1 September to 31 March

6 points. LL=40 UL=85.



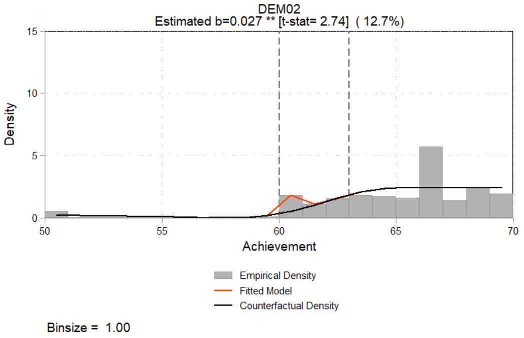
COPD10: The percentage of patients with COPD with a record of FeV1 in the previous 15 months

7 points. LL=40 UL=70.



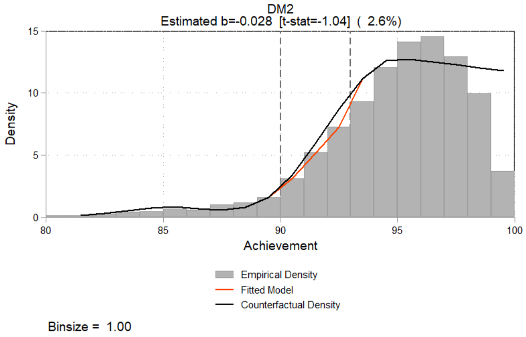
COPD13: The percentage of patients with COPD who have had a review, undertaken by a healthcare professional, including an assessment of breathlessness using the MRC dyspnoea score in the preceding 15 months

9 points. LL=50 UL=90.



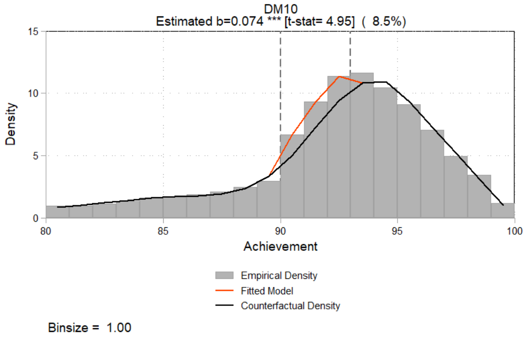
DEM02: The percentage of patients diagnosed with dementia whose care has been reviewed in the previous 15 months *15 points. LL=25 UL=60.*

Table E3: Bunching graphs of QOF clinical indicators which did not change between 2009-2010 and 2011 (Continued)



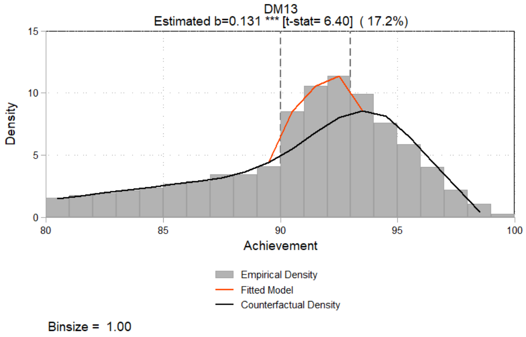
DM02: The percentage of patients with diabetes whose notes record BMI in the previous 15 months

3 points. LL=40 UL=90.



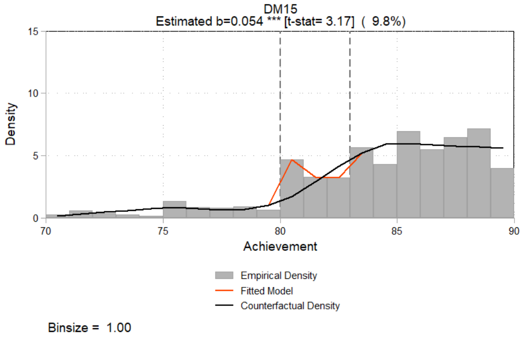
DM10: The percentage of patients with diabetes with a record of neuropathy testing in the previous 15 months

3 points. LL=40 UL=90.



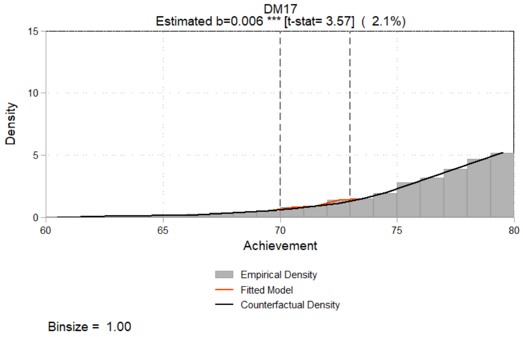
DM13: The percentage of patients with diabetes who have a record of micro-albuminuria testing in the previous 15 months (exception reporting for patients with proteinuria)

3 points. LL=40 UL=90.



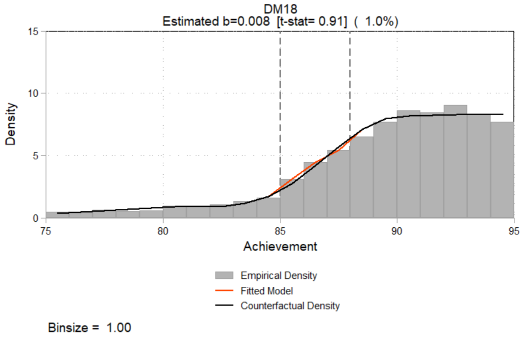
DM15: The percentage of patients with diabetes with proteinuria or micro-albuminuria who are treated with ACE inhibitors (or A2 antagonists) *3 points. LL=40 UL=80.*

Table E3: Bunching graphs of QOF clinical indicators which did not change between 2009-2010 and 2011 (Continued)



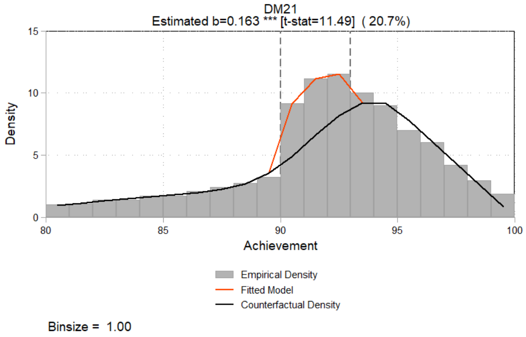
DM17: The percentage of patients with diabetes whose last measured total cholesterol within the previous 15 months is 5mmol/l or less

6 points. LL=40 UL=70.



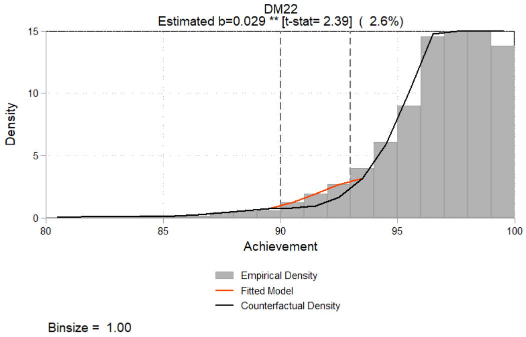
DM18: The percentage of patients with diabetes who have had influenza immunisation in the preceding 1 September to 31 March

3 points. LL=40 UL=85.



DM21: The percentage of patients with diabetes who have a record of retinal screening in the previous 15 months

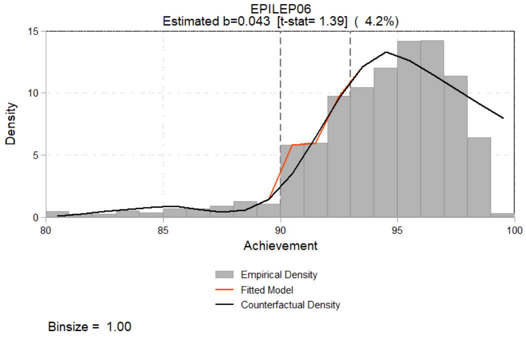
5 points. LL=40 UL=90.



DM22: The percentage of patients with diabetes who have a record of estimated glomerular filtration rate (eGFR) or serum creatinine testing in the previous 15 months

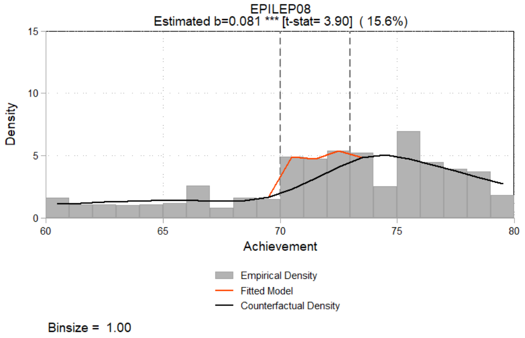
3 points. LL=40 UL=90.

Table E3: Bunching graphs of QOF clinical indicators which did not change between 2009-2010 and 2011 (Continued)



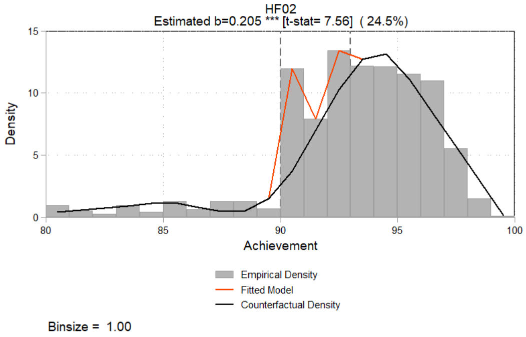
EPILEP06: The percentage of patients aged 18 years and over on drug treatment for epilepsy who have a record of seizure frequency in the previous 15 months

4 points. LL=40 UL=90.



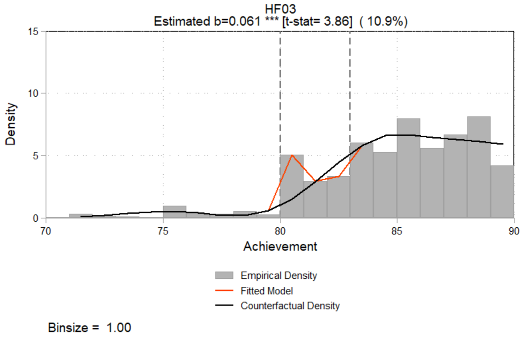
EPILEP08: The percentage of patients aged 18 years and over on drug treatment for epilepsy who have been seizure free for the last 12 months recorded in the previous 15 months

6 points. LL=40 UL=70.



HF02: The percentage of patients with a diagnosis of heart failure (diagnosed after the 1st April 2006) which has been confirmed by an echocardiogram or by specialist assessment

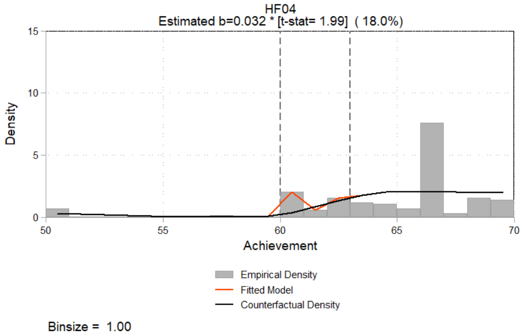
6 points. LL=40 UL=90.



HF03: The percentage of patients with a current diagnosis of heart failure due to LVD who are currently treated with an ACE inhibitor or Angiotensin Receptor Blocker (unless a contraindication or side effects are recorded)

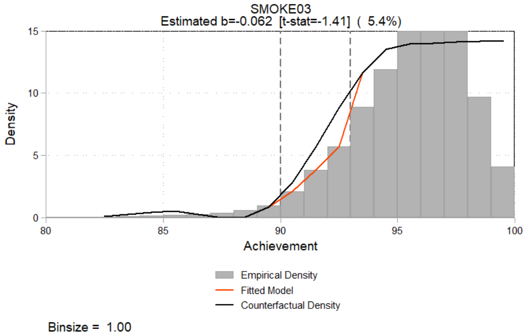
10 points. LL=40 UL=80.

Table E3: Bunching graphs of QOF clinical indicators which did not change between 2009-2010 and 2011 (Continued)



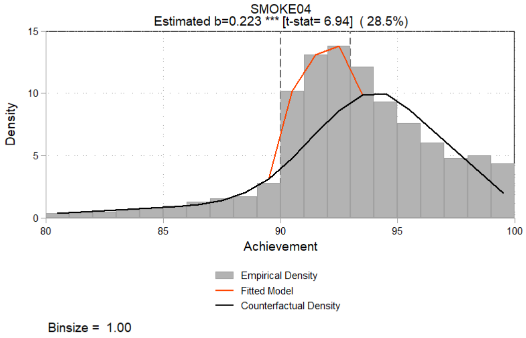
HF04: The percentage of patients with a current diagnosis of heart failure due to LVD who are currently treated with an ACE inhibitor or Angiotensin Receptor Blocker, who are additionally treated with a beta-blocker licensed for heart failure, or recorded as intolerant to or having a contraindication to beta-blockers

9 points. LL=40 UL=60.



SMOKE03: The percentage of patients with any or any combination of the following conditions: coronary heart disease, stroke or TIA, hypertension, diabetes, COPD, CKD, asthma, schizophrenia, bipolar affective disorder or other psychoses whose notes record smoking status in the previous 15 months (except those who have never smoked where smoking status need only be recorded once since diagnosis)

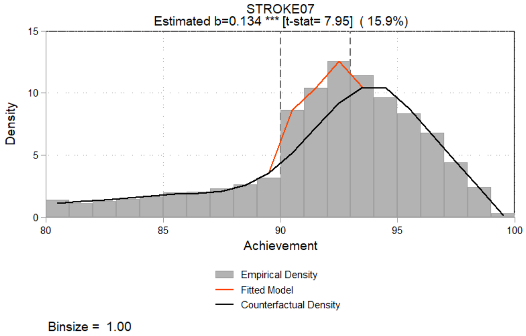
30 points. LL=40 UL=90.



SMOKE04: The percentage of patients with any or any combination of the following conditions: coronary heart disease, stroke or TIA, hypertension, diabetes, COPD, CKD, asthma, schizophrenia, bipolar affective disorder or other psychoses who smoke whose notes contain a record that smoking cessation advice or referral to a specialist service, where available, has been offered within the previous 15 months

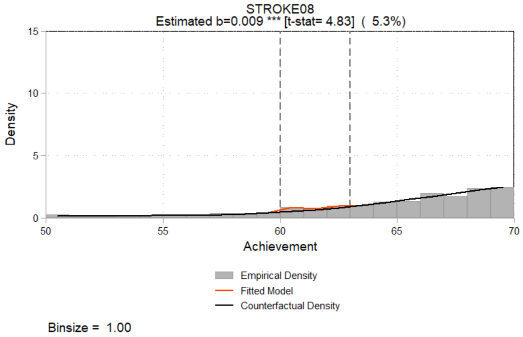
30 points. LL=40 UL=90.

Table E3: Bunching graphs of QOF clinical indicators which did not change between 2009-2010 and 2011 (Continued)



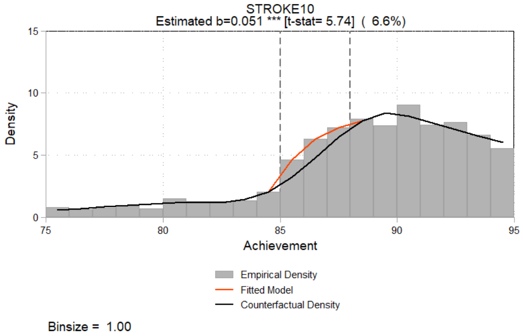
STROKE07: The percentage of patients with TIA or stroke who have a record of total cholesterol in the previous 15 months

2 points. LL=40 UL=90.



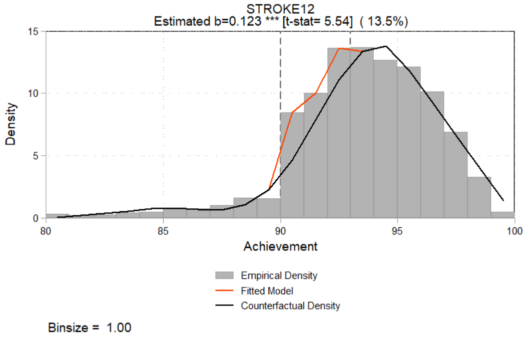
STROKE08: The percentage of patients with TIA or stroke whose last measured total cholesterol (measured in the previous 15 months) is 5 mmol/l or less

5 points. LL=40 UL=60.



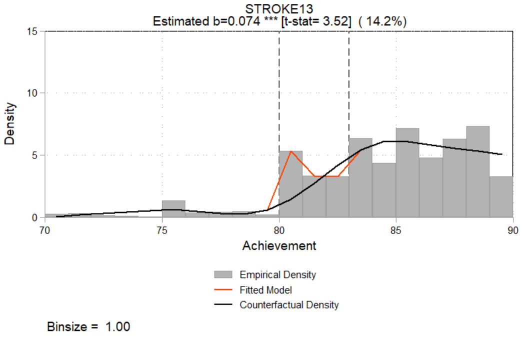
STROKE10: The percentage of patients with TIA or stroke who have had influenza immunisation in the preceding 1 September to 31 March

2 points. LL=40 UL=85.



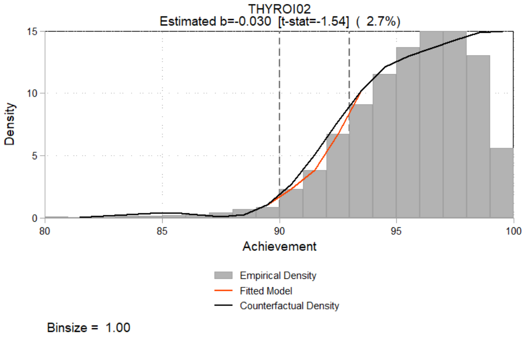
STROKE12: The percentage of patients with a stroke shown to be non-haemorrhagic, or a history of TIA, who have a record that an anti-platelet agent (aspirin, clopidogrel, dipyridamole or a combination), or an anti-coagulant is being taken (unless a contraindication or side-effects are recorded) *4 points. LL=40 UL=90.*

Table E3: Bunching graphs of QOF clinical indicators which did not change between 2009-2010 and 2011 (Continued)



STROKE13: The percentage of new patients with a stroke or TIA who have been referred for further investigation

2 points. LL=40 UL=80.



THYROI02: The percentage of patients with hypothyroidism with thyroid function tests recorded in the previous 15 months

6 points. LL=40 UL=90.

"

F Correlated unobservables and multiple hypothesis testing

The QOF clinical indicators are grouped by specific diseases (e.g. diabetes, chronic heart disease). Our DiD regressions can be modified to include in a system the regressions of all the indicators which are part of the same disease group, potentially improving efficiency by taking advantage of the correlation amongst unobservables of the indicators of the same disease group. We allow for correlation on unobservable characteristics across indicators by implementing a seemingly unrelated regression (SUR) per disease group. For this, we augment the econometric model specified in regression (5) of the main text to cover the entire domain of x_j: from 0 to 100. The model is:

$$\begin{aligned} x_{jit} - x_{jit-1} \quad = \quad & \alpha_{1j} \mathbb{1}\left\{x_{jit-1} \in [UL_j - 10, UL_j)\right\} + \alpha_{2j} \mathbb{1}\left\{t = 3\right\} + \alpha_{3j} \mathbb{1}\left\{x_{jit-1} \in [UL_j - 10, UL_j)\right\} \cdot \mathbb{1}\left\{t = 3\right\} \\ & + \alpha_{4j} \mathbb{1}\left\{x_{jit-1} \in [0, UL_j - 10)\right\} + \alpha_{5j} \mathbb{1}\left\{x_{jit-1} \in [0, UL_j - 10)\right\} \cdot \mathbb{1}\left\{t = 3\right\} \\ & + \alpha_{6j} \mathbb{1}\left\{x_{jit-1} \in (UL_j + 3, 100]\right\} + \alpha_{7j} \mathbb{1}\left\{x_{jit-1} \in (UL_j + 3, 100]\right\} \cdot \mathbb{1}\left\{t = 3\right\} + \alpha_{0j} + v_{jit}. \end{aligned}$$

The parameter α_3 is still the main parameter of interest as the comparison of interest is the difference between the set of sensitive PHCs, $\mathbb{1}\left\{x_{jit-1} \in [UL_j - 10, UL_j)\right\}$, and the set of insensitive ones, $\mathbb{1}\left\{x_{jit-1} \in [UL_j, UL_j + 3]\right\}$, in 2011 once such difference in 2010 has been netted out. The terms $\mathbb{1}\left\{x_{jit-1} \in [0, UL_j - 10)\right\}$ and $\mathbb{1}\left\{x_{jit-1} \in (UL_j + 3, 100]\right\}$ which do not appear when we do the regressions indicator by indicator -regression (5) of the main text- are included because there are PHCs which are located close to the kink for some indicators but not for others. These additional terms allows us to estimate the system for all the indicators of a disease group without restricting the sample to those PHCs which are close to the kink for all the indicators.

An additional concern is the possibility of false positives because of multiple hypothesis testing due to the number of indicators. In order to take it into account, we implemented a Romano-Wolf correction over the p-values of the main specification. In this scenario, we consider each illness group separately. This reduces the chances to reject the null hypothesis on illness groups which include several indicators such as asthma (3 indicators), diabetes (3 indicators) and stroke (5 indicators). (Romano and Wolf, 2016, 2005a,b).[3]

[3] P-values were derived after 1000 bootstrap repetitions, from a routine modified from Clarke (2016) STATA rwolf module.

Table F1: Estimates of the interaction term in regression (4) on the QOF dataset. Adjustment for multiple hypothesis testing.

Estimate of α_3 under the sample in $[UL - l, UL + k]$

Presents only indicators for which $\alpha_3 = 0$ is rejected in at least one specification.

	Entire interval				Removing $[UL - 1, UL + 1]$			
	k=3 pp. above UL		k=2 pp. above UL		k=3 pp. above UL		k=2 pp. above UL	
Indicator	l=10	l=5	l=10	l=5	l=10	l=5	l=10	l=5
AF03	-0.008^{**}	-0.004	-0.009^{**}	-0.004	-0.011^{**}	-0.005	-0.012^{**}	-0.006
	(0.004)	(0.003)	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)
	[0.044]							
AF04	-0.013^{**}	-0.014^{***}	-0.012^{**}	-0.014^{**}	-0.019^{**}	-0.027^{***}	-0.015	-0.023^{**}
	(0.006)	(0.005)	(0.006)	(0.005)	(0.009)	(0.009)	(0.010)	(0.009)
	[0.044]							
ASTHMA06	-0.012^{**}	-0.005	-0.009	-0.002	-0.016^{**}	-0.006	-0.013	-0.003
	(0.006)	(0.006)	(0.007)	(0.007)	(0.008)	(0.009)	(0.008)	(0.009)
	[0.1439]							
CKD06	-0.018^{***}	-0.015^{**}	-0.014^{**}	-0.011^{*}	-0.017^{***}	-0.014^{**}	-0.009	-0.006
	(0.005)	(0.006)	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)	(0.008)
	[0.000]							
COPD13	-0.007	-0.010^{*}	-0.006	-0.009^{*}	-0.006	-0.008	-0.006	-0.008
	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)	(0.007)	(0.006)	(0.007)
	[0.122]							
DM13	-0.005^{**}	-0.004^{*}	-0.003	-0.003	-0.007^{***}	-0.006^{*}	-0.006^{**}	-0.005
	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)
	[0.110]							
SMOKE04	-0.003	-0.004	-0.003	-0.005	-0.004	-0.007	-0.005	-0.008^{*}
	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)
	[0.364]							

Notes: Own calculations based on QOF data. Indicators presented in the table are those which were significant at the 90% level in at least one specification. Regressions are estimated using a SUR system as described in Appendix E. Standard errors clustered at PHC level in parenthesis. Significance levels (and p-values in brackets for column 1) were computed adjusting for multiple hypothesis testing at the disease group level, also see Appendix E for further details. Significance: * 1%, ** 5%, *** 1%.

G The role of the variance on the estimated results

The variance of the shocks impact the test ability to identify whether tasks are complements or substitutes. In this appendix we show, with simple simulations, how the estimates in both steps are affected when the variance increases. Simulations are based on the following scenario. This is the case of two tasks with a unique kink in the linear reward function for the first one. The optimization problem is:

$$\max_{e_1,e_2\in[0,1]} U = E_{\varepsilon_1}\left[(a_1^L \cdot x_1) \times \mathbb{1}(x_1 < UL) + (a_1^R * x_1 + (a_1^L - a_1^R) * UL) \times \mathbb{1}(x_1 \geq UL) + a_2 x_2\right]$$

$$s.t.$$

$$x_1 = e_1 + \varepsilon_1$$

$$x_2 = e_2$$

$$C(e_1, e_2; z) = \frac{1}{z} \cdot \left(\frac{1}{2} \cdot \left(c_1 \cdot e_1^2 + c_2 \cdot e_2^2\right) + \delta \cdot e_1 \cdot e_2\right)$$

where x_i corresponds to the realized achievement for task $i \in \{1, 2\}$, given an underlying effort e_i. For task 1, there is uncertainty on the achievement as it depends as well on a normally distributed random shock $\varepsilon_1 \sim N(0, \sigma^2)$. We consider a risk-neutral utility function for the provider using a quadratic cost function defined by parameters c_1, c_2 and δ. The optimization uses the SUBPLEX derivative-free algorithm implemented in NLOpt (similar results with other algorithms) (Johnson, 2007; Rowan, 1990).

Step 1. We compute decision rules (policy functions) on e_1 and e_2, given the value of z and the vector of parameters $(UL, c_1, c_2, \delta, a_1^R, a_1^L, a_2, \sigma)$. A Gauss-Legendre numerical integration with 40 nodes approximates the objective function expectation, assuming a domain for the shock between -10σ and 10σ.

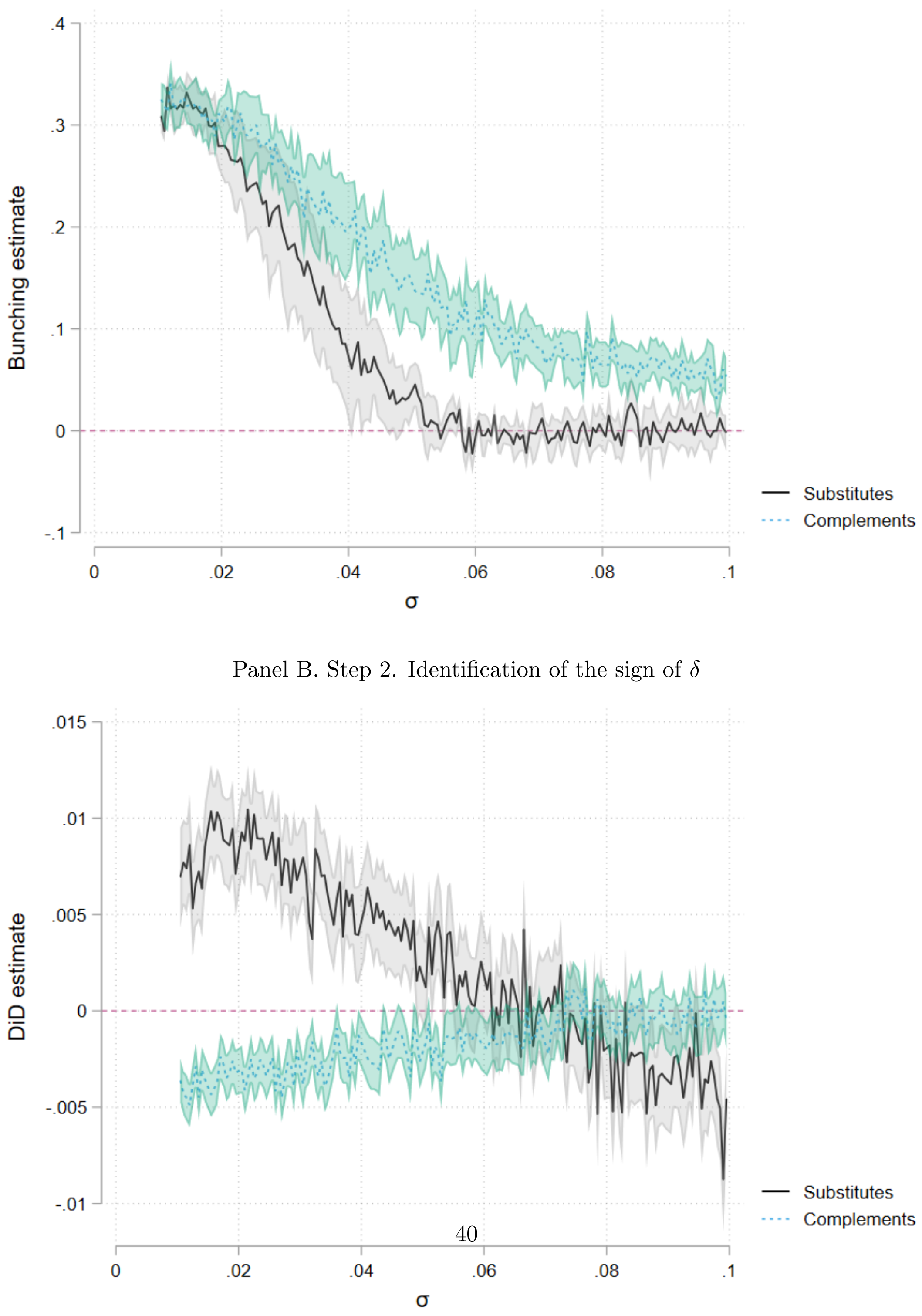
Step 2. We draw 10.000 observations of z from a uniform distribution between 0.1 and 4.98. For each, we compute the values of e_1, x_1, x_2 using the policy rules, considering three periods. Periods 1 and 2 with the same parameters, different shocks ε, but for period 3, there is a drop in the payment for task 2: $a_2^{t=3} < a_2^{t=1,2}$.

Step 3. Given the simulated dataset, we run the two empirical strategy steps: (I) detection of bunching and (ii) detection of the relative response to the change in incentives for task 2 on choices about effort 1.

We consider $UL = 0.5$ and bound our analysis for values of $x_1 \in (0, 1)$, dropping all other units. As a base scenario, we consider $a_1^L = 2.2$, $a_1^R = 1.2$, $a_2^{(t=1,2)} = 1.3$, $a_2^{(t=3)} = 1$. For the case of complements, cost parameters are $c_1 = 2, c_2 = 2, \delta = -1$, and for substitutes $c_1 = 8, c_2 = 8, \delta = 1$. We look for bunching in the interval $x_1 \in [0.45, 0.55]$ in the first step. For the second step, our reference group (insensible units at the kink) is composed of units exactly at $x_1 = 0.5$ in the second year; the 'treated' group (sensitive units) units are located in the range $x_1 \in [20, 40]$.

Figure G1 shows two graphs resulting from the simulations where we vary the size of the shock variance along the interval $\sigma \in [0.01, 0.1]$. In panel A, the estimated bunching size declines with a larger shock variance for both substitutes and complements. Panel B shows that the coefficient that indicates whether tasks are substitutes or complements also declines with variance. Hence, once bunching is not detected in Step 1, the coefficient in Step 2 is no longer informative about the type substitutability or complementarity of tasks.

Figure G1: Estimates of steps 1 and 2 with respect to the variance of the shocks on achievement. Evidence from simulations



Panel A. Step 1. Detection of bunching

Panel B. Step 2. Identification of the sign of δ

References

BMA (2013). Focus on qof payments. https://www.bma.org.uk/-/media/files/pdfs/practical%20advice%20at%20work/contracts/independent%20contractors/qof%20guidance/focusonqofpaymentsnov2013.pdf. Accesed: 2016-08-09.

Clarke, D. (2016). Rwolf: Stata module to calculate romano-wolf stepdown p-values for multiple hypothesis testing.

Johnson, S. G. (2007). The NLopt nonlinear-optimization package. https://github.com/stevengj/nlopt.

Kleven, H. J. (2016). Bunching. *Annual Review of Economics 8*(1).

Romano, J. P. and M. Wolf (2005a). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association 100*(469), 94–108.

Romano, J. P. and M. Wolf (2005b). Stepwise multiple testing as formalized data snooping. *Econometrica 73*(4), 1237–1282.

Romano, J. P. and M. Wolf (2016). Efficient computation of adjusted p-values for resampling-based stepdown multiple testing. *Statistics & Probability Letters 113*, 38–40.

Rowan, T. H. (1990). *Functional stability analysis of numerical algorithms.* Ph. D. thesis, Department of Computer Science, University of Texas at Austin, Austin, TX.

Saez, E. (2010). Do taxpayers bunch at kink points? *American Economic Journal: Economic Policy 2*(3), 180–212.