Quality-based payment in health care: Theory and practice

Quality-based payment in health care: Theory and practice

A dissertation presented

by

Samuel Starr Richardson

to

The Committee on Higher Degrees in Health Policy

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Health Policy

Harvard University

Cambridge, Massachusetts

July 2013

Dissertation Advisor: Thomas G. McGuire Samuel Starr Richardson

Quality-based payment in health care: Theory and practice

Abstract

Quality-based payment in healthcare—also known as pay-for-performance—is a popular policy intervention aimed at improving healthcare quality. However, there has been little theoretical work characterizing the underlying quality problem or the interaction between pay-for-performance and existing payment mechanisms. Furthermore, there is little empirical evidence that pay-for-performance has a substantial effect on healthcare quality.

In chapter 1, I develop a model of provider competition on two dimensions of quality and show that the efficient pay-for-performance contract corrects a market failure by rewarding dimensions of quality that are under-supplied in the existing system. I argue that provider allocation of effort to various tasks is inefficient without pay-for-performance, a multitasking problem that can be mitigated by an optimally designed pay-for-performance contract.

In 2004, U.K. National Health Service implemented the Quality and Outcomes Framework (QOF), a new contract that rewarded primary care practices based on a wide range of quality measures. In chapters 2 and 3, I use electronic medical record data from 357 practices to analyze how practices responded to the QOF.

Chapter 2 analyzes practice performance on quality of care for coronary heart disease, diabetes, and chronic kidney disease. I find that over the first two to three years of the QOF, overall quality of care improved in each domain, with improvements in both care processes and intermediate outcomes.

Chapter 3 investigates the specificity with which practices responded to the new incentives under QOF, focusing on the following thresholds: (1) performance ceilings above which practices do not receive additional payment, (2) test score thresholds that define success or failure on a quality indicator, and (3) end-of-fiscal-year effects introduced by annual reporting of results. I find discrete differences in provider behavior around each of these thresholds.

This dissertation makes three major contributions to the literature on quality-based payment, showing the following: (1) the design of quality-based payment contracts should consider interactions with the existing payment system; (2) a new quality-based payment contract in the United Kingdom was associated with improved quality of care; and (3) GP practices responded to specific marginal incentives, implying that quality-based payment should align incentives with desired provider responses.

Contents

Author List

Amber Batata contributed to Chapters 2 and 3.

Acknowledgments

I AM DEEPLY INDEBTED TO Tom McGuire, Meredith Rosenthal, and Amitabh Chandra for their generosity in providing guidance throughout the dissertation process. Harshvinder Bhullar provided invaluable insight into the THIN dataset, and Jingying Zhou provided excellent programming support with the data extraction process. Christopher Avery, Julian Jamison, and Martin Anderson provided insightful advice and suggestions. I am also grateful for helpful comments from seminar participants at the Harvard PhD in Health Policy Research Seminar and the Harvard Health Economics Lunch Seminar.

I am happy that my son, Toby, learned to accept that Daddy didn't always open the door to his home office, even when a cute toddler came knocking. Finally, I am eternally grateful to my wife, Robin, for putting up with six New England winters on my behalf, and for all her wonderful support throughout my doctoral career.

"The object of the medical profession today is to secure an income for the private doctor; and to this consideration all concern for science and public health must give way when the two come into conflict. Fortunately, they are not always in conflict."

— George Bernard Shaw

1

Integrating Pay-for-Performance into Health Care Payment Systems

1.1 Introduction

Health care quality in the U.S. and U.K. is widely believed to be inefficiently low, considering the resources devoted to the health care sector (McGlynn et al., 2003; Seddon et al., 2001). One commonly cited reason for this inefficiency is the inability of health care consumers to observe quality of care, which leads to low demand response to provider quality (Weisbrod, 1991). Pay-for-performance is one popular potential solution, whereby a payer directly rewards quality. The most ambitious pay-for-performance program to date is the U.K. National Health Service's (NHS) Quality and Outcomes Framework (QOF), where about a quarter of payments to general practitioners are based on performance on a range of quality indicators (Roland, 2004). In the U.S., accountable care organizations (ACOs), part of the Affordable Care Act of 2010, are an attempt to lower costs and improve quality by augmenting fee-for-service payment with quality-based payment and provider rewards (or penalties) for lower (or higher) than expected costs (McClellan et al., 2010). In this paper, I consider how payers should approach quality-based payment, in light of the market failures such payments are addressing and the interaction of quality-based payment with existing payment mechanisms.

The fundamental insight that I explore in this paper stems from the recognition that some attributes of health care are easier for consumers to observe than others. We should expect traditional payment systems to result in under-provision of those attributes that are poorly-observed by consumers, but not those that are well-observed. This insight seems not to have been considered by major payers implementing pay-for-performance contracts, which have included payments for those aspects of quality that are *best* observed by patients. In the QOF, up to 15% of quality-based payments have been based on patient experience, and the Institute of Medicine recommended that payments for patient-centeredness of care comprise roughly one-third of quality-based reimbursement for providers (Institute of Medicine, 2006).

Major payers seem to be providing quality-based payment based on the value associated with improved quality. However, as long as consumers base their choice of provider partially on provider quality,[1] and providers receive more net income when they attract more patients, providers already have incentives to invest in quality. In this paper, I recognize that providers have existing incentives to set quality, and treat optimal quality-based payment as a mechanism design problem. Incentives for quality may need correcting, but to do so we first need to characterize those incentives and understand the nature of the inefficiencies in existing payment systems. As mentioned above, existing payment systems may result in inefficiently low quality choices by providers due to imperfect observability of quality by consumers.[2] But in this case, bundled (prospective)

[1]There is extensive evidence that, at least in the case of hospital care, there is meaningful demand response to quality (Howard, 2006; Tay, 2003).

[2]We might also worry that provider market power will result in inefficient quality choices, even in the absence of imperfect information, as illustrated in the classic paper by Spence (1975). In this paper, I assume equal marginal

payments provide incentives to invest in well-observed aspects of quality (to which demand is responsive), and not in poorly observed aspects of quality. This is a classic multitasking problem, occurring in the absence of any quality-based payment.

Multitasking—in which contracts rewarding one dimension of effort reduce effort on other, unrewarded, dimensions—has long been a concern with pay-for-performance contracts (Holmstrom and Milgrom, 1991). Only some aspects of medical care are measurable by payers (and therefore contractible); depending on the health care production function, payment based on the measurable aspects of quality may reduce effort on unmeasurable (but still important) aspects. However, a health care payer observes consumer demand as well as some aspects of quality: depending on the nature of the demand function, quality-based payment may in fact be used to reduce multitasking problems associated with traditional, demand-based, payment contracts. Other papers have considered quality-based payment as a response to the multitasking problem in health care, but these papers have not addressed the possibility of demand response to quality informing optimal payment contracts (Eggleston, 2005; Kaarboe and Siciliani, 2011).

In a 1994 paper, Ma showed that bundled payment can achieve the efficient level of quality, so long as there is some positive demand response to quality. However, his conclusions depend on a one-dimensional model of quality; as I show in this paper, these conclusions fail when there are multiple dimensions of quality that are differently observable to patients. This is likely the case in most health care markets: for example, most patients observe how much time a physician spends with them (dimension 1), but few patients know whether the physician prescribed an appropriate medication (dimension 2). Demand will then be relatively more responsive to time spent with patients, compared to appropriate prescribing.[3] Once we go beyond one-dimensional

benefits across consumers, so facing informed consumers, the providers would choose efficient quality levels.

[3]Note that in some cases, such as in this example, the dimension that is relatively well-observed by consumers will be one that is poorly observed by payers. Unqualified references to "observability" in this paper will denote observability by consumers. Note also that observability by the payer depends on technology. For example,

quality, bundled payment is no longer sufficient to achieve the efficient quality level: there is a multitasking problem in which providers have an incentive to over-provide the dimensions of quality that are well-observed by patients, relative to dimensions of quality that patients observe poorly.

In this paper, I consider a model of quality competition between two profit-maximizing providers, in which there are two dimensions of quality that are imperfectly observed by patients. A payer offers a payment contract to the providers, attempting to induce efficient provider choices on both dimensions of quality. Under a traditional bundled (demand-based) payment system, providers over-invest in the better-observed dimension of quality. However, once a payer can implement pay-for-performance, it is possible to induce the efficient level of both dimensions of quality. Critically, this depends on rewarding the poorly observed dimension of quality (or penalizing the well-observed dimension of quality); it does not involve rewarding the dimension of quality that contributes more to patient health outcomes or patient utility.

In their chapter on physician pay-for-performance, Golden and Sloan (2008) present a list of system design questions, including the following: "Which outcome measures will be part of the payment scheme, and by implication, which are considered either less important or too difficult to measure reliably?" This reflects the common understanding that when designing pay-for-performance contracts, we should reward dimensions of quality that are important (presumably in terms of improving health outcomes). In this paper, I will argue that this understanding is incomplete: quality-based payments should reward those aspects of quality with the greatest inefficiencies caused by existing payment mechanisms. These targeted dimensions of quality may or may not be the most important in terms of their effect on health outcomes.

with near-universal adoption of electronic medical records among primary care practices in the U.K., it is feasible for the NHS to observe details of medical records that would be prohibitively costly for Medicare to observe. Finally, observability by payers may depend on satisfying incentives for honest reporting of treatment by providers and patients (Ma and McGuire, 1997).

1.2 Experience with pay-for-performance

Although there has been enthusiasm for pay-for-performance from both public and private payers (Institute of Medicine, 2006), the evidence on provider responses to quality-based payment is decidedly mixed (Rosenthal and Frank, 2006). Most studies tend to find little or no quality improvements associated with pay-for-performance, but nor does multitasking seem to have been a big problem in general. That said, all pay-for-performance programs thus far have been plagued by low marginal incentives for providers: either there is little money on the table, relative to provider income; or targets have been set such that most providers do not need to improve quality to meet the targets.

Starting in 2002, major private payers in California began providing direct incentives for performance on a variety of process-based quality measures: PacifiCare implemented its Quality Improvement Program (QIP) in 2002, and was joined by several other payers (the Integrated Healthcare Association, or IHA) in 2003. Payment under these programs was relatively low; although the IHA was responsible for about 60% of medical groups' capitated revenues, annual quality-based payments per patient never rose above $18. Two papers analyzing the QIP and IHA initiatives find little evidence of overall improvement in performance or of multitasking problems (Mullen et al., 2010; Rosenthal et al., 2005). Interestingly, with the QIP, the lowest-performing medical groups improved quality the most, despite the contract providing little additional marginal incentive for them to improve. (Payment was given to practices that reached a set threshold, so few of the low-performing practices were likely to receive payment.) Although this could be attributable to regression to the mean, it perhaps suggests that factors other than marginal financial incentives or public reporting are affecting the results (in this program, public reporting had been in place for several years before pay-for-performance implementation).

Several papers have attempted to evaluate the effects of the NHS QOF, and there is suggestive evidence that some processes of care may have improved, with perhaps only minor multitasking problems (Campbell et al., 2009; Doran et al., 2011). However, the highest-powered and best-identified study to date estimates a precise zero-effect on various processes and health outcomes among patients with hypertension (Serumaga et al., 2011). The quality-based payments under the QOF are large; starting with full implementation in 2005, the maximum payment for an average practice was roughly £131,250 ($230,000). However, it seems that the marginal incentives for improvement were small: the median practice earned 95.5% of the available payments in the first year after implementation (Doran et al., 2006).[4]

1.3 Model

I consider a model with two-dimensional quality, (q_1, q_2). A social planner is the only payer for health care services, and chooses a payment schedule. Depending on regulatory and informational constraints, payment may be based on patient demand (quantity), cost, and/or quality. There are two providers, each having the same constant-returns-to-scale technology and choosing quality levels (q_1, q_2), with strictly convex cost per patient $c(q_1, q_2)$. Quantity demanded is determined by quality competition between providers at fixed locations, with provider a located at point 0, provider b located at point 1, and unit measure of consumers distributed uniformly on $[0, 1]$. Each consumer has unit demand,[5] and a consumer traveling distance d to provider j receives utility $q_{1j} + q_{2j} - td$, where t is a known, strictly positive cost-of-travel parameter. Each consumer has independent probabilities (λ_1, λ_2) of observing quality dimensions (q_1, q_2) of both providers.[6] Each

[4] It is possible that the marginal incentives were large, and practices simply improved drastically in the first year after implementation; however, it is undeniable that marginal incentives for improvement were small in all years after the first.

[5] Here we can think of a consumer demanding health care for a given period of time, and signing up with a provider for that time. Alternatively, we may think of a sick consumer demanding health care from a single provider for a given episode of treatment. In either case, aspects of care that are often considered to be quantity (such as number of visits or number of procedures) will contribute to the *quality* of the single unit of care demanded. When providers have the ability to set quantity, as is typically assumed to be the case in many health care markets (McGuire, 2000), there is no fundamental distinction between quantity- and quality-based competition (Tirole, 1988).

[6] In my model, it is this imperfect observability of quality that results in low demand response to quality (relative to what demand response would be if consumers fully valued the benefit they received from higher quality care). However, the conclusions of this paper are not dependent on the mechanism underlying the low demand

consumer has beliefs about any quality dimensions that are unobserved to that consumer (and in equilibrium these beliefs will be correct), and chooses the provider who maximizes the consumer's utility. Prices faced by consumers are administratively set such that no patient chooses an outside option.

Thus, the game proceeds in three stages:

1. Social planner chooses a provider payment contract (in each case I consider below, several of these payment parameters will be constrained to equal zero):

   (a) Bundled payment p_b is paid per unit of demand.

   (b) Quality payments per unit demand p_1 and p_2 are paid per unit of q_1 and q_2, respectively.

   (c) Cost-based reimbursement $0 \leq p_c < 1$ is reimbursement as a percentage of provider costs.

2. Profit-maximizing[7] providers a and b choose non-negative quality vectors (q_{1a}, q_{2a}) and (q_{1b}, q_{2b}), respectively. Provider j's cost per patient is a strictly increasing, strictly convex, continuously differentiable function $c\left(q_{1j}, q_{2j}\right)$, with $c\,(0,0) = 0$, $c_1\,(0,x) = 0$, and $c_2\,(x,0) = 0$, $x \in [0, \infty)$. (Note that c_i represents the partial derivative of $c\left(q_{1j}, q_{2j}\right)$ with respect to its i'th argument.) Alternatively, providers may choose to exit the market and receive zero profits.

3. Consumers have independent probabilities $\lambda_1, \lambda_2 \in (0,1)$ of observing quality dimensions (q_1, q_2) of both providers. A consumer at point i choosing provider a receives utility $u\,(q_{1a}, q_{2a}, i) = q_{1a} + q_{2a} - ti$; the same consumer choosing provider b

response to quality. Other plausible mechanisms include the presence of positive externalities to medical treatment or lack of information about the value associated with higher quality care. Note further that in my model, the independence of λ_1 and λ_2 is not necessary for any of the conclusions to go through; this assumption simply affords greater ease of exposition.

[7] Most models of healthcare providers' utility functions include some degree of altruism (McGuire, 2000); the profit-maximization assumption does not substantially affect my results, unless provider altruism favors one dimension of quality over the other. So long as altruism is modeled as the patient's utility entering the provider's utility function, altruism will simply result in lower payments being required to attain efficient quality levels.

receives utility $u\left(q_{1b}, q_{2b}, i\right) = q_{1b} + q_{2b} - t \cdot (1-i)$. Each consumer has beliefs about provider quality choices that are unobserved by that consumer, and according to those beliefs chooses the provider that maximizes the consumer's utility.

Consumer choices in stage 3 imply a demand function for provider j of $\mu_j\left(q_{1j}, q_{2j}, q_{1,-j}, q_{2,-j}\right)$. Profit to provider j is:

$$
\begin{aligned}
\pi_j &= p_1 q_{1j}\mu_j + p_2 q_{2j}\mu_j + p_b \cdot \mu_j + p_c \cdot c\left(q_{1j}, q_{2j}\right) \cdot \mu_j - c\left(q_{1j}, q_{2j}\right) \cdot \mu_j \\
&= \mu_j \cdot \left[p_1 q_{1j} + p_2 q_{2j} + p_b - (1-p_c) \cdot c\left(q_{1j}, q_{2j}\right)\right]
\end{aligned}
$$

Since payment to providers is simply a transfer, net social welfare is given by:

$$
SW\left(q_{1a}, q_{2a}, q_{1b}, q_{2b}\right) = \int_{i=0}^{1} u_i\left(q_{1a}, q_{2a}, q_{1b}, q_{2b}\right) di - \mu_a \cdot c\left(q_{1a}, q_{2a}\right) - \mu_b \cdot c\left(q_{1b}, q_{2b}\right)
$$

1.4 Characterization of equilibrium

In this section, I solve for Perfect Bayesian Equilibrium under different regulatory regimes. Depending on contractibility, the social planner will be constrained to set some of the payment parameters equal to zero. In the subsections below, I derive demand, solve for providers' profit-maximizing quality choices, and solve for the planner's constrained optimum under various regulatory regimes.

1.4.1 Demand

First, I derive demand for provider j as a function of $\left(q_{1a}, q_{2a}, q_{1b}, q_{2b}\right)$. I denote as $\tilde{q}_{kj}$ the consumer's belief about provider j's choice on quality dimension k. The probability that consumer i chooses provider a is:[8]

$$\begin{aligned}
\mu_{ia} &= (1-\lambda_1)(1-\lambda_2)\,\mathbb{1}_{[\tilde{q}_{1a}+\tilde{q}_{2a}-ti\geq\tilde{q}_{1b}+\tilde{q}_{2b}-t(1-i)]}+\\
&\quad \lambda_1(1-\lambda_2)\,\mathbb{1}_{[q_{1a}+\tilde{q}_{2a}-ti\geq q_{1b}+\tilde{q}_{2b}-t(1-i)]}+\\
&\quad (1-\lambda_1)\,\lambda_2\,\mathbb{1}_{[\tilde{q}_{1a}+q_{2a}-ti\geq\tilde{q}_{1b}+q_{2b}-t(1-i)]}+\\
&\quad \lambda_1\lambda_2\,\mathbb{1}_{[q_{1a}+q_{2a}-ti\geq q_{1b}+q_{2b}-t(1-i)]}
\end{aligned}$$

Assuming that absent additional information, consumers believe the two providers choose the same quality vectors,[9] this reduces to:

$$\begin{aligned}
\mu_{ia} &= (1-\lambda_1)(1-\lambda_2)\,\mathbb{1}_{\left[i\leq\frac{1}{2}\right]}+\lambda_1(1-\lambda_2)\,\mathbb{1}_{\left[i\leq\frac{1}{2}+\frac{q_{1a}-q_{1b}}{2t}\right]}+\\
&\quad (1-\lambda_1)\,\lambda_2\,\mathbb{1}_{\left[i\leq\frac{1}{2}+\frac{q_{2a}-q_{2b}}{2t}\right]}+\lambda_1\lambda_2\,\mathbb{1}_{\left[i\leq\frac{1}{2}+\frac{q_{1a}+q_{2a}-q_{1b}-q_{2b}}{2t}\right]}
\end{aligned}$$

Define $g(x)=\max(0,\min(1,x))\ \ \forall x\in\mathbb{R}$. Demand for provider a is given by the following semi-differentiable function:

$$\begin{aligned}
\mu_a &= \int_{i=0}^{1}\mu_{ia}di\\
&= (1-\lambda_1)(1-\lambda_2)\frac{1}{2}+\lambda_1(1-\lambda_2)\cdot g\left(\frac{1}{2}+\frac{q_{1a}-q_{1b}}{2t}\right)+\\
&\quad (1-\lambda_1)\,\lambda_2\cdot g\left(\frac{1}{2}+\frac{q_{2a}-q_{2b}}{2t}\right)+\lambda_1\lambda_2\cdot g\left(\frac{1}{2}+\frac{q_{1a}+q_{2a}-q_{1b}-q_{2b}}{2t}\right)
\end{aligned}$$

Since there is unit measure of consumers, each with unit demand, demand for provider b is given by $\mu_b=1-\mu_a$. Note that with $(\lambda_1,\lambda_2)\ll(1,1)$, demand is strictly positive for both providers. Furthermore, define $\frac{\partial\mu_a}{\partial q_{ka}}=\frac{\partial\mu_b}{\partial q_{kb}}=\frac{\partial\mu}{\partial q_k}$, $k\in\{1,2\}$. Finally, note that where

[8] $\mathbb{1}_{[\textit{logical expression}]}$ is an indicator function that takes the value 1 if *logical expression* is true and 0 otherwise.

[9] Note that there could be other reasonable off-equilibrium-path beliefs (for example, a consumer observing a provider choosing higher-than-expected q_1 might also believe the provider chose higher-than-expected q_2. However, if a patient observing one dimension of quality can infer the other dimension, both dimensions have become equally observable.

$\frac{\partial \mu}{\partial q_k}$ is undefined, we have $\frac{\partial_+ \mu_a}{\partial q_{ka}} = \frac{\partial_- \mu_b}{\partial q_{kb}}$ and $\frac{\partial_- \mu_a}{\partial q_{ka}} = \frac{\partial_+ \mu_b}{\partial q_{kb}}$.

1.4.2 Provider profit-maximization

Provider a's profit-maximization problem is then given as:

$$\begin{aligned}&\max_{q_{1a},q_{2a}} \pi_a \left(q_{1a}, q_{2a}; q_{1b}, q_{2b}, \lambda_1, \lambda_2, t\right) \\ &= \mu_a \left(q_{1a}, q_{2a}; q_{1b}, q_{2b}, \lambda_1, \lambda_2, t\right) \cdot \left[p_1 q_{1a} + p_2 q_{2a} + p_b - (1 - p_c) \cdot c\left(q_{1a}, q_{2a}\right)\right]\end{aligned}$$

First-order conditions for a maximum are given by:

$$\begin{aligned} k \in \{1, 2\}, \ \frac{\partial \pi_a}{\partial q_{ka}} &= \frac{\partial \mu_a}{\partial q_{ka}} \cdot \left[p_1 q_{1a} + p_2 q_{2a} + p_b - (1 - p_c) \cdot c\left(q_{1a}, q_{2a}\right)\right] + \\ &\quad \mu_a \cdot \left[p_k - (1 - p_c) \cdot \frac{\partial c}{\partial q_{ka}}\right] \leq 0 \qquad (1.4.1) \end{aligned}$$

$$q_{ka} \cdot \frac{\partial \pi_a}{\partial q_{ka}} = 0$$

Given $c(0,0) = 0$, $c_1(0,x) = 0$, and $c_2(x,0) = 0$, $x \in [0, \infty)$: if p_1, p_2, and p_b are non-negative and if $\frac{\partial \mu_a}{\partial q_{ka}}$ is defined then expression (1.4.1) must hold with equality in equilibrium.[10] The critical points where $\frac{\partial \mu_a}{\partial q_{ka}}$ is undefined occur where $q_{ka} = q_{kb} \pm t$ or $q_{1a} + q_{2a} = q_{1b} + q_{2b} \pm t$.

The physician's individual rationality constraint is given by
$p_1 q_1 + p_2 q_2 + p_b - (1 - p_c) \cdot c(q_1, q_2) \geq 0$.

Proposition. *Given a single payment contract, any equilibrium must be symmetric:* $q_{1a} = q_{1b}$; $q_{2a} = q_{2b}$. Proof in appendix.

Note that at a symmetric equilibrium, we have a linear demand curve with respect to

[10] Later in the paper, we will consider cases where the planner sets a negative value for p_1, p_2, or p_b. However, it will never be socially optimal for the planner to set these values sufficiently negative that a corner solution is induced. This follows from $c(0,0) = 0$, $c_1(0,x) = 0$, and $c_2(x,0) = 0$.

each quality dimension: $\frac{\partial \mu}{\partial q_k} = \frac{\lambda_k}{2t}$. Substituting into our first-order conditions yields:

$$\frac{\lambda_k}{2t} \cdot [p_1 q_{1a} + p_2 q_{2a} + p_b - (1 - p_c) \cdot c(q_{1a}, q_{2a})] + \mu_a \left[p_k - (1 - p_c) \cdot \frac{\partial c}{\partial q_{ka}} \right] \leq 0$$

Second-order conditions are given by:

$$\begin{aligned}
SOC_1: \quad & \tfrac{\lambda_1}{t} \cdot [p_1 - (1 - p_c)\, c_1] \leq \mu \cdot (1 - p_c)\, c_{11} \\
SOC_2: \quad & \tfrac{\lambda_2}{t} \cdot [p_2 - (1 - p_c)\, c_2] \leq \mu \cdot (1 - p_c)\, c_{22} \\
SOC_3: \quad & \left[\tfrac{\lambda_1}{t} \cdot [p_1 - (1 - p_c)\, c_1] - \mu \cdot (1 - p_c)\, c_{11} \right] \cdot \\
& \left[\tfrac{\lambda_2}{t} \cdot [p_2 - (1 - p_c)\, c_2] - \mu \cdot (1 - p_c)\, c_{22} \right] \geq \\
& \left[\tfrac{\lambda_1}{2t} [p_2 - (1 - p_c)\, c_2] + \tfrac{\lambda_2}{2t} [p_1 - (1 - p_c)\, c_1] - \mu \cdot (1 - p_c)\, c_{12} \right]^2
\end{aligned}$$

At any stationary point, the following two conditions are sufficient to satisfy the second-order conditions: (1) $p_c \leq 1$, and (2) $p_k \leq (1 - p_c)\, c_k,\ k \in \{1, 2\}$. Condition (1) holds at any solution to the first-order conditions. Condition (2) will hold at a solution to the first-order conditions when $p_1 q_{1a} + p_2 q_{2a} + p_b - (1 - p_c) \cdot c(q_{1a}, q_{2a}) \geq 0$; that is, the net income per patient is positive, which is required to satisfy the provider's individual rationality constraint.

1.4.3 Regulation

The social planner's problem is to maximize social welfare subject to the constraints imposed by provider profit-maximizing. Note that given a symmetric solution (and correct beliefs on the equilibrium path), all consumers with $i \leq \frac{1}{2}$ will choose provider a, with the remainder choosing provider b. This implies that the planner cannot affect the travel costs to consumers, and the social planner's objective reduces to:

$$\max\ SW = q_1 + q_2 - c(q_1, q_2)$$

This is a strictly concave objective function, and the first-best solution has $c_1 = c_2 = 1$. The social planner's constraints for an interior solution are given by:

$$\begin{aligned}
FOC_1: &\quad (1-p_c)\cdot\frac{\partial c}{\partial q_1} - p_1 = \frac{\lambda_1}{t}\cdot\left[p_1 q_1 + p_2 q_2 + p_b - (1-p_c)\cdot c\left(q_1, q_2\right)\right] \\
FOC_2: &\quad (1-p_c)\cdot\frac{\partial c}{\partial q_2} - p_2 = \frac{\lambda_2}{t}\cdot\left[p_1 q_1 + p_2 q_2 + p_b - (1-p_c)\cdot c\left(q_1, q_2\right)\right] \\
IR: &\quad p_1 q_1 + p_2 q_2 + p_b - (1-p_c)\cdot c\left(q_1, q_2\right) \geq 0
\end{aligned}$$

This can be thought of as an instruments and targets problem: the socially optimal values of q_1 and q_2 are the two targets, and the planner will generally need two independent instruments to achieve the targets.

1.4.4 Policy with a mix of bundled and cost-based payment (the traditional NHS model)

From the inception of the NHS until 2004, primary care practices were paid almost entirely based on capitation (bundled payment). In this subsection, I will consider a case where the planner is constrained to set all reimbursement other than p_b and p_c equal to zero, and show that no such mix of bundled and cost-based payment can induce efficient quality choices by providers unless $\lambda_1 = \lambda_2$. (The traditional NHS model is the specific case where $p_c = 0$.) The planner's problem is:

$$\max_{p_b,\, p_c} SW = q_1 + q_2 - c\left(q_1, q_2\right)$$

$$\begin{aligned}
s.t.\ FOC_1: &\quad (1-p_c)\cdot c_1 = \frac{\lambda_1}{t}\left[p_b - (1-p_c)\, c\right] \\
FOC_2: &\quad (1-p_c)\cdot c_2 = \frac{\lambda_2}{t}\left[p_b - (1-p_c)\, c\right] \\
IR: &\quad p_b \geq (1-p_c)\, c
\end{aligned}$$

If $\lambda_1 = \lambda_2 = \lambda$, then the first-best is achievable by setting $\frac{p_b}{1-p_c} = c^* + \frac{t}{\lambda}$, where c^* is the value of the cost function where $c_1 = c_2 = 1$. Otherwise, the first-best is not achievable, since with $\lambda_1 \neq \lambda_2$, $c_1 \neq c_2$. Although we have two instruments (p_b and p_c) and two targets (q_1 and q_2) the instruments are collinear with respect to the targets. In this case, providers choose quality such that $\frac{c_1}{c_2} = \frac{\lambda_1}{\lambda_2}$: there is a multitasking problem, with over-investment in the better-observed dimension of quality, relative to the less-observed dimension of quality.

Ma (1994) showed that bundled payment can be set to achieve the right overall *level* of quality, with a similar result that the optimal bundled payment varies with the inverse of the demand response to quality. However, my model adds an additional dimension of quality, and when there is differential demand response to different dimensions of quality ($\lambda_1 \neq \lambda_2$), it is impossible to achieve the first-best with bundled payment. Bundled payment is thus unable to address the multitasking problem associated with differential demand response to different aspects of quality. In order to do so, we will need an additional policy instrument.

1.4.5 Policy with treatment-intensity-based payment (the traditional Medicare model)

Since the implementation of the Resource Based Relative Value Scale for Part B Medicare payments in 1992, physician payments have been based on the total number of procedures and Medicare's estimate of the average resources needed to provide those procedures. Note that this is not cost-based reimbursement, since a physician who uses fewer resources to provide a given procedure pockets the difference in resources used. This can be modeled by thinking of the number and intensity of procedures provided to a patient as q_1 and the quality of those procedures (or possibly coordination of care) as q_2. In this case the planner is constrained to set all reimbursement other than p_1 equal to zero. The planner's problem is:

$$\max_{p_1} SW = q_1 + q_2 - c\left(q_1, q_2\right)$$

$$
\begin{aligned}
s.t.\ FOC_1: &\quad c_1 - p_1 = \tfrac{\lambda_1}{t}\left[p_1 q_1 - c\right] \\
FOC_2: &\quad c_2 = \tfrac{\lambda_2}{t}\left[p_1 q_1 - c\right] \\
IR: &\quad p_1 q_1 \geq c
\end{aligned}
$$

The first-best is achievable if and only if $\lambda_2 = \lambda_1 + \frac{\lambda_2 c^* + t}{q_1^*} = \frac{t + \lambda_1 q_1^*}{q_1^* - c^*} > \lambda_1$. Unless λ_1 is much lower than λ_2, treatment-intensity-based payment will result in over-provision of q_1 relative to q_2.[11]

Compared to the results in subsection 1.4.4 above where p_b rewards the dimension of quality to which demand is more responsive, in this case, p_1 rewards both overall demand and q_1 specifically. Here we have providers choosing quality such that $\frac{c_1 - p_1}{c_2} = \frac{\lambda_1}{\lambda_2}$, compared to $\frac{c_1}{c_2} = \frac{\lambda_1}{\lambda_2}$ in subsection 1.4.4. This implies that, compared to bundled payment, payment based on q_1 will get us closer to the social optimum where $c_1 = c_2 = 1$ only when λ_1 is substantially lower than λ_2.

1.4.6 Policy with bundled and quality-based payment (the new Medicare and NHS models)

Under the NHS QOF, payment to primary care practices is a mix of capitation and quality-based payment. The QOF pays for a wide range of quality measures, but no quality-based payment contract can pay for all dimensions of quality.

Payment under Medicare ACOs is based on Medicare's traditional intensity-based payment (paying on q_1). However, a provider who costs less than expected (provides a lower q_1) shares in the savings to the system, and a provider who costs more than expected (provides a higher q_1) is not paid as much as under traditional Medicare payment.[12] This can be modeled as bundled payment, plus lower p_1 than was the case in

[11] Note that I have assumed linear quality-based rewards: the payment function $f(q_1) = p_1 q_1$. If we remove the restriction of linearity and allow any payment function, we can achieve the first-best by inserting a step discontinuity at q_1^* and setting $f(q_1^*) = c^* + \frac{t}{\lambda_2}$. If $\lambda_1 \geq \lambda_2$, then it is necessary for $f(q_1)$ to decrease above q_1^*.

[12] This is a description of one of the two payment options for ACOs, in which providers have symmetric incentives above and below the traditional level of q_1. Another option allows providers to share in savings, but

the traditional Medicare payment model.[13]

In either system, the new payment model allows the planner to choose positive values for p_b and p_1, but must set all other payment parameters to zero. The planner's problem is now:

$$\max_{p_b, p_1} SW = q_1 + q_2 - c\left(q_1, q_2\right)$$

$$\begin{aligned} s.t.\ FOC_1: &\quad c_1 - p_1 = \tfrac{\lambda_1}{t}\left(p_b + p_1 q_1 - c\right) \\ FOC_2: &\quad c_2 = \tfrac{\lambda_2}{t}\left(p_b + p_1 q_1 - c\right) \\ IR: &\quad p_b + p_1 q_1 \geq c \end{aligned}$$

Can the social planner achieve the point where $c_1 = c_2 = 1$?

$$\begin{aligned} FOC_1: &\quad 1 - p_1 = \tfrac{\lambda_1}{t}\left(p_b + p_1 q_1 - c\right) \\ FOC_2: &\quad 1 = \tfrac{\lambda_2}{t}\left(p_b + p_1 q_1 - c\right) \end{aligned}$$

Solving yields:

$$\begin{aligned} p_1 &= 1 - \frac{\lambda_1}{\lambda_2} \\ p_b &= c^* + \frac{t}{\lambda_2} - p_1 q_1^* = c^* + \frac{t}{\lambda_2} - \left(1 - \frac{\lambda_1}{\lambda_2}\right) q_1^* \end{aligned}$$

This contract achieves the first-best quality choices by providers, but note that when $\lambda_1 > \lambda_2$, $p_1 < 0$. If there is a non-negativity constraint on p_1, then $p_1^* = 0$, and the first-best is not achievable (there is over-investment in q_1, relative to q_2, as in subsection 1.4.4).

---

not be accountable for increased costs; this can be modeled as bundled payment plus payment based on q_1, where the quality-based payment formula is kinked at the level of q_1 from the traditional payment system.

[13]Note that this treats the market as if consumers sign up with a given provider. In practice, consumers will be assigned to ACOs based on where they receive most of their treatment (*ie* where they receive higher q_1). This clearly provides incentives to increase q_1 to be attributed with patients (or possibly to reduce q_1 to avoid attribution). It is beyond the scope of this paper to consider the trade-off between this type of manipulation and increased consumer choice of providers.

Furthermore, for extreme parameter values with $\lambda_1 < \lambda_2$, it can be the case that $p_b^* < 0$.

Similar to the results from Ma (1994), if $\lambda_1 = \lambda_2$, then quality can be thought of as uni-dimensional, and the optimal contract includes only bundled payment. If we decrease λ_1, then bundled payment results in a multitasking problem: providers over-invest in q_2 relative to q_1. Increasing p_1 provides additional incentive to increase q_1, and by reducing p_b by $p_1 q_1$, we maintain the right overall level of reimbursement. Increasing λ_1 results in the reverse problem, and we need to impose negative p_1 while increasing p_b.

Note that $p_1 < 1$, implying that the optimal payment for a marginal unit of quality is less than the benefit of that marginal unit to patients. At the optimal quality level, the marginal benefit of quality to the consumer is equal to the marginal cost to the provider. However, the quality-based payment to the provider need not cover the entire marginal cost (and, in fact, should not, since setting $p_1 = 1$ would result in inefficiently high provision of q_1). Only in the case of zero demand response to q_1 is it optimal to set marginal quality-based payment equal to marginal benefit. When there is a demand response, the provider is rewarded for increasing q_1 not only through the direct payment for quality p_1, but also through an increase in demand. Here there is an important interaction between quality-based payment and the existing incentive structure.

In this section, I've shown that quality-based payment can be used as a policy instrument to address inefficiencies associated with traditional payment models. However, the level of quality-based payment is dependent on the nature of the inefficiency (in this case the unequal observability of the different quality dimensions). Optimal quality-based payment does not involve paying based on the marginal benefit of quality; rather, it varies with the demand response to different dimensions of quality, and it is plausible for the optimal quality-based payment to be negative. Quality-based payment should be targeted at those aspects of quality that are least rewarded by the existing payment system, relative to their social benefit.

1.5 Extensions

The model above admittedly leaves out many potentially important characteristics of health care markets, but the basic point that quality-based payment should be targeted at addressing inefficiencies in the existing payment system is likely robust. The result in section 1.4.6 that the planner can achieve the first best clearly depends on the number of independent instruments equaling or exceeding the number of targets. A more realistic case is where there are more than two dimensions of quality, and the number of instruments is less than the number of targets, in which case the solution will be second-best. However, the optimal policy will still depend on the observability of different quality dimensions.[14] Other modifications to the model with fairly straightforward implications include allowing for provider altruism or considering effects of payment contracts on provider entry, exit, and location decisions. In the rest of this section, I consider two generalizations: patient heterogeneity in health status, and deadweight losses associated with raising funds.

1.5.1 Patient heterogeneity in health status

Consider the following model of heterogeneity in patient health status. The unit measure of consumers is divided into ς sick types and $(1-\varsigma)$ healthy types; $0<\varsigma<1$. Sick types have the same utility function as above, but healthy types only receive benefit from quality dimension q_1: a healthy consumer traveling distance d to a physician choosing quality (q_1, q_2) receives utility $q_1 - ti$. Each provider's cost per sick patient is $c\left(q_1, q_2\right)$, and the cost per healthy patient is $c\left(q_1, 0\right)$. Note that I assume providers must choose the same value of q_1 for all patients, but for all practical purposes choose $q_2=0$ for healthy types.[15]

[14] When there are fewer instruments than targets, the optimal contract will also depend on whether different dimensions of quality are substitutes or complements. The planner will want to reward poorly observed dimensions of quality, as well as dimensions that are complements of other poorly observed dimensions.

[15] In this context, we can think of q_1 as representing screenings and other care provided to all patients, and q_2 as representing chronic care management that is only provided to sick types.

Now there are two sources of market failure if payment takes the form of flat (non-risk-adjusted) bundled payment. The multitasking issues arising from differential demand response that were highlighted in the previous section are still in force. In addition, there is a creaming/skimping problem (as described in Ellis, 1998) in which providers over-invest in q_1 relative to q_2 in order to attract healthier, less costly patients. If $\lambda_1 < \lambda_2$, these distortions move in opposite directions and the relative magnitude of the distortions will determine whether there is over- or under-investment in q_1 relative to q_2. Otherwise, if $\lambda_1 \geq \lambda_2$, then there is over-investment in q_1.

Several different payment instruments could be used to address these distortions (note that there are still two targets, so we will generally need two independent instruments). As in section 1.4.6, we can achieve the first-best through implementing bundled payment plus payment on one dimension of quality. The level of quality-based payment here is dependent both on the relative observability of the two quality dimensions, as well as on the difference in cost between healthy and sick types. In this case, the optimal quality-based payment rewards the less-observed dimension of quality, as well as the dimension of quality that sick types care about relatively more.

Alternatively, if the planner can observe patient type, risk-adjusted bundled payment can lead to the first-best. Note that only in the case where $\lambda_1 = \lambda_2$ will optimal risk adjustment result in payment equal to cost for each type plus a constant. In fact, in some cases with $\lambda_1 > \lambda_2$, optimal payments for treating healthy types will be below the cost of providing care to those types. This observation that risk adjustment can drive quality choices as well as selection incentives has been illustrated by Glazer and McGuire (2000).

Another issue that arises once we have heterogeneity in health status is the possibility that the planner should induce specialization by providers, with one provider focusing on sick patients and the other provider focusing on healthy patients. Physician specialization will tend to increase social welfare when provider market power (t) is small, when there are large numbers of each type of consumer ($\varsigma \approx \frac{1}{2}$), when patients are able to sort well (λ_1 and λ_2 are large), and when q_1 and q_2 are substitutes. Generally, if providers are specializing, different physicians should be assigned to different payment contracts (or multiple contracts should be offered, with physicians selecting different options).

1.5.2 Distortionary taxation

Up until this point, I have assumed that payments to providers are simply a transfer, and thus do not reduce social welfare. However, it is worth noting that when quality is poorly observed, payments to providers are much higher than the cost of providing care. In a world of uniform, risk-neutral providers, it would be possible to raise all funds above the true cost of care in a non-distortionary way by charging a flat entry fee (provider profits net of entry fees could be set to zero).[16]

A more interesting case involves deadweight loss associated with all taxation: here the planner wants to minimize the cost of achieving socially optimal values of q_1 and q_2 (and, accounting for deadweight loss, the optimal quality levels will be reduced). Consider first the case where the planner is free to choose positive values for p_b, p_1, and p_2. In this case there are more instruments than targets, and it is straightforward to see that the cost-minimizing contract will set p_b as low as possible, paying based on quality as much as possible.[17] This follows from the equation for the provider's marginal profit associated with changes in quality dimension k:

$$\frac{\partial \pi}{\partial q_k} = \frac{\partial \mu}{\partial q_k}\left(p_b + p_1 q_1 + p_2 q_2 - c\right) - \mu \cdot \frac{\partial c}{\partial q_k} + \mu p_k$$

For any given level of payment, the marginal profit is increasing more quickly in $p_k q_k$ than in p_b: the planner can induce higher quality at the same cost by increasing $p_k q_k$ and reducing p_b by equal amounts. Note that if p_b is unconstrained, the planner can induce optimal quality choices while setting provider profit arbitrarily close to zero, by sending

[16] In this light, it is curious that medical education is subsidized in most developed countries, rather than being taxed.

[17] Setting $\frac{1-p_1}{1-p_2} = \frac{\lambda_1}{\lambda_2}$ will achieve the right balance between q_1 and q_2. Note that this payment formula includes higher rewards for the less-observed dimension of quality, consistent with the results from the base model.

p_b towards negative infinity.

In section 1.4.6, we found that when the planner can choose values of p_b and p_1, it is optimal to set a positive value for p_1 if and only if q_2 is better observed than q_1 (that is, $\lambda_2 > \lambda_1$). When there is deadweight loss from taxation, $\lambda_2 > \lambda_1$ is sufficient but not necessary for the optimal value of p_1 to be positive. As the marginal cost of funds increases, and as λ_1 and λ_2 decrease, inducing quality through bundled payment becomes more costly relative to using quality-based payment. Consider the following extreme case: as λ_2 goes to zero, it becomes impossible to induce higher values of q_2 through bundled payment. In this case, the planner will be able to achieve any given level of q_1 more cheaply by using quality-based payment than by using bundled payment.

Another way the planner can reduce the cost of implementing a given quality level is by using non-linear quality-based payments. Consider a case where provider revenues are given by $\mu \cdot (p_b + f(q_1))$, where f is a function chosen by the planner. Where there is no uncertainty in the link between effort and measured quality, the planner's optimal f will include a step discontinuity at q_1^* (the optimal level of q_1). If we introduce uncertainty in measured quality (by adding normally-distributed noise, for example), it will still be optimal for f to provide the greatest marginal incentives in the neighborhood of q_1^*.

This discussion also has implications for public quality reporting, which presumably would increase the observability of the quality dimensions that are reported. So long as the payer adjusts payment contracts optimally, we should always increase any λ_k if it is costless to do so, because this will reduce the cost of implementing the optimal quality levels. However, consider a case where $\lambda_1 > \lambda_2$, and payment is bundled: increasing λ_1 would exacerbate the multitasking problem associated with differential observability of quality. The important point here is that (similar to the case of pay-for-performance) the effects of quality reporting will be dependent on the characteristics of payment contracts and the inefficiencies associated therewith.

1.6 Conclusion

My goal in this paper has been to address how we should think about pay-for-performance and the market failures it can mitigate, not to provide a recipe for implementation of a specific quality-based payment contract. As such, I have presented a simple model of competition between health care providers, sufficient to illustrate a basic point: quality-based payment should be used to address specific market failures in the existing payment system. Quality-based payment is one instrument of many that payers can use to come nearer to targets of provider quality.

In order to use quality-based payment effectively, designers of payment systems need to characterize the inefficiencies they aim to correct. We should expect that rewarding some dimensions of quality will result in providers focusing on those dimensions.[18] Thus, quality-based payment should focus on the dimensions of quality that are most underprovided in the existing payment system (as well as dimensions that are complements of other under-provided dimensions, or substitutes to over-provided dimensions of quality).

One major implication of this paper is that, in a payment system that already rewards providers based on consumer demand, we should not additionally reward aspects of quality to which there is relatively high demand response. Although for many quality dimensions, it is difficult to estimate demand response, we can state with some confidence that we should not reward patient satisfaction or patient experience measures. (If demand responds to *anything*, it ought to respond to patient satisfaction.) I even argue that if it is politically feasible, providers should be financially penalized for having better patient satisfaction scores, which is the opposite of current practice: both the QIP/IHI initiative in California and the QOF in the U.K. include payment for patient satisfaction and/or use of patient satisfaction surveys.

[18] Whether this results in reductions in other dimensions of quality depends on the nature of the providers' production functions, and whether different dimensions of quality are complements or substitutes.

The insights from my model can be applied beyond multitasking problems arising from failures of demand response to quality. In section 1.5.1, I argued that quality-based payment can be used to combat the problem of creaming and skimping—over-providing quality to profitable consumers and under-providing quality to unprofitable consumers. The QOF is built on a capitated payment system with very little risk adjustment, providing incentives for practices to attract healthy patients and avoid sick patients. All of the QOF clinical measures (about half of all payments) are based on care for patients with chronic disease, and payment is scaled by the number of patients with the disease in a practice's register. In this case, the QOF seems to be providing appropriate incentives to combat selection of healthy patients by practices.

The current paradigm of pay-for-performance neglects the interaction between quality-based payment and existing payment mechanisms. However, we cannot correctly implement supply-side incentives for quality without understanding the incentives that arise from demand-side responses within the existing payment system. When existing systems reward some aspects of health care quality more than others (which in my model arises from differential observability of quality dimensions), there are multitasking problems in the absence of quality-based payment. Future research should seek to more fully characterize the nature of the quality problem in health care markets, identifying specific aspects of quality to be targeted by pay-for-performance contracts. Pay-for-performance should then be used to address these specific inefficiencies, rather than as a blunt instrument to simply "improve quality".

2

Effect of payment reform on quality of primary care for coronary heart disease, diabetes, and chronic kidney disease in the United Kingdom

2.1 Introduction

In 2004, the National Health Service (NHS) implemented the Quality and Outcomes Framework (QOF), a new contract for General Practitioners (GP's) that added pay-for-performance (P4P) components to the existing capitation-based contract. Background on the history and details of the QOF are presented in section 2.2. Although there has been enthusiasm for P4P from both public and private payers (Institute of Medicine, 2006), the evidence on provider responses to quality-based payment is decidedly mixed (Rosenthal and Frank, 2006).

There is some evidence that quality of care has improved under the new QOF contract. Researchers from the National Primary Care Research and Development Centre (NPCRDC), at the University of Manchester, have published several papers investigating the effects of the QOF on health care quality. Doran et al. (2006) analyzed outcomes of the QOF after the first year, and found that the median practice earned 95.5% of the available points, as compared to the 75% that the NHS expected (and budgeted for). Since there were no reliable baseline data, it was unclear whether the benchmarks were set too low or whether quality of care rapidly improved after implementation of the QOF. In two closely-related papers (Campbell et al., 2009, 2007), the authors reviewed small random samples of medical records at each of 42 practices in England in 1998, 2003, 2005, and 2007. They found that the QOF was associated with a short-lived acceleration in quality improvement for asthma and diabetes, but that there was no such effect on care for coronary heart disease (CHD). They also analyzed whether there was a difference in trend between rewarded and unrewarded aspects of quality, finding that rewarded aspects improved more quickly.

Two studies using large databases extracted from EMR systems of GP practices have found conflicting results. Doran et al. (2011) found that the new contract was associated with significant improvement across a wide range of quality indicators, in a dataset including 148 English practices. Serumaga et al. (2011) used a different database from 358 practices across the U.K. to analyze quality of care for hypertensive patients. They precisely estimated zero effect of the new contract on blood pressure monitoring, control, treatment intensity, or health outcomes.

In 2008, a hospital P4P program was implemented in northwestern England rewarding facilities based on 28 quality measures in five clinical areas. In a well-identified analysis (Sutton et al., 2012) found that in-hospital mortality decreased significantly for patients admitted with rewarded conditions.

Outside the U.K., most studies tend to find small effects or no effect of P4P on quality of care, though many programs provide smaller incentives than those seen in U.K. P4P schemes (Rosenthal and Frank, 2006). For example, private payers in California implemented P4P for medical groups starting in 2002 and 2003, but annual quality-based payments per patient never rose above $18. Two papers on analyzing the California P4P experience did not find any major effects of P4P on either rewarded or unrewarded quality (Mullen et al., 2010; Rosenthal et al., 2005).

On the inpatient side, the Center for Medicare and Medicaid Services recently completed a six-year demonstration of P4P, in collaboration with Premier Inc. There is evidence of initial improvements in care processes (Lindenauer et al., 2007; Werner et al., 2011), but there seems to have been little or no effect on patient outcomes (Jha et al., 2012; Glickman et al., 2007).

In this study, we use the same dataset Serumaga et al. used to analyze hypertension treatment, but analyze a range of quality of care measures for CHD, diabetes, and CKD. Our work improves on the methods of Campbell et al. (2007) and (2009) by including a much larger sample in terms of both number of practices and number of patients, and by calculating performance monthly, as opposed to at four different time points across ten years. Our work also adds to the results from Doran et al. (2011) by analyzing intermediate outcomes quality measures, and not just process measures, which may be only weakly associated with patient outcomes (Ryan et al., 2009).

2.2 Background on the Quality and Outcomes Framework

Tony Blair's Labour government prioritized investment in the NHS, with total national health spending increasing from 6.6% of GDP in 1997 to 8.4% in 2007 (OECD, 2011). There were concerns about performance throughout the NHS, but measuring and improving the quality of primary care was a particular focus for the government (Department of Health, 2000). The first step in this direction was the Performance Assessment Framework (PAF), which assessed performance of Primary Care Trusts, but this measurement at the level of the local health authority was seen as largely ineffectual in improving health care quality (Department of Health, 1999; Chang, 2007). However, the PAF set the stage for increased focus on performance of primary care providers, and by 2004 the NHS implemented the QOF for GP's, likely the most ambitious P4P contract ever implemented in a health care setting.

By the early 2000's, it was generally accepted that the quality of primary care in the NHS left much to be desired, and that GP's in the system felt overworked and underpaid (Seddon et al., 2001; British Medical Association, 2001). One of the government's priorities from its 2000 NHS Plan was to increase the supply of GP's by 7.5% within four years (Department of Health, 2000; Doran and Roland, 2010). It was clear that the existing General Medical Services (GMS) contract was insufficient to encourage the level of GP entry and retention that would be necessary to meet the target of increased GP supply.[1]

After two years of negotiations between the NHS and British Medical Association (BMA), including an initial agreement that was rejected by doctors, a new contract was signed in June 2003. The new contract bundled payments at the practice level, as opposed to contracting with individual GP's, and provided for increased payments to GP practices beginning in fiscal year 2003 (FY03).[2] Practices were also given increased flexibility to define their hours and scope of services beginning in FY04.

As with the previous contract, the majority of payment was based on a capitated rate, with rudimentary case mix adjustment.[3] However, the new contract added the QOF, which had a budget of £1.8 billion ($3.3 billion) over three years, as compared to an overall annual primary care budget of £4.9 billion in FY02 (Doran et al., 2006; National Audit Office, 2008). When the QOF was implemented, practices could receive additional

[1]GP's are either paid under a national GMS contract, or locally-negotiated Personal Medical Services (PMS) contracts, which can vary considerably, but typically use the GMS as their framework.

[2]The U.K. fiscal year runs April through March.

[3]The capitation rate is adjusted based on patient age and sex, whether the patient is new to the practice, whether the patient is in a nursing or residential home, mortality and long-term limiting illness rates in the ward where the patient lives, wages in the practice's ward, and rurality of the practice.

payments for their performance on 146 quality indicators, divided into three sections: clinical care quality (including process and intermediate outcomes measures) for 10 chronic diseases, organization of care, and patient experience. Up to 1050 points could be awarded to a practice based on its performance on the indicators, and the practice received annual payment based on the number of points achieved (after adjusting for practice size and chronic disease prevalence among the practice's patients). For an average practice, the QOF payment was £76 per point in FY04, increasing to £125 per point in FY05.

Since 2004, the NHS has made minor changes to the QOF, but the overall structure has remained consistent. The largest changes came in 2006, when several new chronic diseases—including chronic kidney disease (CKD)—were added to the list of clinical areas in the QOF and the total number of points available was reduced to 1000. Also, due to higher performance on the measure than was originally expected, the lower thresholds for quality payment were increased (practices would generally begin receiving payments at 40% performance on the measures, instead of 25%). Some smaller changes to the list of quality indicators were implemented in 2009, along with minor adjustments to the formula that determines a practice's payment per QOF point.

The main stated purpose of the QOF was to improve quality of primary care in the NHS. However, regardless of the effects on health care quality, there were other important effects of the new contract. Spending on primary care increased from £4.9 billion in FY02 to £7.7 billion in FY05 (National Audit Office, 2008).[4] Between FY02 and FY05, net (pre-tax) income for non-salaried GP's increased by over 40% in real terms (The Health and Social Care Information Centre, 2009). Since then, real gross practice income has declined slowly, with real net income decreasing by 13.5% between FY05 and FY09 (The Health and Social Care Information Centre, 2011b). Although the effect of the new contract on GP morale is unclear, the total supply of GP's increased by over 4000 (15%) in the first four years of the contract (British Medical Association, 2007; Whalley et al., 2008;

[4]The government initially budgeted £6.9 billion for FY05; the overrun was largely caused by higher-than-expected performance on the QOF.

The Health and Social Care Information Centre, 2011a).[5] Finally, the QOF required practices to report performance through an electronic medical record (EMR) system and reimbursed all of the costs of EMR adoption. Although over 95% of practices used EMRs by 2000, many practices have upgraded their systems and interoperability of systems has increased markedly (Payne et al., 2011).

2.3 Methods

2.3.1 Data

We use The Health Improvement Network (THIN) data from Cegedim Strategic Data (CSD, http://csdmruk.cegedim.com/), which are comprehensive electronic medical records data from 514 UK GP practices, representing almost 6% of the UK population. The THIN data include year of birth, sex, postcode-level socioeconomic measures, medical encounters, diagnoses (Read codes), prescriptions, medical test values, referrals, and transfers into and out of the practice.

We use data from all 357 practices that began using the Vision practice management software by January 1, 2002, and had acceptable mortality reporting (as deemed by CSD) by March 1, 2003. Our analysis dataset starts in April 2000, and we begin including a practice in our analyses when the practice has been using the Vision software for at least 15 months and has acceptable mortality recording. We dropped 3866 practice-months from 206 separate practices whose data did not qualify for inclusion starting in 2000 (mostly due to adoption of Vision after January 1999).

In our analysis dataset, each observation represents a single patient in a single month, from April 2000 through March 2011 (FY00-FY10). For inclusion, a patient must be permanently registered with the practice for at least part of the month and must be at least

[5] Over 90% of the new GP's were salaried GP's, who tend to work fewer hours and be lower paid than practice partners. Salaried GP's represented under 3% of the GP workforce in FY02, and nearly 15% of the workforce in FY06.

18 years old.[6] We use Read codes (and, for CKD, creatinine values) to identify patients as having any of 27 diagnoses. See table 2.3.1 for descriptive statistics of our sample and information on what sources we used to identify Read codes indicating different diagnoses.

2.3.2 Quality scores

We use the QOF Business Rules (available at http://www.pcc-cic.org.uk/) to calculate monthly patient-level performance on the CHD, diabetes, and CKD quality indicators. We classify each indicator as structure, process, or intermediate outcome; we run separate analyses on processes, intermediate outcomes, and all indicators combined. The quality score for process and intermediate outcome measures is the percentage of relevant indicators met for the given patient. When running analyses on all indicators combined, we weight CHD and diabetes performance according to the QOF quality weights as of April 2004, and CKD performance according to the April 2006 weights. Each quality score is the weighted percentage of relevant indicators met for the given patient. See tables 2.3.2 through 2.3.4 for a list of quality indictors used in the analyses. We omit CHD and diabetes indicators 3 and 4, which are smoking-related indicators that were calculated differently starting in 2006.

2.3.3 Analysis

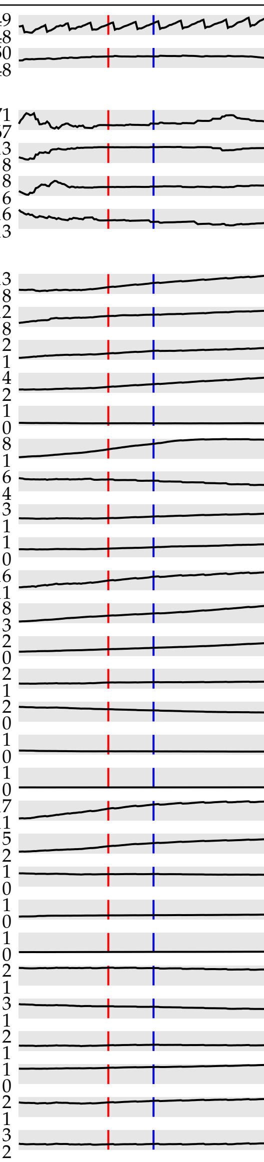
We present results from two separate sets of linear regression analyses, carried out at the patient-month level. In the first set of descriptive, non-parametric regressions, we estimate the following patient fixed-effects regression equation:

$$quality_{it} = \alpha + \gamma_t + diagnoses_{it} \cdot \beta + c_i + \varepsilon_{it} \tag{2.3.1}$$

[6] For human subjects protection, the THIN data only include year of birth, not exact date of birth. We include a patient starting in the month when her expected age is at least 18. Note that this results in an influx of 18-year-old patients in July of each year, resulting in the sawtooth pattern of average age in table 2.3.1

Table 2.3.1: Descriptive statistics at the patient-month level (*N*= 267,240,821 patient-months; 4,042,890 unique patients in 357 practices)

Variable		Mean or percent over time Apr '00 '04 '06 '11
Age: mean (SD)	48.5 (18.3)	49 48
Male	49.1%	50 48
Urban/Rural classification		
Urban	68.5%	71 67
Town & fringe	11.7%	13 8
Rural	6.9%	8 6
Missing	12.9%	16 13
Diagnosis present:		
Anxiety disorders	10.6%	13 8
Asthma	10.4%	12 8
Atrial fibrillation	1.4%	2 1
Cancer (any except non-melanotic skin cancer)	2.9%	4 2
metastatic cancer	0.1%	1 0
Chronic kidney disease	5.7%	8 1
Coronary heart disease	5.0%	6 4
Chronic obstructive pulmonary disease	1.8%	3 1
Dementia	0.5%	1 0
Depression	14.1%	16 11
Diabetes (any)	5.4%	8 3
diabetes with complications	0.8%	2 0
Epilepsy	1.3%	2 1
Heart failure	1.1%	2 0
Hemiplegia	0.2%	1 0
HIV/AIDS	0.01%	1 0
Hypertension	15.2%	17 11
Hypothyroidism	3.4%	5 2
Left ventricular dysfunction	0.6%	1 0
Liver disease, mild	0.2%	1 0
Liver disease, moderate to severe	0.04%	1 0
Myocardial infarction	1.9%	2 1
Peptic ulcer disease	2.1%	3 1
Peripheral vascular disease	1.3%	2 1
Psychosis	0.9%	1 0
Rheumatological disease	1.8%	2 1
Stroke or TIA	2.3%	3 2



Diagnoses in **bold** are the focus of this paper. Diagnoses in *italics* are coded based on QOF guidelines except CKD, for which we supplement the QOF diagnosis codes with eGFR (see Denburg et al., 2011). All other diagnoses are coded using the Khan et al. (2010) adaptation of the Charlson Index, except anxiety disorders and diabetes, which are coded based on CSD's list of relevant Read codes.

Table 2.3.2: List of QOF quality indicators used for CHD analyses

Indicator		Type	Points
CHD 1.	The practice can produce a register of patients with CHD	Structure (S)	6
CHD 2.	The percentage of patients with newly diagnosed angina who are referred for exercise testing and/or specialist assessment	Process (P)	7
CHD 5.	The percentage of patients with CHD who have a record of blood pressure in previous 15 months	P	7
CHD 6.	The percentage of patients with CHD, in whom the last blood pressure reading (in last 15 months) is 150/90 or less	Intermediate Outcome (IO)	19
CHD 7.	The percentage of patients with CHD who have a recorded total cholesterol in previous 15 months	P	7
CHD 8.	The percentage of patients with CHD whose last measured total cholesterol (in last 15 months) is 5 mmol/l or less	IO	16
CHD 9.	The percentage of patients with CHD with a record in the last 15 months that aspirin, an alternative anti-platelet therapy, or an anti-coagulant is being taken (unless contraindicated)	P	7
CHD 10.	The percentage of patients with CHD who are currently treated with a beta blocker (unless contraindicated)	P	7
CHD 11.	The percentage of patients with a history of MI (diagnosed after 1 April 2003) who are currently treated with an ACE inhibitor	P	7
CHD 12.	The percentage of patients with CHD who have a record of influenza vaccination in the preceding 1 September to 31 March	P	7

Table 2.3.3: List of QOF quality indicators used for diabetes analyses

Indicator		Type	Points
DM 1.	The practice can produce a register of all patients with diabetes mellitus	Structure (S)	6
DM 2.	The percentage of patients with diabetes whose notes record BMI in the previous 15 months	Process (P)	3
DM 5.	The percentage of diabetic patients who have a record of HbA1c or equivalent in the previous 15 months	P	3
DM 6.	The percentage of patients with diabates in whom the last HbA1c is 7.4 or less in last 15 months	Intermediate Outcome (IO)	16
DM 7.	The percentage of patients with diabates in whom the last HbA1c is 10 or less in last 15 months	IO	11
DM 8.	The percentage of patients with diabetes who have a record of retinal screening in the previous 15 months	P	5
DM 9.	The percentage of patients with diabetes with a record of presence or absence of peripheral pulses in the previous 15 months	P	3
DM 10.	The percentage of patients with diabetes with a record of neuropathy testing in the previous 15 months	P	3
DM 11.	The percentage of patients with diabetes who have a record of blood pressure in the previous 15 months	P	3
DM 12.	The percentage of patients with diabetes in whom the last blood pressure is 145/85 or less	IO	17
DM 13.	The percentage of patients with diabetes with a record of micro-albuminuria testing in previous 15 months (exception for patients with proteinuria)	P	3
DM 14.	The percentage of patients with diabetes who have a record of serum creatinine testing in previous 15 months	P	3
DM 15.	The percentage of patients with diabetes with proteinuria or micro-albuminuria who are treated with ACE inhibitors (or A2 antagonists)	P	3
DM 16.	The percentage of patients with diabetes who have a recorded total cholesterol in previous 15 months	P	3
DM 17.	The percentage of patients with diabetes whose last measured total cholesterol (in last 15 months) is 5 mmol/l or less	IO	6
DM 18.	The percentage of patients with diabetes who have a record of influenza vaccination in the preceding 1 September to 31 March	P	3

Table 2.3.4: List of QOF quality indicators used for CKD analyses

Indicator		Type	Points
CKD 1.	The practice can produce a register of patients with CKD	Structure (S)	6
CKD 2.	The percentage of patients on the CKD register whose notes have a record of blood pressure in the previous 15 months	Process (P)	6
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	Intermediate Outcome (IO)	11
CKD 4.	The percentage of patients on the CKD register with hypertension who are treated with an angiotensin converting enzyme inhibitor (ACE-I) or angiotensin receptor blocker (ARB), unless a contraindication or side effects are recorded	P	4

where t is the month (ranging from -48 for April 2000 to 83 for March 2011), the γ vector contains the coefficients of interest, and $diagnoses_{it}$ is a vector with one indicator for each diagnosis given in table 2.3.1.

Our estimates of the effect of the new GP contract on quality of care come from the following patient fixed-effects model:

$$\begin{aligned} quality_{it} = {} & \alpha + \beta_1 \cdot month_t + \beta_2 \cdot month_t^2 + \beta_3 \cdot pre\_post_t + \\ & \beta_4 \cdot pre\_post_t \cdot month_t + \beta_5 \cdot pre\_post_t \cdot month_t^2 + \\ & month\_of\_year_t \cdot \eta_1 + month\_of\_year_t \cdot pre\_post_t \cdot \eta_2 + \\ & diagnosis_{it} \cdot \eta + c_i + \varepsilon_{it} \end{aligned} \tag{2.3.2}$$

where $month_t$ is the month (ranging from -48 for April 2000 to 83 for March 2011) and $month\_of\_year_t$ is the calendar month, to control for seasonality. pre_post_t is a vector with an indicator for each possible policy regime: for CHD and diabetes analysis, the pre-policy period is FY00-FY02, FY03 is the phase-in period, and FY04-FY10 is the post-policy period. CKD was added to the QOF in April 2006, so for CKD analyses,

pre_post_t divides FY04-FY10 into three additional periods: April 2004 to November 2005 is the post-QOF, pre-CKD period; December 2005 to March 2006 is a phase-in period for CKD;[7] and FY06-FY10 is post-QOF, post-CKD. All reported standard errors are robust to heteroskedasticity and clustering at the practice level.

2.3.4 Defining what we mean by an effect of the QOF

If there is an effect of the QOF on quality of care, then that effect need not be restricted to a short time period after implementation of the QOF. Indeed, it is plausible that the QOF payment reforms cause long-run differences compared to what would have been the case in the absence of payment reform, and these quality differences may not be constant over time. Thus the ideal analysis would calculate the present discounted value of infinite stream of quality provided under the QOF, compared to the present discounted value of the stream of quality that would have occurred absent payment reform.

However, putting aside our inability to estimate these differences into the future, our analysis is hampered by lack of a strong source of identification of shorter-term effects. If the QOF had been implemented for some practices but not for others, or if implementation had been staggered across time—in short, if we had a control group for our analyses—we could empirically estimate the counterfactual with more confidence.[8] Without a control group, the best we can do is identify effects off of the extrapolation of trends from the pre-QOF period.

Economic theory is unable to provide much guidance for our modeling decisions in this regard. If the market is simply moving from one static equilibrium to another, any effect of the QOF will be a jump at the time of implementation, and the effect size will be constant over time. However, it is plausible that practices are in a dynamic equilibrium, characterized by secular changes in quality over time. Furthermore, payment reform might result in dynamic changes (for example, causing practices to innovate more

[7] CKD indicators were made public in late November 2005.

[8] Sadly, policies are often implemented without evaluation in mind.

quickly); in this case, the long-term effect of payment reform may be different from the effect at implementation. Our baseline analysis allows differing quadratic trends in each time period, but it is unclear *ex ante* whether to expect dynamic changes, and models using quadratic trends may be overly sensitive if there are changes in curvature.

For these reasons, we present results for a variety of alternative model specifications that allow for differing extrapolation of pre-QOF trends. We present results for the our estimated effect on quality in the first month as well as the effect averaged over different post-implementation time periods. (results in the Appendix).

2.4 Results

Results for CHD and diabetes quality are presented in figures 2.4.1 and 2.4.2, and in table 2.4.1. Payment reform was associated with meaningful improvement in rewarded dimensions of quality of care for CHD and diabetes. Overall performance on CHD measures increased by 9.9 percentage points ($SE = 0.69$), and diabetes performance increased by 8.3 percentage points ($SE = 0.98$). We find improvement in both process-based measures and intermediate outcomes, though for diabetes the process measures showed greater improvement than outcomes measures.

Results for CKD quality are presented in figure 2.4.3 and table 2.4.2. We find that overall CKD quality improved by 5.4 percentage points ($SE = 0.27$) in 2006 when CKD was added to the list of QOF clinical domains. Both process and outcomes measures improved significantly, but the improvement in process measures was relatively small.

CKD quality also improved in 2004, when there were incentives for quality of care for other clinical groups, but not for CKD patients. One potential complication with interpreting these positive effects is that many patients with CKD also have another condition that was rewarded under the QOF in 2004. However, separate analyses of CKD patients without any other relevant diagnoses also find significant quality improvement in 2004 (see table A.2.1 in the Appendix), pointing to the possibility of positive spillovers

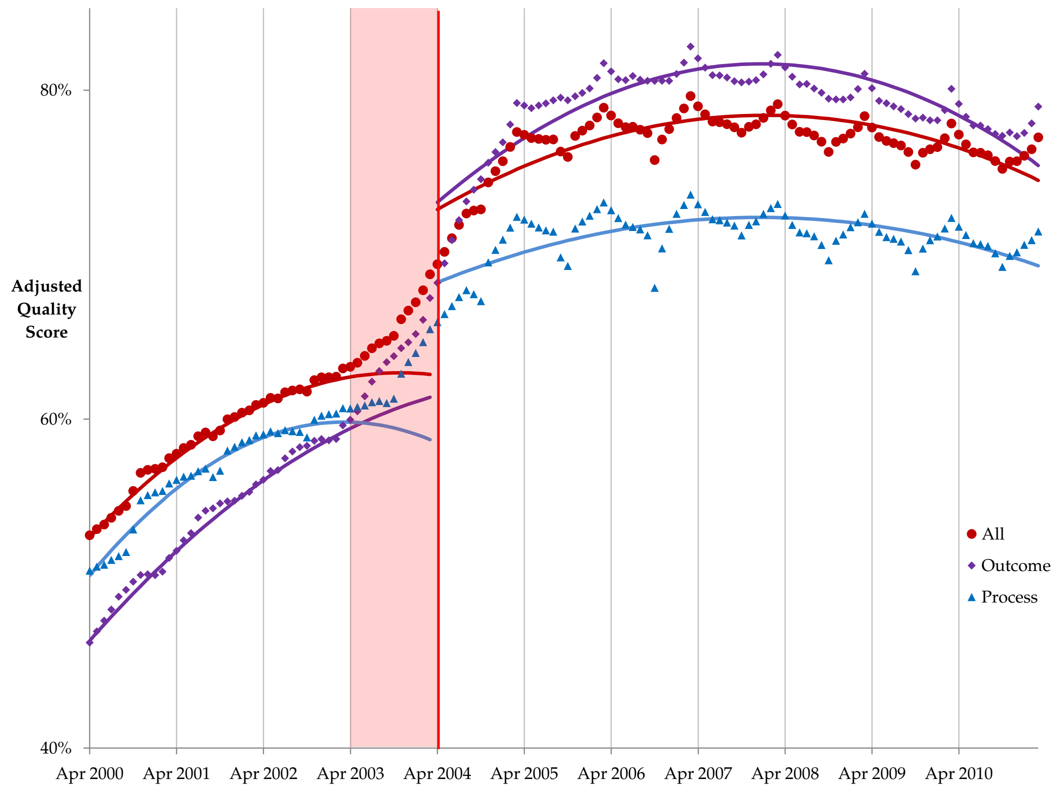


Figure 2.4.1: Mean CHD quality scores by month, adjusted for diagnoses and patient fixed effects. Data points are taken from month dummies in a non-parametric regression model. Trend lines are taken from a seasonality-adjusted parametric regression model allowing for different pre-QOF and post-QOF quadratic trends and intercepts. Estimates for the phase-in period (April 2003 to March 2004) are based on extrapolating quadratic trends from the pre-QOF period.

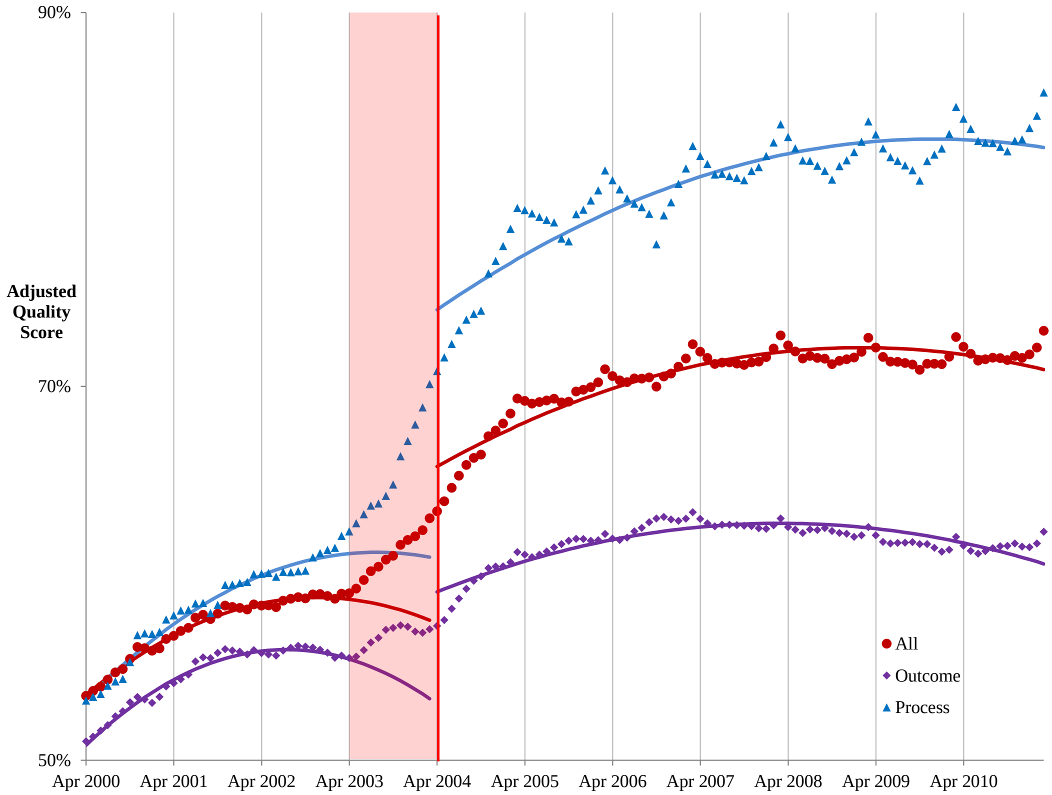


Figure 2.4.2: Mean diabetes quality scores by month, adjusted for diagnoses and patient fixed effects. Data points are taken from month dummies in a non-parametric regression model. Trend lines are taken from a seasonality-adjusted parametric regression model allowing for different pre-QOF and post-QOF quadratic trends and intercepts. Estimates for the phase-in period (April 2003 to March 2004) are based on extrapolating quadratic trends from the pre-QOF period.

Table 2.4.1: Regression results for CHD and diabetes quality

	CHD			Diabetes		
Variable	**All** β (SE)	**Outcome** β (SE)	**Process** β (SE)	**All** β (SE)	**Outcome** β (SE)	**Process** β (SE)
month	-0.0566	0.120	-0.195^{***}	-0.169^{*}	-0.282^{**}	-0.0825
	(0.0527)	(0.0806)	(0.0442)	(0.0729)	(0.103)	(0.0840)
month^2	-0.0055^{***}	-0.0040^{**}	-0.0076^{***}	-0.0052^{***}	-0.0068^{***}	-0.0050^{**}
	(0.0010)	(0.0015)	(0.0008)	(0.0013)	(0.0018)	(0.0015)
post	10.1^{***}	11.7^{***}	9.74^{***}	8.39^{***}	5.99^{***}	13.3^{***}
	(0.666)	(1.02)	(0.559)	(0.975)	(1.36)	(1.12)
post·month	0.309^{***}	0.257^{***}	0.372^{***}	0.388^{***}	0.438^{***}	0.351^{***}
	(0.0095)	(0.0131)	(0.0089)	(0.0150)	(0.0190)	(0.0158)
post·month^2	0.00267^{***}	-0.0002	0.0056^{***}	0.0033^{***}	0.0052^{***}	0.0030^{***}
	(0.00009)	(0.0001)	(0.00009)	(0.0001)	(0.0002)	(0.0002)
Seasonally-adjusted effect estimates across varying post-reform time periods						
1 month	9.91^{***}	10.9^{***}	9.96^{***}	8.32^{***}	5.79^{***}	13.4^{***}
	(0.685)	(1.06)	(0.572)	(0.986)	(1.34)	(1.17)
6 months	10.7^{***}	11.6^{***}	10.9^{***}	9.32^{***}	6.94^{***}	14.4^{***}
	(0.810)	(1.25)	(0.675)	(1.17)	(1.59)	(1.37)
1 year	11.7^{***}	12.3^{***}	12.2^{***}	10.6^{***}	8.42^{***}	15.5^{***}
	(0.985)	(1.52)	(0.821)	(1.41)	(1.94)	(1.65)
2 years	13.9^{***}	13.9^{***}	15.2^{***}	13.4^{***}	11.8^{***}	18.0^{***}
	(1.41)	(2.18)	(1.18)	(2.00)	(2.78)	(2.32)
3 years	16.4^{***}	15.3^{***}	18.8^{***}	16.5^{***}	15.6^{***}	20.8^{***}
	(1.94)	(2.97)	(1.62)	(2.72)	(3.79)	(3.14)

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Units are percentage points of quality scores. Standard errors are robust to heteroskedasticity and clustering at the practice level. April 2004 is coded as month = 0. All regressions include patient fixed-effects, month-of-year dummies (to control for pre-QOF seasonality), month-of-year dummies interacted with *post* (to control for post-QOF seasonality), and diagnosis indicators. Coefficients for trends and intercept in the phase-in period are suppressed.

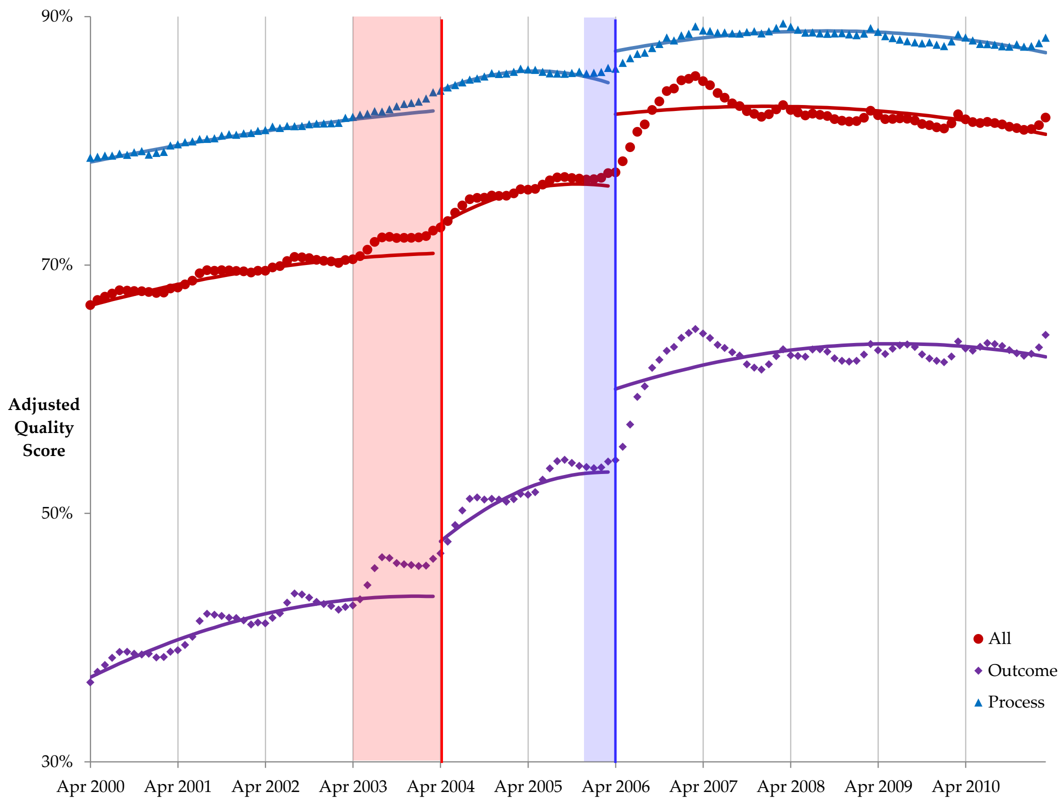


Figure 2.4.3: Mean CKD quality scores by month, adjusted for diagnoses and patient fixed effects. Data points are taken from month dummies in a non-parametric regression model. Trend lines are taken from a seasonality-adjusted parametric regression model allowing for different quadratic trends and intercepts in three time periods: FY00-FY03, FY04-FY05, FY06-FY10. Estimates for the two phase-in periods (April 2003 to March 2004, and December 2005 to March 2006) are based on extrapolating quadratic trends from the previous period.

Table 2.4.2: Regression results for CKD quality

Variable	All β (SE)	Outcome β (SE)	Process β (SE)
month	-0.0170	-0.0818	-0.0301
	(0.0490)	(0.0992)	(0.0840)
month^2	-0.0015	-0.0037^{*}	-0.0013
	(0.0009)	(0.0016)	(0.0011)
2004-2005	2.34^{***}	4.25^{***}	1.15
	(0.596)	(1.12)	(0.687)
2004-2005·month	0.327^{***}	0.450^{***}	0.119
	(0.0224)	(0.0986)	(0.113)
2004-2005·month^2	-0.0077^{***}	-0.0035	-0.0036
	(0.0010)	(0.0035)	(0.0040)
post_2006	8.82^{***}	7.37^{**}	-4.66^{*}
	(0.866)	(2.12)	(2.26)
post_2006·month	0.118^{***}	0.491^{***}	0.427^{***}
	(0.0216)	(0.0741)	(0.0928)
post_2006·month^2	-0.0001	-0.0006	-0.0037^{***}
	(0.0002)	(0.0007)	(0.0010)
Seasonally-adjusted effect estimates across varying post-reform time periods			
FY06 vs FY04			
1 month	5.44^{***}	6.19^{***}	2.66^{***}
	(0.274)	(0.442)	(0.199)
1 year	6.63^{***}	7.73^{***}	4.73^{***}
	(0.398)	(0.753)	(0.393)
2 years	8.62^{***}	10.1^{***}	7.67^{***}
	(0.676)	(1.33)	(0.729)
FY06 vs pre			
1 month	11.1^{***}	18.4^{***}	3.62
	(2.25)	(4.21)	(2.43)
1 year	11.7^{***}	20.5^{***}	4.08
	(2.77)	(5.18)	(3.00)
2 years	12.4^{***}	22.8^{***}	4.47
	(3.42)	(6.40)	(3.72)
FY04 vs pre			
1 month	2.19^{**}	4.29^{***}	1.40^{*}
	(0.609)	(1.16)	(0.636)
1 year	3.67^{***}	6.78^{***}	2.08^{*}
	(0.903)	(1.71)	(0.956)

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Units are percentage points of quality scores. Standard errors are robust to heteroskedasticity and clustering at the practice level. April 2004 is coded as month = 0. All regressions include patient fixed-effects, month-of-year dummies (to control for pre-QOF seasonality), month-of-year dummies interacted with *2004-2005* and *post_2006* (to control for differences in seasonality in each time period), and diagnosis indicators. Coefficients for trends and intercept in the phase-in periods are suppressed.

from other clinical groups to patients with CKD. It is possible that some practices may have anticipated the inclusion of CKD in the QOF; however, we are unaware of any public information pointing to the addition of CKD until late November 2005.

Across all regressions, we find that quality scores flatten out after initial improvements in the first two to three years after implementation. Also notable across clinical domains is change in seasonality of quality scores associated with payment reform.[9] Quality scores at the end of the fiscal year improved significantly more than quality scores in the middle of the fiscal year.[10] The most likely explanation for this phenomenon is that QOF payments are rewarded based on performance at the end of the fiscal year, so there is greater incentive to improve quality of care in March than at any other time. This is compounded by a 15-month look-back on many of the quality indicators: a test done between January and March can "count" for two years of QOF calculations.

Appendix tables A.2.2 through A.2.11 present results from alternative model specifications and from practice-level fixed effects models. We consistently find positive results across different models, though the size of the effects varies. We do not place much stock in the extrapolations more than the first two years past implementation of payment reform, since the estimates become increasingly sensitive to the assumed counterfactual.[11] With one exception,[12] when our baseline variable quadratic model finds positive effects averaged across any time period less than two years, all other models also find positive effects.

We find somewhat higher estimated effects using practice-level fixed-effects models than using patient fixed-effects models, particularly in the case of diabetes. This is consistent with the possibility of case mix changes, where diabetic patients diagnosed in

---

[9]Pre-QOF seasonality in outcomes scores are most likely explained by natural seasonal variation in blood pressure.

[10]Estimates of components of the η_2 vector in regression equation (2.3.2) are highly significant (results not shown).

[11]In some cases, extrapolation of the pre-reform quadratic trend far enough implies (impossible) negative quality scores by 2011.

[12]The exception is the practice-level fixed-effects model for CKD process quality (table A.2.11).

the post-reform period are healthier or more compliant than those who were diagnosed before payment reform.

2.5 Discussion

With the adoption of payment reform in 2004, primary care practices in the NHS improved rewarded process- and outcomes-based quality scores across several clinical domains. Furthermore, there were small improvements in care for CKD after payment reform and before quality of care for CKD patients was included in the QOF. While there are other possible ways practices could have diverted effort from unrewarded aspects of quality, this provides some evidence that multitasking may not have been a major problem with the QOF.

Our results are broadly consistent with those from Campbell et al. (2007) and (2009), though unlike the previous studies, we find significant improvements in CHD quality. Previous studies have looked at quality trends through 2007 and noted that quality leveled off after improving in the first year of payment reform.

This points to the importance of not interpreting evaluations of NHS payment reform as reflecting the effect of ”pay-for-performance”. The overall payment reforms also included an initial huge increase in average payments per patient; it could be the case that practices responded to an increase in payment, rather than to P4P. The clean experiment of the effectiveness of P4P, per se, would be to assign some practices to the QOF and others to comparably increased capitation levels. In at least one case where this approach was taken, there were quality improvements in the P4P group relative to the group that received higher non-quality-based payment (Gertler and Vermeersch, 2013).

Another major result of payment reform was investment in of EMRs by practices. Since all practices in our sample had robust EMR systems prior to payment reform, we are unable to measure the degree to which this EMR adoption resulted in quality improvements. However, this is arguably a strength of our paper, since we are able to disentangle QOF-related effects from effects due to technology adoption.

2.6 Conclusion

We found initial improvements in quality of care with payment reform in the NHS, but quality of care was fairly constant after the initial improvement. Further research is warranted to understand the degree to which initial improvements in quality are attributable to P4P, rather than other aspects of the 2004 payment reforms. Researchers should also explore whether different, possibly dynamic, incentives could have encouraged further improvement after the first couple years of implementation.

3

Specificity of healthcare providers' responses to pay-for-performance incentives in the United Kingdom

3.1 Introduction

In 2004, the National Health Service (NHS) in the United Kingdom (UK) implemented the Quality and Outcomes Framework (QOF) for primary care practices, perhaps the most ambitious pay-for-performance (P4P) scheme ever tried in a health care setting. The QOF builds on the previous capitation payment system by increasing financial rewards for practices that provide higher quality care, as measured by a range of structural, process-based, and outcomes-based indicators.

As originally implemented, practices received additional payment for their performance on 146 quality indicators, divided into three sections: clinical care quality for

10 chronic diseases, organization of care, and patient experience. Up to 1050 points could be awarded to a practice based on its performance on the indicators, and the practice received annual payment based on the number of points achieved (after adjusting for practice size and chronic disease prevalence among the practice's patients). Since 2004, there have been minor changes to the QOF, including payment on clinical care quality for several additional chronic diseases, a reduction in total points available to 1000, and minor changes to the formulas that translate performance into QOF points and calculate payment per QOF point.

There is evidence that payment reform was associated with substantial improvements in quality of care for a range of conditions (Campbell et al., 2009; Doran et al., 2011; Richardson, 2013). It is unclear, however, whether improvements in quality of care were due to increased average payments per patient, public reporting of quality data, or the specific quality-based payment incentives in the QOF. To begin to attribute effects to the P4P aspects of the QOF, it is necessary to look more precisely at which practices responded to incentives, and how they responded. Furthermore, the QOF is also expensive: spending on primary care increased from £4.9 billion in fiscal year 2002 (FY02) to £7.7 billion in FY05,[1] and the NHS now spends approximately £1 billion per year on the QOF (National Audit Office, 2008; Cashin, 2011). We do not know how efficiently the QOF translates this increased spending into improved quality and whether the incentive structure in the QOF could be altered to achieve similar quality levels with lower spending.

There is mixed evidence on how well providers base their decisions on the specific marginal financial incentives they face. For example, in the Quality Improvement Program in California described in Rosenthal et al. (2005), the medical groups who were the worst performers at baseline improved the most despite having almost no financial incentive to do so, having started so far below the targets for receiving payment. If providers fail to respond strategically to the incentives they face under P4P programs, there are important implications for the design of such programs. On the one hand, if

[1]The U.K. fiscal year runs April through March.

providers engage in "schmeduling"—responding heuristically to imperfectly perceived incentives—we will be unable to use the traditional principal-agent optimization framework to predict the effects of any given set of incentives (Liebman and Zeckhauser, 2004). On the other hand, such dulled responses to specific incentives might imply that the precise details of P4P programs are not terribly important.

3.2 Specific incentives in the QOF

Broadly speaking, the QOF financially rewards higher performance on a wide range of quality indicators. However, at a more granular level, the marginal incentives facing practices at different levels of baseline performance are quite different. For each quality indicator k, payment is linearly increasing between a floor ($\underline{q}$) and a ceiling ($\overline{q}_k$). Figure 3.2.1 shows how points are calculated for each QOF indicator, where each indicator k is associated with a maximum number of points ϕ_k, and points translate linearly into payment in pounds. Practices at the ceiling do not receive additional payment for further improvement, but there are strong marginal incentives for those operating between $\underline{q}$ and $\overline{q}_k$.

One type of behavior where there are markedly different incentives facing practices above and below the ceilings is exception reporting, which is reporting that patients are inappropriate for certain performance measures. When a practice claims an exception for a patient on a given indicator, the patient is removed from the denominator for calculation of performance on that indicator. A practice operating above the ceiling gains nothing when it reports an exception (and in fact reduces payment, since the size of the denominator enters the formula calculating the payment per point). However, a practice operating between the floor and the ceiling increases its payment by claiming an exception on a patient for whom the practice is failing to satisfy the indicator. Gravelle et al. (2008) found that practices above and below payment ceilings did in fact behave differently when reporting exceptions, with those below the threshold inflating their exception reports by

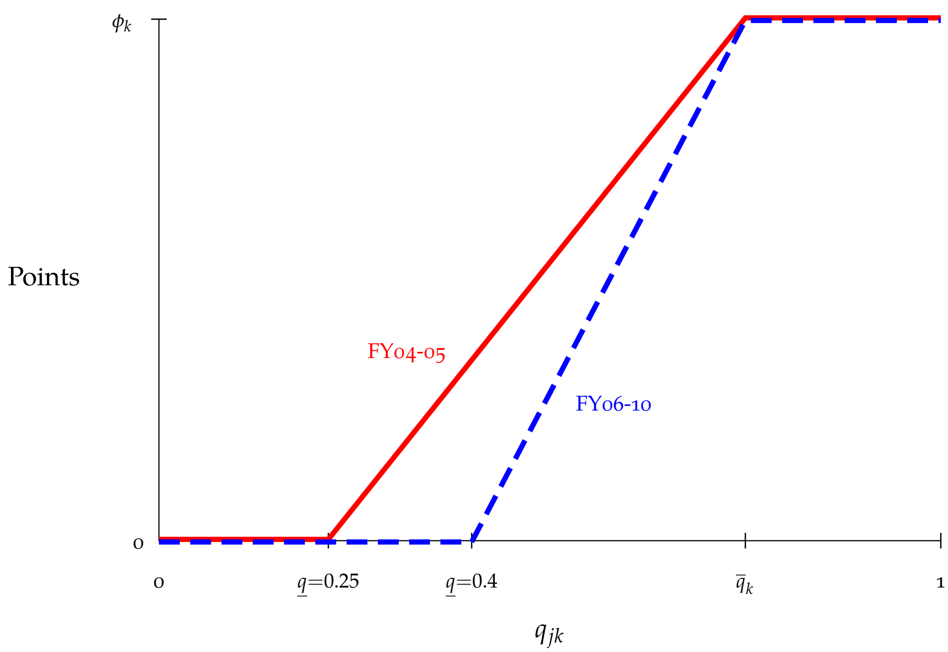


Figure 3.2.1: Calculation of QOF points achieved on each indicator.

some 18%.

Beyond the discontinuity created by payment ceilings, many of the indicators themselves create discontinuous incentives. The intermediate outcomes indicators reward a practice for the percentage of relevant patients whose last test value in a fiscal year is below a threshold. See table 3.2.1 for a list of the CHD and diabetes intermediate outcomes measures included during the first 7 years of the QOF. Collectively, these indicators represent between 8% and 9% of all QOF points available, translating to roughly £11,000 per year for an average practice. The distributions of practice performance on these indicators are presented in appendix figures A.3.1 through A.3.10.

Consider the cholesterol indicators (CHD 8 and DM 17), which reward practices based on the percentage of patients whose last total cholesterol measurement in the fiscal year is below 5 mmol/l (193 mg/dl). A practice could potentially increase its payment by aggressively retesting patients whose numbers are just above 5 mmol/l, but not patients whose numbers are just below 5 mmol/l, before the end of the fiscal year. Furthermore, only practices operating between the floor and ceiling have an incentive to engage in this type of behavior. It is unclear whether practices will respond to the discrete 5 mmol/l threshold, and if they do, whether it is only the practices that are between the payment floor and ceiling.

Finally, the QOF rewards performance as measured at the end of the fiscal year (in March). Richardson (2013) presents evidence that the QOF introduced clear seasonality in quality scores, with the highest scores consistently coming in March. Again, if practices are responding to the true marginal incentives they are facing, we should only observe this type of seasonality in practices operating below the performance ceiling.

In this paper, we investigate the degree to which practices respond to the specifics of QOF incentives introduced by payment ceilings, test value thresholds, and end-of-fiscal-year measurement.

Table 3.2.1: Intermediate outcomes quality measures for CHD and diabetes

Indicator		**Years**	**Ceiling***	**Max points***
CHD 6.	The percentage of patients with CHD, in whom the last blood pressure reading (in last 15 months) is 150/90 or less	All	70	19 | 19 | 17
CHD 8.	The percentage of patients with CHD whose last measured total cholesterol (in last 15 months) is 5 mmol/l or less	All	60 | 70 | 70	16 | 17 | 17
DM 23.	The percentage of patients with diabates in whom the last HbA1c is 7 or less in last 15 months	FY09-10	50	17
DM 6.	The percentage of patients with diabates in whom the last HbA1c is 7.4 or less in last 15 months	FY04-05	50	16
DM 20.	The percentage of patients with diabates in whom the last HbA1c is 7.5 or less in last 15 months	FY06-08	50	17
DM 24.	The percentage of patients with diabates in whom the last HbA1c is 8 or less in last 15 months	FY09-10	70	8
DM 25.	The percentage of patients with diabates in whom the last HbA1c is 9 or less in last 15 months	FY09-10	90	10
DM 7.	The percentage of patients with diabates in whom the last HbA1c is 10 or less in last 15 months	FY04-08	85 | 90 | NA	11
DM 12.	The percentage of patients with diabetes in whom the last blood pressure (in last 15 months) is 145/85 or less	All	55 | 60 | 60	17 | 18 | 18
DM 17.	The percentage of patients with diabetes whose last measured total cholesterol (in last 15 months) is 5 mmol/l or less	All	60 | 70 | 70	6

* Several of the ceilings and maximum points were changed in FY06 or in FY09. In these cases, the first number refers to FY04-05, the second number refers to FY06-08, and the third number refers to FY09-10.

3.3 Theoretical framework

The following model is our framework for considering how we expect GP practices to respond to the change from largely capitated payment to the new contract's mix of capitation and quality-based payment. We assume that there are J providers in monopolistic competition for patients who have varying health care conditions (denoted by K-dimensional vector ρ, where element k of vector ρ is one if the patient has condition k and zero otherwise). Each provider chooses K-dimensional quality vector q provided to all patients, and each patient chooses a provider, resulting in increasing and weakly concave demand for each provider $\mu(q)$.[2]

Provider j's cost of providing care of quality q_j to consumer i is given by $c_i\left(q_j\right) = c\left(q_j\right) \cdot \sum_k \omega_k \rho_{ik}$, where c is a strictly convex cost function and ω_k is a cost weight given to condition k.[3] We define the prevalence of condition k among patients of practice j as

$$P_{jk} \equiv \frac{\sum_{i=1}^{\mu_j} \rho_{ik}}{\mu_j}$$

where μ_j is demand for provider j, in number of patients. A practice's average cost per patient is then $c\left(q_j\right) = c\left(q_j\right) \cdot \sum_k \omega_k P_{jk}$.

3.3.1 Provider quality choices and profits before payment reform

We abstract from the details of case-mix adjustment and relatively small fee-based payments, and model payment as a simple bundled (capitated) rate per patient p_b. Profit[4]

[2] We do not explicitly model patient choices, but we assume a positive demand response to quality.

[3] The additivity of conditions' weights is a simplifying assumption that will have little effect in the empirical section. The salient point is that patients with any given condition cost more than those without the condition, *ceteris paribus*.

[4] We will assume profit-maximization by providers, but the model is robust to at least some forms of provider altruism. We can think of c as the cost function net of the provider's altruism-related benefit per patient. If the provider's altruism-related benefit per patient is a weakly concave function of q, then c will remain convex, and all our conclusions will continue to hold.

to a constant-returns-to-scale provider j before payment reform is

$$\pi_j\left(q_j\right) = \mu\left(q_j\right) \cdot \left[p_b - c\left(q_j\right) \cdot \sum_k \omega_k P_{jk}\right]. \tag{3.3.1}$$

Assuming an interior solution, the K first-order conditions for a maximum are $\forall k$:

$$\begin{aligned} \frac{\partial \pi_j}{\partial q_{jk}} &= \frac{\partial \mu_j}{\partial q_{jk}} \cdot \left[p_b - c\left(q_j\right) \cdot \sum_l \omega_l P_{jl}\right] - \\ &\quad \mu_j \cdot \left[\frac{\partial c}{\partial q_{jk}} \cdot \sum_l \omega_l P_{jl} + c\left(q_j\right) \cdot \sum_l \omega_l \frac{\partial P_{jl}}{\partial q_{jk}}\right] = 0 \end{aligned} \tag{3.3.2}$$

We will assume that condition prevalences are independent of provider quality choices (*i.e.* $P_{jk} \perp q_j,\ \forall k$). That is, the size of the practice can depend on q, but the mix of patients in the practice does not. In this case, the first-order conditions simplify to the following marginal-revenue-equals-marginal-cost expression, $\forall k$:

$$\frac{\partial \mu_j}{\partial q_{jk}} \cdot p_b = \frac{\partial \mu_j}{\partial q_{jk}} \cdot c\left(q_j\right) \cdot \sum_l \omega_l P_{jl} + \mu_j \cdot \frac{\partial c}{\partial q_{jk}} \tag{3.3.3}$$

Note the implication that $\frac{\partial q_{jk}}{\partial p_b} > 0$: even in the absence of P4P, there will be positive quality choices, and we should expect increased payment per patient to result in increased quality. Furthermore (and again assuming $P_{jk} \perp q_j\ \forall k$), providers serving patient populations with higher condition prevalences should provide lower quality care.

3.3.2 Provider quality choices and profits after payment reform

The 2004 payment reforms included payments for provider performance on 146 quality indicators, with 65 of the indicators rewarding processes or intermediate outcomes for patients with particular conditions. Each indicator k is associated with a maximum number of points (ϕ_k),[5] and we use q_{jk} to denote the percentage of practice j's relevant

[5]Unrewarded dimensions of quality can be included in the model with $\phi_k = 0$.

patients for which the quality indicator was met. Each indicator has a floor ($\underline{q}$) below which the practice receives no points and a ceiling ($\overline{q}_k$) above which the practice receives no additional points beyond ϕ_k.[6] Points are awarded linearly between the floor and the ceiling, so the points awarded on indicator k are given by

$$g\left(\frac{q_{jk}-\underline{q}}{\overline{q}_k-\underline{q}}\right)$$

where we define $g(x) = \max\left(0, \min\left(1, x\right)\right)\ \forall x \in \mathbb{R}$. Each point is worth a fixed amount v for the average-sized practice, and the payment for practice j is scaled linearly with the practice size, μ_j. Condition-specific indicators are further scaled by the square root of the practice's prevalence of the condition, $\sqrt{P_{jk}}$, relative to the nationwide average square root of the prevalence, $\overline{\sqrt{P_k}} \equiv J^{-1} \sum_j \sqrt{\frac{\sum_{i=1}^{\mu_j} \rho_{ijk}}{\mu_j}}$.

Thus, after payment reform, provider j's profit is

$$\pi_j\left(q_j\right) = \mu\left(q_j\right) \cdot \left[\left[p_b - c\left(q_j\right) \cdot \sum_k \omega_k P_{jk}\right] + \frac{v}{\overline{\mu}} \cdot \sum_k \phi_k \cdot g\left(\frac{q_{jk}-\underline{q}}{\overline{q}_k-\underline{q}}\right) \cdot \frac{\sqrt{P_{jk}}}{\overline{\sqrt{P_k}}}\right] \tag{3.3.4}$$

where $\overline{\mu}$ is the size of the average practice, so $\frac{v}{\overline{\mu}}$ represents the per-patient payment per QOF point.

The K first-order conditions for an interior maximum are $\forall k$:

[6] For all indicators, the floor in FY04-FY05 was 25%, and the floor was increased to 40% beginning in FY06. The ceilings for indicators range from 50% to 90%.

$$
\begin{aligned}
\frac{\partial \pi_j}{\partial q_{jk}} = & \frac{\partial \mu_j}{\partial q_{jk}} \cdot \left[p_b - c\left(q\right) \cdot \sum_l \omega_l P_{jl} \right] - \qquad (3.3.5)\\
& \mu_j \cdot \left[\frac{\partial c}{\partial q_{jk}} \cdot \sum_l \omega_l P_{jl} + c\left(q_j\right) \cdot \sum_l \omega_l \frac{\partial P_{jl}}{\partial q_{jk}} \right] + \\
& \frac{\partial \mu_j}{\partial q_{jk}} \cdot \frac{v}{\overline{\mu}} \cdot \sum_l \phi_l \cdot g\left(\frac{q_{jl} - \underline{q}}{\overline{q}_l - \underline{q}} \right) \cdot \frac{\sqrt{P_{jl}}}{\overline{\sqrt{P_l}}} + \\
& \mu_j \frac{v}{\overline{\mu}} \cdot \left[\phi_k \cdot \frac{\mathbb{1}_{\left[\underline{q} < q_{jk} < \overline{q}_k\right]}}{\overline{q}_k - \underline{q}} \cdot \frac{\sqrt{P_{jk}}}{\overline{\sqrt{P_k}}} + \sum_l \phi_l \cdot g\left(\frac{q_{jl} - \underline{q}}{\overline{q}_l - \underline{q}} \right) \cdot \frac{\frac{1}{2 \cdot \sqrt{P_{jl}}} \cdot \frac{\partial P_{jl}}{\partial q_{jk}}}{\overline{\sqrt{P_l}}} \right] = 0
\end{aligned}
$$

As in section 3.3.1, we consider the special case where $P_{jk} \perp q_j \; \forall k$, reducing the first-order conditions to the following marginal-revenue-equals-marginal-cost expression, $\forall k$:

$$
\begin{aligned}
& \frac{\partial \mu_j}{\partial q_{jk}} \cdot \left[p_b + \frac{v}{\overline{\mu}} \cdot \sum_l \phi_l \cdot g\left(\frac{q_{jl} - \underline{q}}{\overline{q}_l - \underline{q}} \right) \cdot \frac{\sqrt{P_{jl}}}{\overline{\sqrt{P_l}}} \right] + \qquad (3.3.6)\\
& \mu_j \cdot \frac{v}{\overline{\mu}} \cdot \phi_k \cdot \frac{\mathbb{1}_{\left[\underline{q} < q_{jk} < \overline{q}_k\right]}}{\overline{q}_k - \underline{q}} \cdot \frac{\sqrt{P_{jk}}}{\overline{\sqrt{P_k}}} \\
= & \frac{\partial \mu_j}{\partial q_{jk}} \cdot c\left(q_j\right) \cdot \sum_l \omega_l P_{jl} + \mu_j \cdot \left[\frac{\partial c}{\partial q_{jk}} \cdot \sum_l \omega_l P_{jl} \right]
\end{aligned}
$$

Note that marginal revenues are discontinuous at $\underline{q}$ and $\overline{q_k}$, and there will be solutions where the first-order conditions do not hold with equality. Figure 3.3.1 depicts marginal revenue as a function of quality on a single rewarded indicator, both before and after payment reform. After payment reform, providers with marginal cost (MC) less than MC_1 will choose $q_j > \overline{q}_k$. Providers with $\mathrm{MC} \in [\mathrm{MC}_1, \mathrm{MC}_2]$ will choose $q_j = \overline{q}_k$. Providers with $\mathrm{MC} \in (\mathrm{MC}_2, \mathrm{MC}_3]$ will choose $q_j \in \left(\underline{q},\ \overline{q}_k\right)$. Finally, providers with $\mathrm{MC} > \mathrm{MC}_3$ will choose $q_j < \underline{q}$.

If the QOF had provided incentives for only a single dimension of quality, the

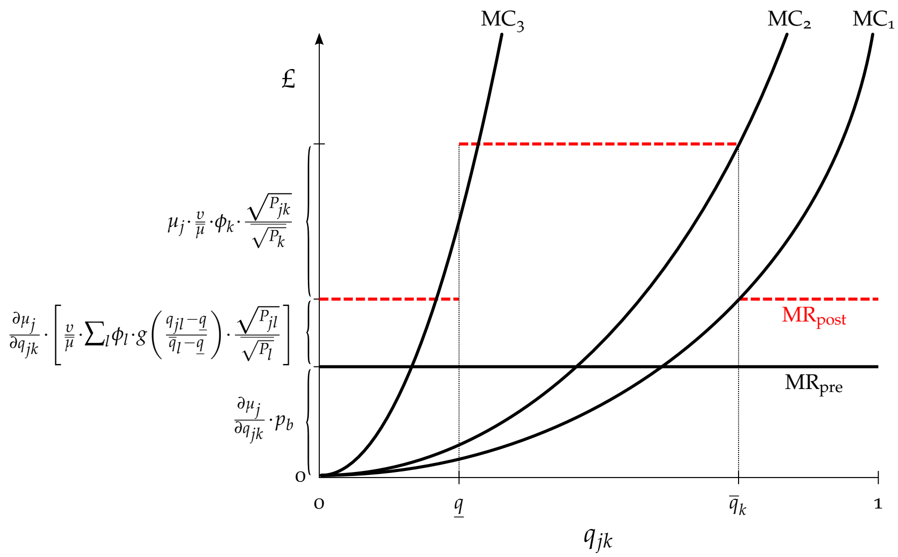


Figure 3.3.1: Marginal revenue (MR) for a single dimension of rewarded quality, before and after QOF, and three possible marginal cost curves.

theoretical predictions would follow easily from the first-order conditions and from figure 3.3.1. We would expect all practices to improve, with practices having MC $\in$ (MC$_2$, MC$_3$) improving more than others (at least in percentage terms). However, Sherry (2012) has shown that when a P4P program rewards multiple dimensions of quality, the expected direction of change is unknown even for rewarded dimensions. This comes from the possibility that different dimensions of quality may be either substitutes or complements; investments in one dimension of quality may change the marginal costs of other dimensions.

For this reason, we focus not only on comparing practice behavior before and after QOF, but also on comparing behavior of practices above and below ceilings for each indicator ($\overline{q}_k$).[7] This gets around the problem noted by Sherry (2012): given a continuous cost function, there is a discontinuity in marginal incentives at the ceiling. Under the

[7] In theory, we would also want to compare practices above and below payment floors, $\underline{q}$, since these practices would also face different marginal incentives. However, it is very rare for practices to perform below $\underline{q}$.

assumption that when averaged over different practices, marginal incentives for a behavior vary continuously except for the discontinuity at $\overline{q}_k$, we can apply a regression discontinuity design to estimate whether practices are responding to the marginal incentives implied by ceilings. This assumption amounts to practices having cost functions such that their marginal cost for condition k is not determinative of their marginal cost for other conditions, and would follow (for example) from a setup where the P_j vector varies stochastically across practices.

As discussed in section 3.1, the structure of marginal incentives around the payment ceilings implies that profit-maximizing providers should behave differently depending on whether they are operating above or below the ceilings. In particular, this paper will investigate whether practices vary their patterns of exception reporting, medical testing, and end-of-fiscal-year behavior based on the specific marginal incentives they face.

3.4 Methods

We use The Health Improvement Network (THIN) data from Cegedim Strategic Data (CSD, http://csdmruk.cegedim.com/), which are comprehensive electronic medical records data from 514 UK GP practices, representing almost 6% of the UK population. The THIN data include year of birth, sex, postcode-level socioeconomic measures, medical encounters, diagnoses (Read codes), prescriptions, medical test values, referrals, and transfers into and out of the practice. These data have been used previously by Serumaga et al. (2011) to analyze effects of the QOF on treatment of hypertension.

We use data from all 357 practices that began using the Vision practice management software by January 1, 2002, and had acceptable mortality reporting (as deemed by CSD) by March 1, 2003. Our analysis dataset starts in April 2000, and we begin including a practice in our analyses when the practice has been using the Vision software for at least 15 months and has acceptable mortality recording. We dropped 3866 practice-months from 206 separate practices whose data did not qualify for inclusion starting in 2000 (mostly due to adoption of Vision after January 1999).

Calculating practice performance. Using version 8.0 (March 15, 2006) of the "New GMS Contract QOF Implementation Dataset and Business Rules", published by the Department of Health, we calculate each practice's performance on all CHD and diabetes quality indicators, except those related to smoking cessation[8]. For each indicator in each month, we determine whether each practice is above or below the payment ceiling.[9] Finally, we create overall measures of marginal incentives for CHD and diabetes by calculating the percent of points in the disease domain for which the practice is above the payment ceiling, denoted $ceil_{CHD}$ and $ceil_{dia}$, respectively.[10]

Results are presented in the following two sections, which use different analysis datasets extracted from the THIN data. Section 3.5 analyzes exception reporting by practices, and uses data at the practice-month level, looking only at the post-QOF period (since there were no exception reports before QOF). Section 3.6 analyzes the probability that a practice will retest a patient's cholesterol, as a function of the baseline cholesterol value. For these analyses we construct a dataset in which each observation is a single cholesterol test taken within two years before or after QOF implementation (FY02-FY05).

3.5 Exception reporting

Performance on each indicator is calculated as the number of patients who meet the indicator (numerator) divided by the number of patients with the relevant condition (denominator). Each practice may claim exceptions, excluding patients from the denominator for various reasons: if the patient fails to visit the practice after several attempts to schedule the patient, if the patient is on maximal drug therapy, or if the patient is otherwise inappropriate for the indicator. However, if a practice claims an

[8] Calculation of denominators for the smoking cessation indicators fundamentally changed in 2006.

[9] Note that some of the payment ceilings change in 2006 or 2009.

[10] A practice above most of the ceilings will have a high value for these variables, and will face low marginal incentives to improve quality, relative to a practice below the ceilings.

exception for a patient, the patient will only be excluded from indicators that the patient fails to meet. That is, if an excepted patient meets a given indicator, she will still be included both in the numerator and denominator for that indicator.

For a practice below the threshold, reporting an exception weakly increases payment: either the patient is meeting the indicator, so the exception does not apply; or the patient fails to meet the indicator, so the exception increases the practice's performance on the indicator (decreasing the denominator by 1). For a practice above the ceiling, claiming an exception weakly reduces payment, because payment per point is increasing in the reported prevalence P_{jk}. Again, if the patient is meeting the indicator, the exception does not make a difference. But if the patient fails to meet the indicator, then the practice's denominator size is reduced, and the improvement in performance does not increase payment, since the marginal payment for performance increases is zero for a practice above the ceiling.

We analyze patterns of exception reporting using data at the practice-month level, subset to post-QOF data (starting April 2004). We estimate the following practice-fixed-effects regression model separately for CHD and for diabetes:

$$Pr\_X_{jkt} = \alpha + \beta \cdot ceil_{jk,t-1} + \gamma_j + \delta_{FY_t} + \rho_{month_t} + \varepsilon_{jkt} \tag{3.5.1}$$

where Pr_X_{jkt} is the proportion of exceptions claimed among patients with condition k (where k is CHD or diabetes) in practice j at time t, where t is denominated in months. δ_{FY_t} and ρ_{month_t} are fixed effects vectors for fiscal years and months of the year, respectively. We expect β to be negative, since a practice should claim fewer exceptions when its lagged performance places it above more ceilings.

Results for CHD and diabetes are presented in table 3.5.1. We find no effect of ceilings on exception reporting for CHD, but find that practices performing above more ceilings for the diabetes indicators are less likely to claim diabetes exceptions. Increasing by one standard deviation (21 percentage points) the percent of diabetes-related points for which

Table 3.5.1: Regression results for CHD and diabetes exception reporting associated with the practice's performance relative to payment ceilings

	CHD		Diabetes	
	β	SE	β	SE
$ceil_{jk,t-1}$	0.000412	(0.001030)	-0.00158^{*}	(0.000754)
FY05	0.000254	(0.000262)	0.00118^{***}	(0.000340)
FY06	-0.0000911	(0.000249)	0.00157^{***}	(0.000421)
FY07	-0.000717^{**}	(0.000236)	0.000889^{*}	(0.000399)
FY08	-0.000682^{*}	(0.000276)	0.000311	(0.000372)
FY09	-0.00114^{***}	(0.000299)	0.000539	(0.000407)
FY10	-0.000947^{**}	(0.000309)	0.000238	(0.000396)
May	0.000699^{**}	(0.000264)	0.000328	(0.000333)
Jun	0.000579^{**}	(0.000212)	0.000280	(0.000260)
Jul	0.000869^{***}	(0.000258)	0.000415	(0.000278)
Aug	0.000658^{*}	(0.000257)	0.000193	(0.000247)
Sep	0.000814^{***}	(0.000238)	0.00102^{***}	(0.000296)
Oct	0.00135^{***}	(0.000248)	0.00143^{***}	(0.000300)
Nov	0.00180^{***}	(0.000265)	0.00221^{***}	(0.000321)
Dec	0.00176^{***}	(0.000314)	0.00201^{***}	(0.000358)
Jan	0.00399^{***}	(0.000389)	0.00548^{***}	(0.000360)
Feb	0.00442^{***}	(0.000361)	0.00823^{***}	(0.000556)
Mar	0.00555^{***}	(0.000496)	0.0130^{***}	(0.000762)
constant	0.00662^{***}	(0.000746)	0.00860^{***}	(0.000603)
Pr_X_{jkt} mean	0.00434		0.00571	

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

$N = 28,233$ practice-months

The dependent variable is the percentage of a practice's patients with the given diagnosis for whom the practice claims an exception during a given month. FY04 and April are the omitted categories. Practices are weighted by the number of patients in the practice with the relevant diagnosis, averaged over the post-QOF period. Standard errors are robust to heteroskedasticity and clustering at the practice level, and regressions include practice fixed effects.

the practice is above ceilings decreases the number of patients for whom the practice claims an exception by 5.8%.[11]

Also of note is the strong seasonality in exception reporting. For both CHD and diabetes, practices report many more exceptions in the final months of each fiscal year (January through March) than they do earlier in the year. This is consistent with practices waiting until near the end of the fiscal year to decide whether it is advantageous to claim exceptions. Table 3.5.2 includes similar regressions in which we interact the $ceil_{jk,t-1}$ variable with months of the year. We find that the seasonal pattern of exception reporting is less evident when practices are above payment ceilings than when they are below ceilings. This suggests that practices use excepetion reporting strategically near the end of the year to improve reported performance on those measures where they are performing below ceilings.

To further analyze the degree to which practices were strategically using exception reporting, we compared each practice's payments at the end of each fiscal year to what the practice would have been paid had the practice not claimed any exceptions.[12] We found that on average, exception reporting increased practices' CHD payments by 0.4% and their diabetes payments by 2.0%. However, over 56% of the time, a practice's CHD QOF payments would have been *higher* had the practice reported no CHD exceptions. Over 10% of practices, through claiming exceptions, lost more than 2.5% of the CHD income they would have received had they claimed no exceptions. Over 31% of the time, a practice's diabetes payments would have been higher had the practice reported no diabetes exceptions.

Although there have been concerns that some practices overused exception reports to boost payments (Gravelle et al., 2008), to our knowledge there was never concern that

[11] This is calculated as $^{-0.00158 \cdot 0.21}/_{0.00571} = 0.058$.

[12] The data are not sufficient to calculate actual payments, but we can accurately determine the percent difference between the payment actually received and the payment that would have been received had the practice not claimed any exceptions. Furthermore, some exceptions apply across several chronic conditions, so we were unable to calculate the full effects of the exception reporting.

Table 3.5.2: Regression results for CHD and diabetes exception reporting associated with the practice's performance relative to payment ceilings

	CHD exceptions		Diabetes exceptions	
	β	SE	β	SE
$ceil_{jk,t-1}$	−0.00164	(0.00143)	−0.00200	(0.00171)
FY05	0.000215	(0.000263)	0.00102**	(0.000352)
FY06	−0.000121	(0.000249)	0.00144***	(0.000427)
FY07	−0.000766**	(0.000238)	0.000724	(0.000409)
FY08	−0.000731**	(0.000276)	0.000120	(0.000380)
FY09	−0.00120***	(0.000300)	0.000320	(0.000418)
FY10	−0.001000**	(0.000310)	−0.0000176	(0.000413)
May	−0.00163	(0.000849)	−0.00206	(0.00133)
Jun	−0.00133	(0.000712)	−0.00168	(0.00132)
Jul	−0.000953	(0.000793)	−0.00115	(0.00136)
Aug	−0.000591	(0.000773)	−0.00217	(0.00128)
Sep	−0.000954	(0.000733)	−0.000447	(0.00154)
Oct	−0.00105	(0.000865)	−0.000695	(0.00145)
Nov	−0.000317	(0.000805)	0.000898	(0.00142)
Dec	−0.000774	(0.000764)	0.00117	(0.00150)
Jan	0.00132	(0.00133)	0.00444**	(0.00157)
Feb	0.00545***	(0.00143)	0.00905***	(0.00194)
Mar	0.00894***	(0.00257)	0.0274***	(0.00351)
May·$ceil_{jk,t-1}$	0.00321*	(0.00130)	0.00370*	(0.00175)
Jun·$ceil_{jk,t-1}$	0.00258*	(0.00108)	0.00309	(0.00175)
Jul·$ceil_{jk,t-1}$	0.00246*	(0.00112)	0.00246	(0.00179)
Aug·$ceil_{jk,t-1}$	0.00160	(0.00117)	0.00377*	(0.00168)
Sep·$ceil_{jk,t-1}$	0.00238*	(0.00110)	0.00228	(0.00200)
Oct·$ceil_{jk,t-1}$	0.00335**	(0.00127)	0.00335	(0.00190)
Nov·$ceil_{jk,t-1}$	0.00292*	(0.00125)	0.00203	(0.00190)
Dec·$ceil_{jk,t-1}$	0.00352**	(0.00110)	0.00124	(0.00212)
Jan·$ceil_{jk,t-1}$	0.00367*	(0.00176)	0.00153	(0.00206)
Feb·$ceil_{jk,t-1}$	−0.00157	(0.00199)	−0.00135	(0.00258)
Mar·$ceil_{jk,t-1}$	−0.00468	(0.00322)	−0.0215***	(0.00463)
constant	0.00422***	(0.00103)	0.00361*	(0.00142)
Pr_X_{jkt} mean	0.00434		0.00571	

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

$N = 28,233$ practice-months

The dependent variable is the percentage of a practice's patients with the given diagnosis for whom the practice claims an exception during a given month. FY04 and April are the omitted categories. Practices are weighted by the number of patients in the practice with the relevant diagnosis, averaged over the post-QOF period. Standard errors are robust to heteroskedasticity and clustering at the practice level, and regressions include practice fixed effects.

practices would inflate their payment by failing to claim exceptions. For this reason it would seem that the expected costs of underreporting exceptions would be low. Hence, our analysis suggests that practices in general were not overly sophisticated in their strategic use of exception reporting.

A related question concerns the budgetary impact to the government of allowing exception reporting. Note that as practices across the country report exceptions, they reduce the average reported prevalence of the relevant conditions ($\overline{\sqrt{P_k}}$ from section 3.3.2), thereby slightly increasing payment for all other practices. Therefore, a close approximation to the budgetary impact of allowing exception reporting is simply the increase in average points attained due to practices reporting exceptions.[13] We find that allowing exception reporting increased practices' mean points attained from 87.5% to 90.2% on CHD indicators, and from 92.5% to 95.6% on diabetes indicators. At least in these two disease domains, allowing exception reporting only had a small budgetary impact.

3.6 Retesting near test score thresholds for payment

The quality indicators in table 3.2.1 create discrete cutoffs in cholesterol, blood pressure, and HbA1c values: above the cutoff the practice gets zero points for the patient, and below the cutoff the practice gets full points. Furthermore, only a patient's most recent test result in a fiscal year matters. Given that cholesterol, blood pressure, and HbA1c values are somewhat noisy, retesting a patient below the cutoff could result in reduced quality scores, and retesting a patient above the cutoff could result in increased quality scores, even in the absence of any change in the patient's underlying health.

In this section, we focus on the cholesterol indicators (CHD 8 and DM 17) and analyze whether the cutoffs introduced a discontinuity in a patient's likelihood of having her cholesterol retested before the end of the fiscal year. We expect patients with scores just

[13] This approximation will be slightly biased downward if practices performing worse are more likely to report exceptions.

below the cutoff to be less likely to be retested than patients with scores just above the cutoff. Analyses in this section are at the patient-month level, and we subset to data from the two years before and after implementation (FY02-05). All standard errors are robust to clustering at the practice level, and to heteroskedasticity (where appropriate).

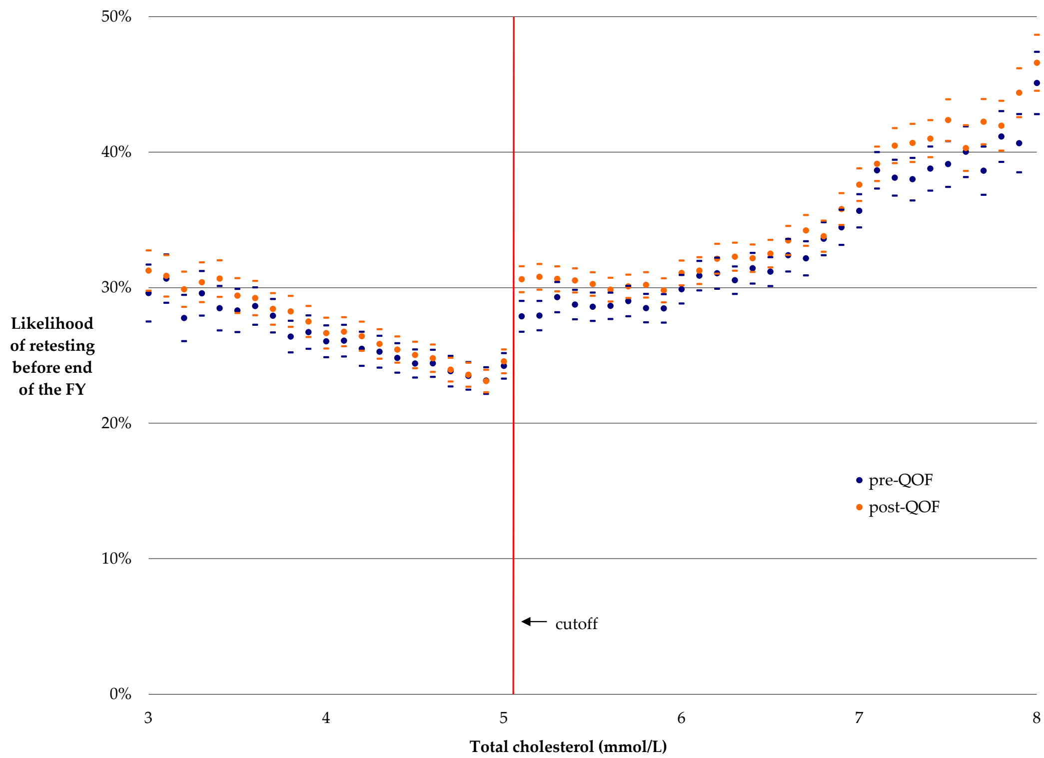


Figure 3.6.1: Unadjusted mean percentage of patients being retested before the end of the fiscal year, by cholesterol value. Tests in the last month of the fiscal year are dropped, and 95% confidence intervals are robust to clustering at the practice level. $N = 1,852,621$ cholesterol tests.

Figure 3.6.1 presents the unadjusted probability of a patient with a given cholesterol value to be retested before the end of the fiscal year, before and after QOF. There is a discontinuity at the cutoff value of 5 mmol/L both in the pre and post period: patients with a cholesterol value of 5.1 mmol/L are substantially more likely than those with a value of 5 mmol/L to be retested before the end of the fiscal year. This discontinuity is larger in the post-QOF period, and the difference between the pre and post probabilities is greatest just above the cutoff. This is consistent with what we would expect from the incentives practices face: the patients with cholesterol values just above the cutoff are most likely to drop below in the next test.

It is certainly possible that patient characteristics other than cholesterol values affect the probability of the patient being retested. For this reason, we analyzed the same data using a patient fixed-effects model, and controlling for the month of the year. Results presented in figure 3.6.2 show similar findings, with the discontinuity at the cutoff is larger in the post period than in the pre period. Note however that the fixed effects matter: probability of retesting in this model is non-decreasing in the initial cholesterol value, and the difference between the pre and post period does not attenuate as cholesterol increases further above the cutoff.

To quantify the change in the discontinuity, we estimate the following patient fixed-effects regression equation:

$$\begin{aligned} \Pr\left[retest_{it}\right] &= \alpha + \beta_1 \cdot postQOF_t + \beta_2 \cdot \mathbb{1}_{[chol_{it}>5]} + \\ &\quad \beta_3 \cdot postQOF_t \cdot \mathbb{1}_{[chol_{it}>5]} + \gamma_i + \rho_{month_t} + \varepsilon_{it} \end{aligned} \tag{3.6.1}$$

Results are presented in table 3.6.1. In the pre-QOF period, a patient with a cholesterol value over 5 mmol/L was 18.6 percentage points more likely to be retested during the same fiscal year, compared to the same patient with a cholesterol value below 5 mmol/L. After the QOF, this difference increased by 5.8 percentage points to 24.4 percentage points.

The QOF had a strong overall effect on patients' probabilities of having their cholesterol retested, but there were only incentives for cholesterol testing among patients with CHD, diabetes, or history of stroke or transient ischemic attack (TIA). Next, we analyze separately patients with and without these diagnoses. Figure 3.6.3 plots the unadjusted likelihood of retesting, comparing between patients with and without the relevant diagnoses before and after QOF. The figure shows that patients with the diagnoses were

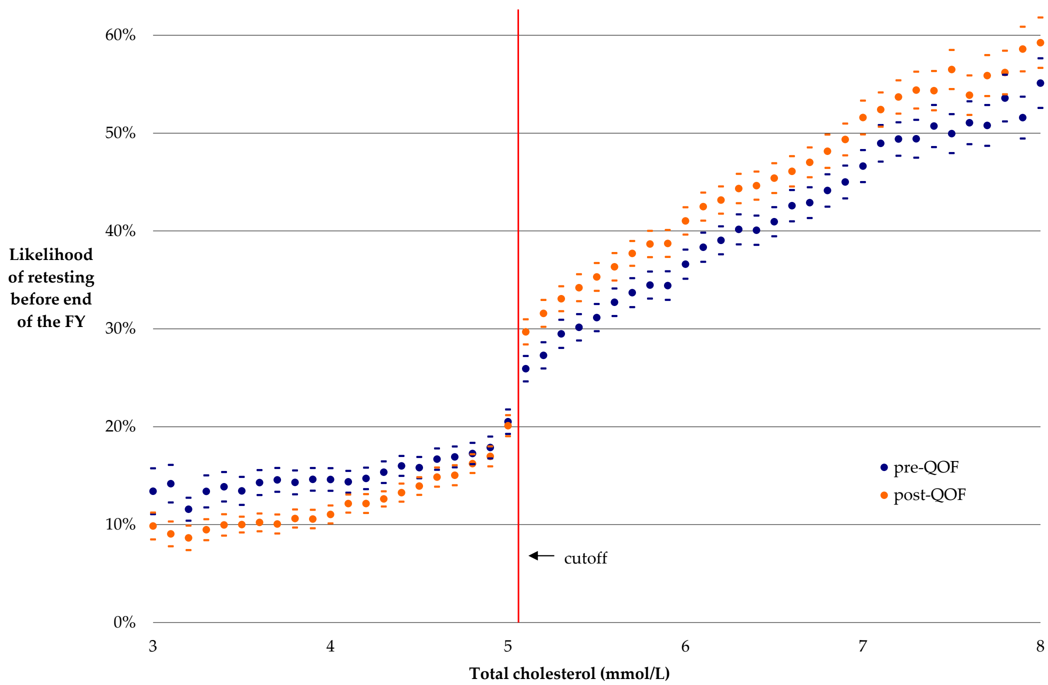


Figure 3.6.2: Results from a linear probability model regression predicting the probability of cholesterol retesting before the end of the fiscal year. We control for patient fixed-effects and month-of-year indicators, and drop tests in the last month of the fiscal year. 95% confidence intervals are robust to heteroskedasticity and clustering at the practice level. $N = 1,852,621$ cholesterol tests from $755,240$ unique patients.

Table 3.6.1: Regression results for cholesterol retesting

	Equation 3.6.1		Equation 3.6.2	
	β	SE	β	SE
$\beta_1 \cdot postQOF$	-0.00668^{*}	(0.00308)	0.00848^{**}	(0.00258)
$\beta_2 \cdot chol > 5$	0.186^{***}	(0.00415)	0.113^{***}	(0.00260)
$\beta_3 \cdot postQOF \cdot chol > 5$	0.0578^{***}	(0.00316)	0.00313	(0.00287)
$\beta_4 \cdot Dx$			0.190^{***}	(0.00494)
$\beta_5 \cdot postQOF \cdot Dx$			-0.00556	(0.00359)
$\beta_6 \cdot chol > 5 \cdot Dx$			0.0154^{**}	(0.00483)
$\beta_7 \cdot postQOF \cdot chol > 5 \cdot Dx$			0.0382^{***}	(0.00439)

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

$N = 1,858,232$ cholesterol tests from $757,412$ unique patients

The dependent variable in each regression is the likelihood of the patient having his or her cholesterol retested before the end of the fiscal year. A constant and indicator variables for months of the year are included in the model and their coefficients are suppressed. Standard errors are robust to heteroskedasticity and clustering at the practice level. Equation 3.6.1 includes patient fixed effects, and equation 3.6.2 accounts for patient-level random effects.

always more likely to be retested, but the change in retesting patterns is specific to the rewarded diagnoses.[14]

To quantify the change in the discontinuity for patients with and without the diagnoses, we estimate the following patient-level random-effects regression equation:

$$
\begin{aligned}
\Pr\left[retest_{it}\right] \quad = \quad & \alpha + \beta_1 \cdot postQOF_t + \beta_2 \cdot \mathbb{1}_{[chol_{it}>5]} + \\
& \beta_3 \cdot postQOF_t \cdot \mathbb{1}_{[chol_{it}>5]} + \beta_4 \cdot Dx_{it} + \beta_5 \cdot postQOF_t \cdot Dx_{it} + \\
& \beta_6 \cdot \mathbb{1}_{[chol_{it}>5]} \cdot Dx_{it} + \beta_7 \cdot postQOF_t \cdot \mathbb{1}_{[chol_{it}>5]} \cdot Dx_{it} + \\
& \rho_{month_t} + \varepsilon_{it}
\end{aligned} \tag{3.6.2}
$$

Results are presented in the rightmost two columns in table 3.6.1. Coefficient β_3 represents the change from pre-QOF to post-QOF in the difference in probability of retesting above and below the 5mmol/L cutoff. The precisely estimated zero implies that

[14] The presence of the diagnoses does not change enough to precisely estimate a patient fixed effects model. A random effects model controlling for the month of the test produces nearly identical results to the unadjusted means.

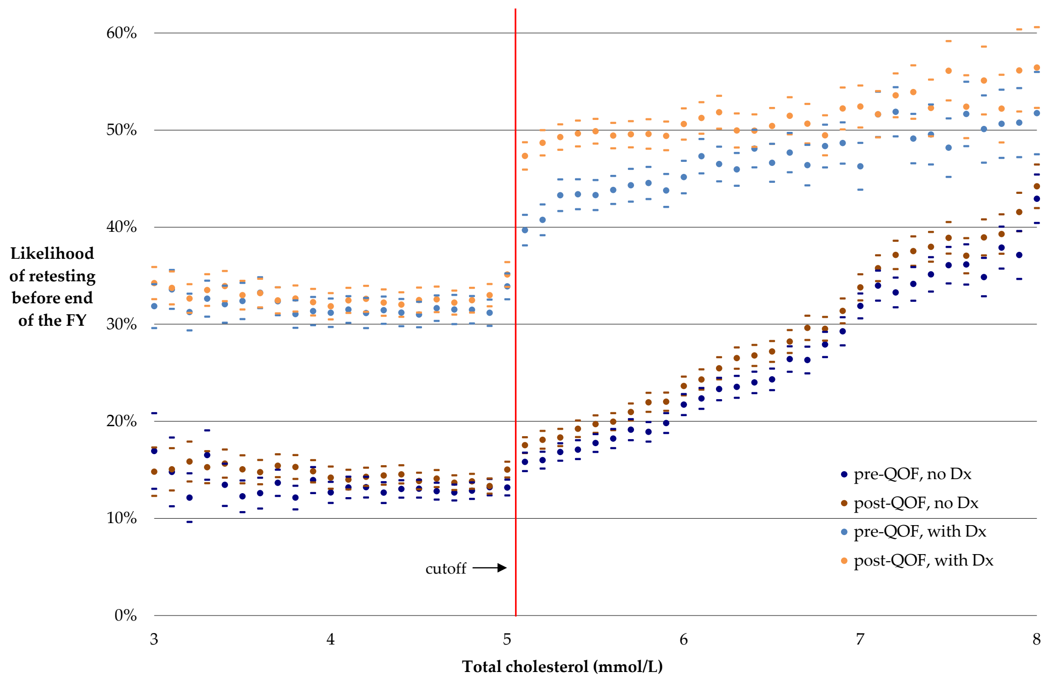


Figure 3.6.3: Unadjusted mean percentage of patients being retested before the end of the fiscal year, by cholesterol value and presence or absence of the following diagnoses: CHD, diabetes, and stroke/TIA. Tests in the last month of the fiscal year are dropped, and 95% confidence intervals are robust to clustering at the practice level. $N = 1,852,621$ cholesterol tests.

the change observed from the pooled analysis—regression equation 3.6.1—only occurred among those with the rewarded diagnoses. Our estimate for the change among those with the relevant diagnoses is $\beta_3 + \beta_7 = 0.0414$ $(SE = 0.00432)$, implying a 4.1 percentage point pre-QOF to post-QOF difference in difference of probabilities of retesting above and below 5 mmol/L.

Furthermore, as with exception reporting, we might expect practices above and below payment ceilings to behave differently when it comes to retesting patients. For this analysis, we subset to patients with CHD and/or diabetes, but without history of stroke or TIA (since we did not extract quality data on stroke measures). For patients with both CHD and diabetes, we code the practice as being below the payment ceiling if below the ceiling for *either* the CHD or diabetes measure (since the practice has marginal incentives to retest that are similar to practices below the ceiling for a patient's only diagnosis, among patients with either CHD or diabetes, but not both). Appendix figure A.3.11 presents results of a patient fixed-effects regression comparing the probability of cholesterol retesting before and after QOF, among practices above or below the relevant payment ceilings given the patient's diagnosis.[15]

Related regression results are presented in appendix table A.3.1, in which we estimate the following patient-level fixed-effects regression equation:

$$\begin{aligned}
\Pr\left[retest_{it}\right] \quad = \quad & \alpha + \beta_1 \cdot postQOF_t + \beta_2 \cdot \mathbb{1}_{[chol_{it} > 5]} + \\
& \beta_3 \cdot postQOF_t \cdot \mathbb{1}_{[chol_{it} > 5]} + \beta_4 \cdot over_ceiling_{it} + \\
& \beta_5 \cdot postQOF_t \cdot over_ceiling_{it} + \beta_6 \cdot \mathbb{1}_{[chol_{it} > 5]} \cdot over_ceiling_{it} + \\
& \beta_7 \cdot postQOF_t \cdot \mathbb{1}_{[chol_{it} > 5]} \cdot over_ceiling_{it} + \rho_{month_t} + \varepsilon_{it}
\end{aligned} \tag{3.6.3}$$

Coefficient β_7 is the relevant triple-difference estimate: it shows the effect of QOF on

[15]Note that before the QOF, there was no defined performance ceiling or target performance on the measures, so we define ceilings as if the practices were facing the FY04 ceilings. We should expect no discrete effect of a practice's performance relative to ceilings before QOF.

the difference between practices above and below ceilings in the jump in retesting at a total cholesterol of 5 mmol/L. We would expect the negative coefficient that we find: practices above the ceiling have less incentive to distort their cholesterol testing behavior around the 5 mmol/L threshold. However, the result is largely driven by pre-QOF differences between practices above and below (theoretical) ceilings.

3.7 Conclusion

In this chapter, we have shown that practices responded significantly to various marginal incentives introduced by the QOF. Practices altered their behavior based on whether they were above or below payment ceilings, based on the month of the fiscal year, and based on a patient's cholesterol value relative to a discrete threshold. Research showing that the QOF was associated with overall improvements in quality cannot attribute those improvements to P4P or public reporting of results, per se. Another plausible explanation is that the QOF drastically increased overall payments to practices, and it was this wealth effect that caused improved performance. In this study we found effects that would be hard to explain as being caused by anything other than the P4P or public reporting aspects of the QOF.

However, practice responses were also not purely sophisticated responses to the financial incentives they faced. For example, practices commonly reported exceptions when the exception actually decreased the practice's payment. Practices also increased retesting of patients with cholesterol values above the 5 mmol/L threshold, but did not decrease retesting of patients with values below the threshold. Furthermore, the increased retesting was not concentrated among patients with values slightly above the threshold, but remained among patients far above the threshold (who were unlikely, upon retesting, to drop below the threshold).

This paper suggests that payers implementing P4P contracts should be aware of the specific incentives they are creating, and consider the possibility of unintended consequences. However, at least in the context of primary care practices in the NHS, some unintended consequences that might have been predicted by economic theory did not materialize.

References

British Medical Association. National survey of GP opinion, London, 2001.

British Medical Association. National survey of GP opinion, London, 2007.

Stephen Campbell, David Reeves, Evangelos Kontopantelis, Elizabeth Middleton, Bonnie Sibbald, and Martin Roland. Quality of primary care in England with the introduction of pay for performance. *The New England Journal of Medicine*, 357(2):181–190, 2007.

Stephen M. Campbell, David Reeves, Evangelos Kontopantelis, Bonnie Sibbald, and Martin Roland. Effects of pay for performance on the quality of primary care in England. *The New England Journal of Medicine*, 361(4):368–378, 2009.

Cheryl Cashin. United Kingdom: Quality and Outcomes Framework (QOF), The World Bank, March 2011. URL `http://www.rbfhealth.org/rbfhealth/system/files/Case%20study%20UK%20QOF.pdf`.

Li-cheng Chang. The NHS performance assessment framework as a balanced scorecard approach: Limitations and implications. *International Journal of Public Sector Management*, 20(2):101–117, 2007.

Michelle R. Denburg, Kevin Haynes, Justine Shults, James D. Lewis, and Mary B. Leonard. Validation of The Health Improvement Network (THIN) database for epidemiologic studies of chronic kidney disease. *Pharmacoepidemiology and Drug Safety*, 20(11): 1138–1149, November 2011.

Department of Health. The NHS performance assessment framework, March 1999. URL `http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_4009190`.

Department of Health. The NHS plan: a plan for investment, a plan for reform, July 2000. URL `http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_4002960`.

Tim Doran and Martin Roland. Lessons from major initiatives to improve primary care in the United Kingdom. *Health Affairs*, 29(5):1023–1029, May 2010.

Tim Doran, Catherine Fullwood, Hugh Gravelle, David Reeves, Evangelos Kontopantelis, Urara Hiroeh, and Martin Roland. Pay-for-performance programs in family practices in the United Kingdom. *The New England Journal of Medicine*, 355(4):375–384, July 2006.

Tim Doran, Evangelos Kontopantelis, Jose M. Valderas, Stephen Campbell, Martin Roland, Chris Salisbury, and David Reeves. Effect of financial incentives on incentivised and non-incentivised clinical activities: longitudinal analysis of data from the UK Quality and Outcomes Framework. *BMJ*, 342, June 2011.

Karen Eggleston. Multitasking and mixed systems for provider payment. *Journal of Health Economics*, 24(1):211–23, 2005.

Randall P. Ellis. Creaming, skimping and dumping: provider competition on the intensive and extensive margins. *Journal of Health Economics*, 17(5):537–555, October 1998.

Paul Gertler and Christel Vermeersch. Using performance incentives to improve medical care productivity and health outcomes. Working Paper 19046, National Bureau of Economic Research, May 2013. URL `http://www.nber.org/papers/w19046`.

Jacob Glazer and Thomas G. McGuire. Optimal risk adjustments in markets with adverse selection: An application to managed care. *American Economic Review*, 90(4):1055–1071, 2000.

Seth W. Glickman, Fang-Shu Ou, Elizabeth R. DeLong, Matthew T. Roe, Barbara L. Lytle, Jyutsna Mulgund, John S. Rumsfeld, W. Brian Gibler, E. Magnus Ohman, Kevin A. Schulman, and Eric D. Peterson. Pay for performance, quality of care, and outcomes in acute myocardial infarction. *JAMA*, 297(21):2373–80, 2007.

Brian R. Golden and Frank A. A. Sloan. Physician pay for performance. In Frank A. Sloan and Hirschel Kasper, editors, *Incentives and Choice in Health Care*, pages 289–317. MIT Press, Cambridge, MA, 2008.

Hugh Gravelle, Matt Sutton, and Ada Ma. *Doctor behaviour under a pay for performance contract: further evidence from the Quality and Outcomes Framework*. Number 34 in CHE Research Paper Series. University of York, Centre for Health Economics, February 2008. URL `http://www.york.ac.uk/media/che/documents/papers/researchpapers/rp34_doctor_behavior_under_a_pay_for_performance_contract_further_evidence.pdf`.

Bengt Holmstrom and Paul Milgrom. Multitask Principal-Agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization*, 7: 24–52, 1991.

David H. Howard. Quality and consumer choice in healthcare: Evidence from kidney transplantation. *Topics in Economic Analysis & Policy*, 5(1):1349, 2006.

Institute of Medicine. *Rewarding provider performance: aligning incentives in Medicare*. Pathways to quality health care. National Academies Press, Washington, DC, 2006.

Ashish K. Jha, Karen E. Joynt, E. John Orav, and Arnold M. Epstein. The Long-Term effect of premier pay for performance on patient outcomes. *The New England Journal of Medicine*, page 120328140019008, March 2012.

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

Nada F. Khan, Rafael Perera, Stephen Harper, and Peter W. Rose. Adaptation and validation of the Charlson Index for Read/OXMIS coded databases. *BMC Family Practice*, 11:1, January 2010.

Jeffrey B. Liebman and Richard J. Zeckhauser. Schmeduling. October 2004. URL `http://www.hks.harvard.edu/jeffreyliebman/schmeduling.pdf`.

Peter K. Lindenauer, Denise Remus, Sheila Roman, Michael B. Rothberg, Evan M. Benjamin, Allen Ma, and Dale W. Bratzler. Public reporting and pay for performance in hospital quality improvement. *JAMA*, 356(5):486–496, 2007.

Ching-to Albert Ma. Health care payment systems: Cost and quality incentives. *Journal of Economics & Management Strategy*, 3(1):93–112, 1994.

Ching-to Albert Ma and T. G. McGuire. Optimal health insurance and provider payment. *American Economic Review*, 87(4):685–704, 1997.

Mark McClellan, Aaron N. McKethan, Julie L. Lewis, Joachim Roski, and Elliott S. Fisher. A national strategy to put accountable care into practice. *Health Affairs (Millwood)*, 29(5): 982–90, 2010.

Elizabeth A. McGlynn, Steven M. Asch, John Adams, Joan Keesey, Jennifer Hicks, Alison DeCristofaro, and Eve A. Kerr. The quality of health care delivered to adults in the united states. *The New England Journal of Medicine*, 348(26):2635–45, 2003.

Thomas G. McGuire. Physician agency. In Anthony J. Culyer and Joseph P. Newhouse, editors, *Handbook of Health Economics*, pages 461–536. Elsevier, North-Holland, 2000.

Kathleen J. Mullen, Richard G. Frank, and Meredith B. Rosenthal. 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, 2010.

National Audit Office. NHS pay modernisation: New contracts for general practice services in England, February 2008. URL `http://www.nao.org.uk/publications/0708/new_contracts_for_general_prac.aspx`.

OECD. OECD health data 2011, June 2011. URL `http://www.oecd.org/health/healthdata`.

Thomas H. Payne, Don E. Detmer, Jeremy C. Wyatt, and Iain E. Buchan. National-Scale clinical information exchange in the united kingdom: Lessons for the united states. *Journal of the American Medical Informatics Association*, 18(1):91–98, January 2011.

Samuel S. Richardson. *Effect of payment reform on quality of primary care for coronary heart disease, diabetes, and chronic kidney disease in the United Kingdom*. PhD thesis, Harvard University, July 2013.

Martin Roland. Linking physicians' pay to the quality of care – a major experiment in the united kingdom. *The New England Journal of Medicine*, 351(14):1448–1454, 2004.

Meredith B. Rosenthal and Richard G. Frank. What is the empirical basis for paying for quality in health care? *Medical Care Research and Review*, 63(2):135–157, 2006.

Meredith B. Rosenthal, Richard G. Frank, Zhonghe Li, and Arnold M. Epstein. Early experience with Pay-for-Performance: from concept to practice. *JAMA*, 294(14): 1788–1793, 2005.

Andrew M. Ryan, James F. Burgess, Christopher P. Tompkins, and Stanley S. Wallack. The relationship between medicare's process of care quality measures and mortality. *Inquiry*, 46(3):274–290, September 2009.

M. E. Seddon, M. N. Marshall, S. M. Campbell, and M. O. Roland. Systematic review of studies of quality of clinical care in general practice in the UK, Australia and New Zealand. *Quality in Health Care*, 10(3):152–8, 2001.

Brian Serumaga, Dennis Ross-Degnan, Anthony J. Avery, Rachel A. Elliott, Sumit R. Majumdar, Fang Zhang, and Stephen B. Soumerai. Effect of pay for performance on the management and outcomes of hypertension in the United Kingdom: interrupted time series study. *BMJ*, 342, January 2011.

Tisamarie B. Sherry. A note on the comparative statics of Pay-for-Performance in health care. May 2012.

A. Michael Spence. Monopoly, quality, and regulation. *The Bell Journal of Economics*, 6(2): 417–429, 1975.

Matt Sutton, Silviya Nikolova, Ruth Boaden, Helen Lester, Ruth McDonald, and Martin Roland. Reduced mortality with hospital pay for performance in england. *The New England Journal of Medicine*, 367(19):1821–1828, 2012.

Abigail Tay. Assessing competition in hospital care markets: The importance of accounting for quality differentiation. *The RAND Journal of Economics*, 34(4):786–814, 2003.

The Health and Social Care Information Centre. GP earnings and expenses enquiry 2006/07, May 2009.

The Health and Social Care Information Centre. General and personal medical services: England, 2000-2010, March 2011a. URL `http://www.ic.nhs.uk/statistics-and-data-collections/workforce/nhs-staff-numbers/nhs-staff-2000--2010-general-practice`.

The Health and Social Care Information Centre. GP earnings and expenses 2009/10, November 2011b.

Jean Tirole. *The Theory of Industrial Organization*. MIT Press, 1988.

Burton A. Weisbrod. The Health-Care quadrilemma - an essay on Technological-Change, insurance, quality of care, and cost containment. *Journal of Economic Literature*, 29(2): 523–552, 1991.

Rachel M. Werner, Jonathan T. Kolstad, Elizabeth A. Stuart, and Daniel Polsky. The effect of Pay-For-Performance in hospitals: Lessons for quality improvement. *Health Affairs*, 30 (4):690–698, April 2011.

Diane Whalley, Hugh Gravelle, and Bonnie Sibbald. Effect of the new contract on GPs' working lives and perceptions of quality of care: a longitudinal survey. *The British Journal of General Practice*, 58(546):8–14, January 2008.

A
Appendices

A.1 Appendix to chapter 1: Proof of Proposition

Proposition. *Given a single payment contract, any equilibrium must be symmetric:* $q_{1a} = q_{1b}$; $q_{2a} = q_{2b}$.

Proof. Assume $(q_{1a}, q_{2a}) \neq (q_{1b}, q_{2b})$ in equilibrium. Define strictly concave average profit function for provider j: $A\pi_j\left(q_{1j}, q_{2j}\right) = p_1 q_{1j} + p_2 q_{2j} + p_b - (1 - p_c) \cdot c\left(q_{1j}, q_{2j}\right)$. Note that in equilibrium, each provider will operate in a region where $A\pi_j$ is weakly decreasing: otherwise, the provider could increase q_1 and/or q_2 to weakly increase demand and strictly increase average profits.

Claim. $A\pi_a > 0$ and $A\pi_b > 0$. If $A\pi_j = 0$ and $A\pi_{-j} > 0$, then provider j can strictly increase profits by choosing $\left(q_{1,-j}, q_{2,-j}\right)$. If $A\pi_a = A\pi_b = 0$, then at least one provider can strictly increase profits by decreasing q_1 or q_2, since providers are operating in a weakly decreasing region of $A\pi_j$. (Note that we are not at the corner where $j \in \{a, b\}$, $k \in \{1, 2\}$ $q_{kj} = 0$ or $\frac{\partial \pi_j}{\partial q_{kj}} = 0$, since $(q_{1a}, q_{2a}) \neq (q_{1b}, q_{2b})$ by assumption.)

Define vector $\mathbf{v} = (q_{1b} - q_{1a},\ q_{2b} - q_{2a})^T$. Since $A\pi_j$ is strictly concave and $(q_{1a}, q_{2a}) \neq (q_{1b}, q_{2b})$, $DA\pi_a\left(q_{1a}, q_{2a}\right) \cdot \mathbf{v} \neq DA\pi_b\left(q_{1b}, q_{2b}\right) \cdot \mathbf{v}$.

Define $D_+F\left(\mathbf{x}\right) \cdot \mathbf{v} = \lim_{h \to 0^+} \frac{F(\mathbf{x}+h\mathbf{v}) - F(\mathbf{x})}{h}$ and $D_-F\left(\mathbf{x}\right) \cdot \mathbf{v} = \lim_{h \to 0^-} \frac{F(\mathbf{x}+h\mathbf{v}) - F(\mathbf{x})}{h}$: these are the one-sided directional derivatives of F at $\mathbf{x}$ in the direction of $\mathbf{v}$.

Since $\frac{\partial_+ \mu_a}{\partial q_{ka}} = \frac{\partial_- \mu_b}{\partial q_{kb}}$ and $\frac{\partial_- \mu_a}{\partial q_{ka}} = \frac{\partial_+ \mu_b}{\partial q_{kb}}$,
$D_+\mu_a\left(q_{1a}, q_{2a}; q_{1b} q_{2b}\right) \cdot \mathbf{v} = D_-\mu_b\left(q_{1b}, q_{2b}; q_{1a} q_{2a}\right) \cdot \mathbf{v} \equiv D_+\mu_a$ and
$D_-\mu_a\left(q_{1a}, q_{2a}; q_{1b} q_{2b}\right) \cdot \mathbf{v} = D_+\mu_b\left(q_{1b}, q_{2b}; q_{1a} q_{2a}\right) \cdot \mathbf{v} \equiv D_-\mu_a$.

If both providers are maximizing profits, then the following four inequalities must hold:

$$\begin{aligned}
\mu_a \cdot \left[DA\pi_a\left(q_{1a}, q_{2a}\right) \cdot \mathbf{v}\right] + A\pi_a\left(D_+\mu_a \cdot \mathbf{v}\right) &\leq 0 \\
-\mu_a \cdot \left[DA\pi_a\left(q_{1a}, q_{2a}\right) \cdot \mathbf{v}\right] - A\pi_a\left(D_-\mu_a \cdot \mathbf{v}\right) &\leq 0 \\
\mu_b \cdot \left[DA\pi_b\left(q_{1b}, q_{2b}\right) \cdot \mathbf{v}\right] + A\pi_b\left(D_-\mu_a \cdot \mathbf{v}\right) &\leq 0 \\
-\mu_b \cdot \left[DA\pi_b\left(q_{1b}, q_{2b}\right) \cdot \mathbf{v}\right] - A\pi_b\left(D_+\mu_a \cdot \mathbf{v}\right) &\leq 0
\end{aligned}$$

$$\begin{aligned}
\Rightarrow A\pi_a\left(D_+\mu_a \cdot \mathbf{v} - D_-\mu_a \cdot \mathbf{v}\right) &\leq 0 \\
A\pi_b\left(D_-\mu_a \cdot \mathbf{v} - D_+\mu_a \cdot \mathbf{v}\right) &\leq 0
\end{aligned}$$

$$\Rightarrow D_+\mu_a \cdot \mathbf{v} \quad = \quad D_-\mu_a \cdot \mathbf{v} \equiv D\mu \cdot \mathbf{v}$$

This implies that the provider first-order conditions for a maximum must hold with equality:

$$\begin{aligned}
\mu_a \cdot \left[DA\pi_a\left(q_{1a}, q_{2a}\right) \cdot \mathbf{v}\right] + A\pi_a\left(D\mu \cdot \mathbf{v}\right) &= 0 \\
\mu_b \cdot \left[DA\pi_b\left(q_{1b}, q_{2b}\right) \cdot \mathbf{v}\right] + A\pi_b\left(D\mu \cdot \mathbf{v}\right) &= 0
\end{aligned}$$

$$\Rightarrow \frac{\mu_a \cdot [DA\pi_a (q_{1a}, q_{2a}) \cdot \mathbf{v}]}{A\pi_a} = \frac{\mu_b \cdot [DA\pi_b (q_{1b}, q_{2b}) \cdot \mathbf{v}]}{A\pi_b}$$

Claim. $\mu_a \neq \mu_b$ and $A\pi_a \neq A\pi_b$. Suppose $\mu_a = \mu_b$: $A\pi_a \neq A\pi_b$, since $DA\pi_a (q_{1a}, q_{2a}) \cdot \mathbf{v} \neq DA\pi_b (q_{1b}, q_{2b}) \cdot \mathbf{v}$. But in this case the provider with lower average profits could strictly increase profits by mimicking the provider with higher average profits. Suppose $A\pi_a = A\pi_b$: again, $\mu_a \neq \mu_b$ since $DA\pi_a (q_{1a}, q_{2a}) \cdot \mathbf{v} \neq DA\pi_b (q_{1b}, q_{2b}) \cdot \mathbf{v}$. In this case the provider with lower demand could strictly increase profits by mimicking the provider with higher demand.

Assume $\mu_a > \mu_b$. This implies that $A\pi_a < A\pi_b$, since otherwise, provider b could strictly increase profits by choosing quality vector (q_{1a}, q_{2a}). Since $A\pi_j$ is strictly concave, $|DA\pi_a (q_{1a}, q_{2a}) \cdot \mathbf{v}| > |DA\pi_b (q_{1b}, q_{2b}) \cdot \mathbf{v}|$. This implies $\left|\frac{\mu_a \cdot [DA\pi_a (q_{1a}, q_{2a}) \cdot \mathbf{v}]}{A\pi_a}\right| > \left|\frac{\mu_b \cdot [DA\pi_b (q_{1b}, q_{2b}) \cdot \mathbf{v}]}{A\pi_b}\right|$, which is a contradiction. By the same line of argument, we cannot have $\mu_b > \mu_a$. Therefore, $(q_{1a}, q_{2a}) = (q_{1b}, q_{2b})$ in equilibrium. $\square$

A.2 Appendix to chapter 2

Table A.2.1: Regression results for CKD quality, subsetting to patients without CHD, hypertension, diabetes, or history of stroke

Variable	**All** β (SE)	**Outcome** β (SE)	**Process** β (SE)
month	-0.295^{**}	-0.432^{**}	-0.333^{*}
	(0.0877)	(0.134)	(0.136)
*month*²	-0.0060^{***}	-0.0083^{***}	-0.0075^{**}
	(0.0015)	(0.0023)	(0.0024)
2004-2005	4.86^{***}	6.67^{***}	6.82^{***}
	(1.13)	(1.74)	(1.75)
2004-2005·month	0.458^{***}	0.591^{***}	0.613^{***}
	(0.0474)	(0.0753)	(0.0685)
*2004-2005·month*²	-0.0041	0.0001	-0.0141^{***}
	(0.0022)	(0.0036)	(0.0032)
post_2006	5.05^{***}	5.39^{**}	-1.17
	(1.26)	(1.95)	(1.92)
post_2006·month	0.490^{***}	0.728^{***}	0.674^{***}
	(0.0234)	(0.0373)	(0.0336)
*post_2006·month*²	0.0028^{***}	0.0044^{***}	0.0027^{***}
	(0.0002)	(0.0003)	(0.0003)
seasonally-adjusted estimate for *2004-2005*	4.68^{***} (1.17)	6.45^{***} (1.79)	6.80^{***} (1.81)
seasonality-adjusted estimate for jump in 2006	4.58^{***} (0.432)	4.06^{***} (0.689)	3.09^{***} (0.618)

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Units are percentage points of quality scores. Standard errors are robust to heteroskedasticity and clustering at the practice level. Unless otherwise noted, all coefficients are significant at $p < 0.001$. April 2004 is coded as month = 0. All regressions include patient fixed-effects, month-of-year dummies (to control for pre-QOF seasonality), month-of-year dummies interacted with *2004-2005* and *post_2006* (to control for differences in seasonality in each time period), and diagnosis indicators. Coefficients for trends and intercept in the phase-in periods are suppressed.

Table A.2.2: Effect estimates for CHD quality using six different patient fixed-effects models over seven different time periods

		Time period for estimated effects, in months						
	Model	**1**	**6**	**12**	**24**	**36**	**60**	**84**
All quality	*1yr*	9.12 (0.276)	9.12 (0.276)	9.12 (0.276)				
	2yr	12.3 (0.283)	12.3 (0.283)	12.3 (0.283)	12.3 (0.283)			
	CL	15.0 (0.323)	15.0 (0.323)	15.0 (0.323)	15.0 (0.323)	15.0 (0.323)	15.0 (0.323)	15.0 (0.323)
	VL	8.58 (0.495)	7.98 (0.527)	7.25 (0.568)	5.81 (0.653)	4.37 (0.744)	1.48$^{\mathrm{NS}}$ (0.932)	-1.41$^{\mathrm{NS}}$ (1.13)
	CQ	6.61 (0.430)	6.61 (0.430)	6.61 (0.430)	6.61 (0.430)	6.61 (0.430)	6.61 (0.430)	6.61 (0.430)
	VQ	9.91 (0.685)	10.7 (0.810)	11.7 (0.985)	13.9 (1.41)	16.4 (1.94)	22.2 (3.27)	28.9 (4.98)
Outcome	*1yr*	12.7 (0.400)	12.7 (0.400)	12.7 (0.400)				
	2yr	17.1 (0.412)	17.1 (0.412)	17.1 (0.412)	17.1 (0.412)			
	CL	20.8 (0.493)	20.8 (0.493)	20.8 (0.493)	20.8 (0.493)	20.8 (0.493)	20.8 (0.493)	20.8 (0.493)
	VL	12.3 (0.736)	11.5 (0.788)	10.5 (0.854)	8.56 (0.990)	6.63 (1.13)	2.78$^{\mathrm{NS}}$ (1.43)	-1.08$^{\mathrm{NS}}$ (1.73)
	CQ	9.01 (0.618)	9.01 (0.618)	9.01 (0.618)	9.01 (0.618)	9.01 (0.618)	9.01 (0.618)	9.01 (0.618)
	VQ	10.9 (1.06)	11.6 (1.25)	12.3 (1.52)	13.9 (2.18)	15.3 (2.97)	18.2 (4.98)	21.0$^{p<.01}$ (7.55)
Process	*1yr*	7.36 (0.239)	7.36 (0.239)	7.36 (0.239)				
	2yr	9.80 (0.245)	9.80 (0.245)	9.80 (0.245)	9.80 (0.245)			
	CL	12.5 (0.304)	12.5 (0.304)	12.5 (0.304)	12.5 (0.304)	12.5 (0.304)	12.5 (0.304)	12.5 (0.304)
	VL	6.17 (0.410)	5.57 (0.435)	4.86 (0.466)	3.44 (0.533)	2.02$^{p<.01}$ (0.603)	-0.829$^{\mathrm{NS}}$ (0.752)	-3.67 (0.906)
	CQ	5.21 (0.362)	5.21 (0.362)	5.21 (0.362)	5.21 (0.362)	5.21 (0.362)	5.21 (0.362)	5.21 (0.362)
	VQ	9.96 (0.572)	10.9 (0.675)	12.2 (0.821)	15.2 (1.18)	18.8 (1.62)	27.5 (2.75)	38.3 (4.20)

All models allow for intercept shifts in each period. *1yr* and *2yr* include no trends (comparing adjusted average quality scores for the one year or two years, respectively, around period changes). *CL* has a constant linear treand, *VL* allows variable linear trends, *CQ* assumes a constant quadratic trend, and *VQ* allows variable quadratic trends, as in the baseline analysis.

Units are percentage points of quality scores. Standard errors are robust to heteroskedasticity and clustering at the practice level. Unless otherwise noted, all estimates are significant at $p < 0.001$. All regressions include diagnosis indicators, and models including linear or quadratic trends control for seasonality.

Table A.2.3: Effect estimates for CHD quality using six different practice-level fixed-effects models over seven different time periods

		Time period for estimated effects, in months						
	Model	**1**	**6**	**12**	**24**	**36**	**60**	**84**
All quality	*1yr*	14.1	14.1	14.1				
		(0.290)	(0.290)	(0.290)				
	2yr	18.5	18.5	18.5	18.5			
		(0.302)	(0.302)	(0.302)	(0.302)			
	CL	17.9	17.9	17.9	17.9	17.9	17.9	17.9
		(0.365)	(0.365)	(0.365)	(0.365)	(0.365)	(0.365)	(0.365)
	VL	9.18	8.40	7.46	5.60	3.73	-0.007^{NS}	$-3.74^{p<.01}$
		(0.510)	(0.544)	(0.586)	(0.676)	(0.770)	(0.966)	(1.17)
	CQ	7.40	7.40	7.40	7.40	7.40	7.40	7.40
		(0.440)	(0.440)	(0.440)	(0.440)	(0.440)	(0.440)	(0.440)
	VQ	10.3	11.0	11.9	13.9	16.1	21.3	27.5
		(0.723)	(0.858)	(1.05)	(1.51)	(2.07)	(3.49)	(5.31)
Outcome	*1yr*	19.7	19.7	19.7				
		(0.423)	(0.423)	(0.423)				
	2yr	25.8	25.8	25.8	25.8			
		(0.440)	(0.440)	(0.440)	(0.440)			
	CL	24.6	24.6	24.6	24.6	24.6	24.6	24.6
		(0.519)	(0.519)	(0.519)	(0.519)	(0.519)	(0.519)	(0.519)
	VL	13.1	12.0	10.8	8.34	5.88	0.943^{NS}	$-3.99^{p<.05}$
		(0.770)	(0.825)	(0.893)	(1.04)	(1.19)	(1.49)	(1.81)
	CQ	10.1	10.1	10.1	10.1	10.1	10.1	10.1
		(0.646)	(0.646)	(0.646)	(0.646)	(0.646)	(0.646)	(0.646)
	VQ	11.4	11.9	12.5	13.5	14.6	$16.6^{p<.01}$	$18.5^{p<.05}$
		(1.12)	(1.33)	(1.63)	(2.33)	(3.20)	(5.37)	(8.15)
Process	*1yr*	11.9	11.9	11.9				
		(0.255)	(0.255)	(0.255)				
	2yr	15.5	15.5	15.5	15.5			
		(0.265)	(0.265)	(0.265)	(0.265)			
	CL	15.4	15.4	15.4	15.4	15.4	15.4	15.4
		(0.332)	(0.332)	(0.332)	(0.332)	(0.332)	(0.332)	(0.332)
	VL	7.15	6.41	5.53	3.77	$2.01^{p<.01}$	-1.51^{NS}	-5.04
		(0.435)	(0.462)	(0.495)	(0.567)	(0.644)	(0.804)	(0.970)
	CQ	6.12	6.12	6.12	6.12	6.12	6.12	6.12
		(0.381)	(0.381)	(0.381)	(0.381)	(0.381)	(0.381)	(0.381)
	VQ	10.7	14.6	12.8	15.7	19.1	27.4	37.7
		(0.612)	(0.725)	(0.884)	(1.27)	(1.75)	(2.98)	(4.54)

All models allow for intercept shifts in each period. *1yr* and *2yr* include no trends (comparing adjusted average quality scores for the one year or two years, respectively, around period changes). *CL* has a constant linear treand, *VL* allows variable linear trends, *CQ* assumes a constant quadratic trend, and *VQ* allows variable quadratic trends, as in the baseline analysis.

Units are percentage points of quality scores. Standard errors are robust to heteroskedasticity and clustering at the practice level. All models include practice fixed effects, and observations are weighted by the practice's average number of CHD patients over the time period. Unless otherwise noted, all estimates are significant at $p < 0.001$. Models including linear or quadratic trends control for seasonality.

Table A.2.4: Effect estimates for diabetes quality using six different patient fixed-effects models over seven different time periods

		Time period for estimated effects, in months						
	Model	**1**	**6**	**12**	**24**	**36**	**60**	**84**
All quality	*1yr*	6.81	6.81	6.81				
		(0.258)	(0.258)	(0.258)				
	2yr	8.64	8.64	8.64	8.64			
		(0.486)	(0.486)	(0.486)	(0.486)			
	CL	8.48	8.48	8.48	8.48	8.48	8.48	8.48
		(0.517)	(0.517)	(0.517)	(0.517)	(0.517)	(0.517)	(0.517)
	VL	6.49	6.31	6.09	5.65	5.21	$4.33^{p<.01}$	$3.45^{p<.05}$
		(0.690)	(0.730)	(0.782)	(0.891)	(1.01)	(1.25)	(1.51)
	CQ	3.73	3.73	3.73	3.73	3.73	3.73	3.73
		(0.518)	(0.518)	(0.518)	(0.518)	(0.518)	(0.518)	(0.518)
	VQ	8.32	9.32	10.6	13.4	16.5	23.6	32.1
		(0.986)	(1.17)	(1.41)	(2.00)	(2.72)	(4.50)	(6.77)
Outcome	*1yr*	2.61	2.61	2.61				
		(0.383)	(0.383)	(0.383)				
	2yr	3.59	3.59	3.59	3.59			
		(0.689)	(0.689)	(0.689)	(0.689)			
	CL	5.19	5.19	5.19	5.19	5.19	5.19	5.19
		(0.732)	(0.732)	(0.732)	(0.732)	(0.732)	(0.732)	(0.732)
	VL	$2.46^{p<.01}$	$2.21^{p<.05}$	1.91^{NS}	1.30^{NS}	0.700^{NS}	-0.508^{NS}	-1.72^{NS}
		(0.909)	(0.963)	(1.03)	(1.18)	(1.34)	(1.68)	(2.03)
	CQ	0.373^{NS}	0.373^{NS}	0.373^{NS}	0.373^{NS}	0.373^{NS}	0.373^{NS}	0.373^{NS}
		(0.712)	(0.712)	(0.712)	(0.712)	(0.712)	(0.712)	(0.712)
	VQ	5.79	6.94	8.42	11.8	15.6	24.8	35.9
		(1.34)	(1.59)	(1.94)	(2.78)	(3.79)	(6.34)	(9.59)
Process	*1yr*	13.3	13.3	13.3				
		(0.495)	(0.495)	(0.495)				
	2yr	16.1	16.1	16.1	16.1			
		(0.533)	(0.533)	(0.533)	(0.533)			
	CL	14.8	14.8	14.8	14.8	14.8	14.8	14.8
		(0.590)	(0.590)	(0.590)	(0.590)	(0.590)	(0.590)	(0.590)
	VL	11.9	11.6	11.3	10.7	10.0	8.73	7.45
		(0.839)	(0.885)	(0.943)	(1.06)	(1.19)	(1.46)	(1.74)
	CQ	9.29	9.29	9.29	9.29	9.29	9.29	9.29
		(0.635)	(0.635)	(0.635)	(0.635)	(0.635)	(0.635)	(0.635)
	VQ	13.4	14.4	15.5	18.0	20.8	27.3	35.0
		(1.17)	(1.37)	(1.65)	(2.32)	(3.14)	(5.19)	(7.80)

All models allow for intercept shifts in each period. *1yr* and *2yr* include no trends (comparing adjusted average quality scores for the one year or two years, respectively, around period changes). *CL* has a constant linear treand, *VL* allows variable linear trends, *CQ* assumes a constant quadratic trend, and *VQ* allows variable quadratic trends, as in the baseline analysis.

Units are percentage points of quality scores. Standard errors are robust to heteroskedasticity and clustering at the practice level. Unless otherwise noted, all estimates are significant at $p < 0.001$. All regressions include diagnosis indicators, and models including linear or quadratic trends control for seasonality.

Table A.2.5: Effect estimates for diabetes quality using six different practice-level fixed-effects models over seven different time periods

		Time period for estimated effects, in months						
	Model	**1**	**6**	**12**	**24**	**36**	**60**	**84**
All quality	*1yr*	12.4	12.4	12.4				
		(0.432)	(0.432)	(0.432)				
	2yr	15.6	15.6	15.6	15.6			
		(0.476)	(0.476)	(0.476)	(0.476)			
	CL	13.7	13.7	13.7	13.7	13.7	13.7	13.7
		(0.534)	(0.534)	(0.534)	(0.534)	(0.534)	(0.534)	(0.534)
	VL	10.0	9.70	9.30	8.52	7.73	6.16	$4.59^{p<.01}$
		(0.754)	(0.803)	(0.864)	(0.992)	(1.13)	(1.41)	(1.70)
	CQ	6.85	6.85	6.85	6.85	6.85	6.85	6.85
		(0.609)	(0.609)	(0.609)	(0.609)	(0.609)	(0.609)	(0.609)
	VQ	10.8	11.9	13.2	15.9	18.9	25.4	32.8
		(1.33)	(1.58)	(1.92)	(2.74)	(3.72)	(6.18)	(9.31)
Outcome	*1yr*	10.2	10.2	10.2				
		(0.524)	(0.524)	(0.524)				
	2yr	13.4	13.4	13.4	13.4			
		(0.580)	(0.580)	(0.580)	(0.580)			
	CL	12.0	12.0	12.0	12.0	12.0	12.0	12.0
		(0.650)	(0.650)	(0.650)	(0.650)	(0.650)	(0.650)	(0.650)
	VL	6.53	6.04	5.46	4.30	$3.13^{p<.05}$	0.808^{NS}	-1.52^{NS}
		(0.883)	(0.947)	(1.03)	(1.20)	(1.38)	(1.75)	(2.13)
	CQ	3.69	3.69	3.69	3.69	3.69	3.69	3.69
		(0.708)	(0.708)	(0.708)	(0.708)	(0.708)	(0.708)	(0.708)
	VQ	7.85	8.87	10.2	13.0	$16.1^{p<.01}$	$23.1^{p<.01}$	$31.2^{p<.05}$
		(1.68)	(2.01)	(2.47)	(3.55)	(4.86)	(8.13)	(12.3)
Process	*1yr*	18.2	18.2	18.2				
		(0.643)	(0.643)	(0.643)				
	2yr	22.0	22.0	22.0	22.0			
		(0.681)	(0.681)	(0.681)	(0.681)			
	CL	20.2	20.2	20.2	20.2	20.2	20.2	20.2
		(0.746)	(0.746)	(0.746)	(0.746)	(0.746)	(0.746)	(0.746)
	VL	16.0	15.6	15.1	14.2	13.4	11.6	9.77
		(1.07)	(1.13)	(1.21)	(1.37)	(1.54)	(1.59)	(2.25)
	CQ	12.8	12.8	12.8	12.8	12.8	12.8	12.8
		(0.885)	(0.885)	(0.885)	(0.885)	(0.885)	(0.885)	(0.885)
	VQ	17.2	18.3	19.7	22.7	25.9	33.3	41.7
		(1.67)	(1.95)	(2.35)	(3.31)	(4.47)	(7.42)	(11.2)

All models allow for intercept shifts in each period. *1yr* and *2yr* include no trends (comparing adjusted average quality scores for the one year or two years, respectively, around period changes). *CL* has a constant linear treand, *VL* allows variable linear trends, *CQ* assumes a constant quadratic trend, and *VQ* allows variable quadratic trends, as in the baseline analysis.

Units are percentage points of quality scores. Standard errors are robust to heteroskedasticity and clustering at the practice level. All models include practice fixed effects, and observations are weighted by the practice's average number of diabetic patients over the time period. Unless otherwise noted, all estimates are significant at $p < 0.001$. Models including linear or quadratic trends control for seasonality.

Table A.2.6: Effect estimates for CKD overall quality using six different patient fixed-effects models over seven different time periods

		Time period for estimated effects, in months					
	Model	**1**	**6**	**12**	**24**	**36**	**60**
FY06 vs FY04	*1yr*	5.28	5.28	5.28			
		(0.215)	(0.215)	(0.215)			
	2yr	5.95	5.95	5.95	5.95		
		(0.194)	(0.194)	(0.194)	(0.194)		
	CL	7.00	7.00	7.00	7.00	7.00	7.00
		(0.277)	(0.277)	(0.277)	(0.277)	(0.277)	(0.277)
	VL	4.67	4.18	3.59	2.41	1.23	-1.12$^{p<.05}$
		(0.237)	(0.241)	(0.253)	(0.289)	(0.341)	(0.466)
	CQ	5.58	5.58	5.58	5.58	5.58	5.58
		(0.263)	(0.263)	(0.263)	(0.263)	(0.263)	(0.263)
	VQ	5.44	5.91	6.63	8.62	11.3	19.0
		(0.274)	(0.311)	(0.398)	(0.676)	(1.06)	(2.11)
FY06 vs pre	*1yr*	10.3	10.3	10.3			
		(0.332)	(0.332)	(0.332)			
	2yr	11.4	11.4	11.4	11.4		
		(0.315)	(0.315)	(0.315)	(0.315)		
	CL	13.8	13.8	13.8	13.8	13.8	13.8
		(0.531)	(0.531)	(0.531)	(0.531)	(0.531)	(0.531)
	VL	7.98	7.66	7.27	6.50	5.73	4.18
		(0.706)	(0.731)	(0.761)	(0.825)	(0.892)	(1.03)
	CQ	7.46	7.46	7.46	7.46	7.46	7.46
		(0.506)	(0.506)	(0.506)	(0.506)	(0.506)	(0.506)
	VQ	11.1	11.4	11.7	12.4	13.1$^{p<.01}$	14.4$^{p<.05}$
		(2.25)	(2.48)	(2.77)	(3.42)	(4.16)	(5.90)
					20m		
FY04 vs pre	*1yr*	5.03	5.03	5.03			
		(0.225)	(0.225)	(0.225)			
	2yr	5.50	5.50	5.50	5.50		
		(0.229)	(0.229)	(0.229)	(0.229)		
	CL	6.82	6.82	6.82	6.82		
		(0.337)	(0.337)	(0.337)	(0.337)		
	VL	1.70	1.86	2.07	2.34		
		(0.363)	(0.386)	(0.418)	(0.468)		
	CQ	1.88	1.88	1.88	1.88		
		(0.340)	(0.340)	(0.340)	(0.340)		
	VQ	2.19	2.94	3.67	4.35		
		(0.609)	(0.734)	(0.903)	(1.16)		

All models allow for intercept shifts in each period. *1yr* and *2yr* include no trends (comparing adjusted average quality scores for the one year or two years, respectively, around period changes). *CL* has a constant linear treand, *VL* allows variable linear trends, *CQ* assumes a constant quadratic trend, and *VQ* allows variable quadratic trends, as in the baseline analysis.

Units are percentage points of quality scores. Standard errors are robust to heteroskedasticity and clustering at the practice level. Unless otherwise noted, all estimates are significant at $p < 0.001$. All regressions include diagnosis indicators, and models including linear or quadratic trends control for seasonality.

Table A.2.7: Effect estimates for CKD outcomes quality using six different patient fixed-effects models over seven different time periods

		Time period for estimated effects, in months					
	Model	**1**	**6**	**12**	**24**	**36**	**60**
FY06 vs FY04	*1yr*	7.76	7.76	7.76			
		(0.334)	(0.334)	(0.334)			
	2yr	9.01	9.01	9.01	9.01		
		(0.324)	(0.324)	(0.324)	(0.324)		
	CL	8.58	8.58	8.58	8.58	8.58	8.58
		(0.420)	(0.420)	(0.420)	(0.420)	(0.420)	(0.420)
	VL	5.69	5.09	4.36	2.92	1.46$^{p<.05}$	-1.43^{NS}
		(0.370)	(0.387)	(0.419)	(0.508)	(0.618)	(0.871)
	CQ	6.92	6.92	6.92	6.92	6.92	6.92
		(0.390)	(0.390)	(0.390)	(0.390)	(0.390)	(0.390)
	VQ	6.19	6.81	7.73	10.1	13.4	22.2
		(0.442)	(0.552)	(0.753)	(1.33)	(2.11)	(4.22)
FY06 vs pre	*1yr*	16.4	16.4	16.4			
		(0.517)	(0.517)	(0.517)			
	2yr	18.1	18.1	18.1	18.1		
		(0.532)	(0.532)	(0.532)	(0.532)		
	CL	16.5	16.5	16.5	16.5	16.5	16.5
		(0.832)	(0.832)	(0.832)	(0.832)	(0.832)	(0.832)
	VL	11.1	10.8	10.4	9.70	8.96	7.48
		(1.32)	(1.38)	(1.45)	(1.59)	(1.75)	(2.06)
	CQ	8.98	8.98	8.98	8.98	8.98	8.98
		(0.792)	(0.792)	(0.792)	(0.792)	(0.792)	(0.792)
	VQ	18.4	19.4	20.5	22.8	25.2$^{p<.01}$	30.2$^{p<.01}$
		(4.21)	(4.63)	(5.18)	(6.40)	(7.77)	(11.0)
					20m		
FY04 vs pre	*1yr*	8.60	8.60	8.60			
		(0.407)	(0.407)	(0.407)			
	2yr	9.11	9.11	9.11	9.11		
		(0.424)	(0.424)	(0.424)	(0.424)		
	CL	7.89	7.89	7.89	7.89		
		(0.592)	(0.592)	(0.592)	(0.592)		
	VL	2.58	2.87	3.23	3.70		
		(0.720)	(0.766)	(0.829)	(0.925)		
	CQ	2.06$^{p<.01}$	2.06$^{p<.01}$	2.06$^{p<.01}$	2.06$^{p<.01}$		
		(0.613)	(0.613)	(0.613)	(0.613)		
	VQ	4.29	5.50	6.78	8.22		
		(1.16)	(1.39)	(1.71)	(2.19)		

All models allow for intercept shifts in each period. *1yr* and *2yr* include no trends (comparing adjusted average quality scores for the one year or two years, respectively, around period changes). *CL* has a constant linear treand, *VL* allows variable linear trends, *CQ* assumes a constant quadratic trend, and *VQ* allows variable quadratic trends, as in the baseline analysis.

Units are percentage points of quality scores. Standard errors are robust to heteroskedasticity and clustering at the practice level. Unless otherwise noted, all estimates are significant at $p < 0.001$. All regressions include diagnosis indicators, and models including linear or quadratic trends control for seasonality.

Table A.2.8: Effect estimates for CKD process quality using six different patient fixed-effects models over seven different time periods

		Time period for estimated effects, in months					
	Model	**1**	**6**	**12**	**24**	**36**	**60**
FY06 vs FY04	*1yr*	1.48	1.48	1.48			
		(0.134)	(0.134)	(0.134)			
	2yr	2.25	2.25	2.25	2.25		
		(0.134)	(0.134)	(0.134)	(0.134)		
	CL	2.78	2.78	2.78	2.78	2.78	2.78
		(0.166)	(0.166)	(0.166)	(0.166)	(0.166)	(0.166)
	VL	2.08	1.92	1.72	1.32	$0.930^{p<.01}$	0.142^{NS}
		(0.165)	(0.182)	(0.206)	(0.263)	(0.327)	(0.464)
	CQ	1.68	1.68	1.68	1.68	1.68	1.68
		(0.183)	(0.183)	(0.183)	(0.183)	(0.183)	(0.183)
	VQ	2.66	3.52	4.73	7.67	11.3	20.8
		(0.199)	(0.271)	(0.393)	(0.729)	(1.17)	(2.39)
FY06 vs pre	*1yr*	4.83	4.83	4.83			
		(0.248)	(0.248)	(0.248)			
	2yr	6.12	6.12	6.12	6.12		
		(0.271)	(0.271)	(0.271)	(0.271)		
	CL	7.38	7.38	7.38	7.38	7.38	7.38
		(0.393)	(0.393)	(0.393)	(0.393)	(0.393)	(0.393)
	VL	2.83	$2.59^{p<.01}$	$2.30^{p<.01}$	1.73^{NS}	1.15^{NS}	-0.006^{NS}
		(0.748)	(0.783)	(0.825)	(0.911)	(1.00)	(1.18)
	CQ	2.45	2.45	2.45	2.45	2.45	2.45
		(0.439)	(0.439)	(0.439)	(0.439)	(0.439)	(0.439)
	VQ	3.62^{NS}	3.84^{NS}	4.08^{NS}	4.47^{NS}	4.74^{NS}	4.93^{NS}
		(2.43)	(2.68)	(3.00)	(3.72)	(4.53)	(6.43)
					20m		
FY04 vs pre	*1yr*	3.35	3.35	3.35			
		(0.226)	(0.226)	(0.226)			
	2yr	3.87	3.87	3.87	3.87		
		(0.245)	(0.245)	(0.245)	(0.245)		
	CL	4.60	4.60	4.60	4.60		
		(0.337)	(0.337)	(0.337)	(0.337)		
	VL	1.48	$1.41^{p<.01}$	$1.31^{p<.01}$	$1.19^{p<.05}$		
		(0.388)	(0.417)	(0.457)	(0.517)		
	CQ	$0.768^{p<.05}$	$0.768^{p<.05}$	$0.768^{p<.05}$	$0.768^{p<.05}$		
		(0.348)	(0.348)	(0.348)	(0.348)		
	VQ	$1.40^{p<.05}$	$1.80^{p<.05}$	$2.08^{p<.05}$	2.13^{NS}		
		(0.636)	(0.771)	(0.956)	(1.24)		

All models allow for intercept shifts in each period. *1yr* and *2yr* include no trends (comparing adjusted average quality scores for the one year or two years, respectively, around period changes). *CL* has a constant linear treand, *VL* allows variable linear trends, *CQ* assumes a constant quadratic trend, and *VQ* allows variable quadratic trends, as in the baseline analysis.

Units are percentage points of quality scores. Standard errors are robust to heteroskedasticity and clustering at the practice level. Unless otherwise noted, all estimates are significant at $p < 0.001$. All regressions include diagnosis indicators, and models including linear or quadratic trends control for seasonality.

Table A.2.9: Effect estimates for CKD overall quality using six different practice-level fixed-effects models over seven different time periods

		Time period for estimated effects, in months					
	Model	**1**	**6**	**12**	**24**	**36**	**60**
FY06 vs FY04	*1yr*	6.59	6.59	6.59			
		(0.219)	(0.219)	(0.219)			
	2yr	7.12	7.12	7.12	7.12		
		(0.191)	(0.191)	(0.191)	(0.191)		
	CL	6.16	6.16	6.16	6.16	6.16	6.16
		(0.265)	(0.265)	(0.265)	(0.265)	(0.265)	(0.265)
	VL	4.73	4.33	3.84	2.87	1.89	-0.058^{NS}
		(0.255)	(0.269)	(0.290)	(0.344)	(0.407)	(0.549)
	CQ	5.53	5.53	5.53	5.53	5.53	5.53
		(0.292)	(0.292)	(0.292)	(0.292)	(0.292)	(0.292)
	VQ	5.92	6.61	7.63	10.4	14.0	24.0
		(0.308)	(0.388)	(0.538)	(0.978)	(1.58)	(3.22)
FY06 vs pre	*1yr*	12.3	12.3	12.3			
		(0.361)	(0.361)	(0.361)			
	2yr	13.3	13.3	13.3	13.3		
		(0.363)	(0.363)	(0.363)	(0.363)		
	CL	12.8	12.8	12.8	12.8	12.8	12.8
		(0.512)	(0.512)	(0.512)	(0.512)	(0.512)	(0.512)
	VL	6.88	6.54	6.12	5.29	4.46$^{p<.01}$	2.80^{NS}
		(1.03)	(1.08)	(1.14)	(1.25)	(1.37)	(1.61)
	CQ	8.00	8.00	8.00	8.00	8.00	8.00
		(0.699)	(0.699)	(0.699)	(0.699)	(0.699)	(0.699)
	VQ	13.8	14.4$^{p<.01}$	15.2$^{p<.01}$	16.8$^{p<.01}$	18.6$^{p<.01}$	22.5$^{p<.05}$
		(3.87)	(4.25)	(4.74)	(5.81)	(7.01)	(9.80)
					20m		
FY04 vs pre	*1yr*	5.70	5.70	5.70			
		(0.291)	(0.291)	(0.291)			
	2yr	6.21	6.21	6.21	6.21		
		(0.318)	(0.318)	(0.318)	(0.318)		
	CL	6.59	6.59	6.59	6.59		
		(0.398)	(0.398)	(0.398)	(0.398)		
	VL	1.58$^{p<.01}$	1.64$^{p<.01}$	1.71$^{p<.01}$	1.80$^{p<.05}$		
		(0.541)	(0.585)	(0.642)	(0.724)		
	CQ	2.47	2.47	2.47	2.47		
		(0.504)	(0.504)	(0.504)	(0.504)		
	VQ	3.07$^{p<.01}$	3.98$^{p<.01}$	4.90$^{p<.01}$	5.82$^{p<.01}$		
		(1.13)	(1.35)	(1.64)	(2.07)		

All models allow for intercept shifts in each period. *1yr* and *2yr* include no trends (comparing adjusted average quality scores for the one year or two years, respectively, around period changes). *CL* has a constant linear treand, *VL* allows variable linear trends, *CQ* assumes a constant quadratic trend, and *VQ* allows variable quadratic trends, as in the baseline analysis.

Units are percentage points of quality scores. Standard errors are robust to heteroskedasticity and clustering at the practice level. All models include practice fixed effects, and observations are weighted by the practice's average number of patients over the time period. Unless otherwise noted, all estimates are significant at $p < 0.001$. Models including linear or quadratic trends control for seasonality.

Table A.2.10: Effect estimates for CKD outcomes quality using six different practice-level fixed-effects models over seven different time periods

	Model	**Time period for estimated effects, in months**					
		1	**6**	**12**	**24**	**36**	**60**
FY06 vs FY04	*1yr*	10.6	10.6	10.6			
		(0.400)	(0.400)	(0.400)			
	2yr	11.5	11.5	11.5	11.5		
		(0.357)	(0.357)	(0.357)	(0.357)		
	CL	9.38	9.38	9.38	9.38	9.38	9.38
		(0.474)	(0.474)	(0.474)	(0.474)	(0.474)	(0.474)
	VL	6.90	6.28	5.54	4.07	2.59$^{p<.01}$	-0.356^{NS}
		(0.475)	(0.504)	(0.548)	(0.658)	(0.787)	(1.07)
	CQ	8.66	8.66	8.66	8.66	8.66	8.66
		(0.501)	(0.501)	(0.501)	(0.501)	(0.501)	(0.501)
	VQ	8.39	9.26	10.6	14.1	18.9	32.0
		(0.592)	(0.758)	(1.06)	(1.91)	(3.07)	(6.26)
FY06 vs pre	*1yr*	20.0	20.0	20.0			
		(0.623)	(0.623)	(0.623)			
	2yr	21.5	21.5	21.5	21.5		
		(0.625)	(0.625)	(0.625)	(0.625)		
	CL	18.2	18.2	18.2	18.2	18.2	18.2
		(0.927)	(0.927)	(0.927)	(0.927)	(0.927)	(0.927)
	VL	12.2	11.9	11.4	10.5	9.65	7.89$^{p<.01}$
		(1.66)	(1.73)	(1.82)	(1.99)	(2.17)	(2.53)
	CQ	12.8	12.8	12.8	12.8	12.8	12.8
		(1.17)	(1.17)	(1.17)	(1.17)	(1.17)	(1.17)
	VQ	21.48$^{p<.01}$	22.3$^{p<.01}$	23.5$^{p<.01}$	26.0$^{p<.05}$	28.6$^{p<.05}$	34.3$^{p<.05}$
		(6.79)	(7.46)	(8.32)	(10.2)	(12.3)	(17.2)
					20m		
FY04 vs pre	*1yr*	9.38	9.38	9.38			
		(0.493)	(0.493)	(0.493)			
	2yr	9.99	9.99	9.99	9.99		
		(0.534)	(0.534)	(0.534)	(0.534)		
	CL	8.87	8.87	8.87	8.87		
		(0.696)	(0.696)	(0.696)	(0.696)		
	VL	2.94$^{p<.01}$	3.19$^{p<.01}$	3.49$^{p<.01}$	3.89$^{p<.01}$		
		(0.913)	(0.977)	(1.06)	(1.19)		
	CQ	4.11	4.11	4.11	4.11		
		(0.865)	(0.865)	(0.865)	(0.865)		
	VQ	4.94$^{p<.05}$	6.33$^{p<.01}$	7.77$^{p<.01}$	9.27$^{p<.05}$		
		(2.00)	(2.39)	(2.91)	(3.68)		

All models allow for intercept shifts in each period. *1yr* and *2yr* include no trends (comparing adjusted average quality scores for the one year or two years, respectively, around period changes). *CL* has a constant linear treand, *VL* allows variable linear trends, *CQ* assumes a constant quadratic trend, and *VQ* allows variable quadratic trends, as in the baseline analysis.

Units are percentage points of quality scores. Standard errors are robust to heteroskedasticity and clustering at the practice level. All models include practice fixed effects, and observations are weighted by the practice's average number of patients over the time period. Unless otherwise noted, all estimates are significant at $p < 0.001$. Models including linear or quadratic trends control for seasonality.

Table A.2.11: Effect estimates for CKD process quality using six different practice-level fixed-effects models over seven different time periods

		Time period for estimated effects, in months					
	Model	**1**	**6**	**12**	**24**	**36**	**60**
FY06 vs FY04	*1yr*	2.37	2.37	2.37			
		(0.150)	(0.150)	(0.150)			
	2yr	3.06	3.06	3.06	3.06		
		(0.154)	(0.154)	(0.154)	(0.154)		
	CL	1.97	1.97	1.97	1.97	1.97	1.97
		(0.229)	(0.229)	(0.229)	(0.229)	(0.229)	(0.229)
	VL	2.22	2.13	2.03	1.82	1.61	1.19$^{p<.05}$
		(0.185)	(0.203)	(0.229)	(0.290)	(0.357)	(0.498)
	CQ	1.38	1.38	1.38	1.38	1.38	1.38
		(0.315)	(0.315)	(0.315)	(0.315)	(0.315)	(0.315)
	VQ	3.11	4.11	5.50	8.90	13.1	24.0
		(0.254)	(0.348)	(0.502)	(0.923)	(1.48)	(3.01)
FY06 vs pre	*1yr*	6.40	6.40	6.40			
		(0.427)	(0.427)	(0.427)			
	2yr	7.70	7.70	7.70	7.70		
		(0.472)	(0.472)	(0.472)	(0.472)		
	CL	6.79	6.79	6.79	6.79	6.79	6.79
		(0.506)	(0.506)	(0.506)	(0.506)	(0.506)	(0.506)
	VL	0.711^{NS}	0.371^{NS}	-0.038^{NS}	-0.854^{NS}	-1.67^{NS}	-3.30^{NS}
		(1.47)	(1.54)	(1.64)	(1.82)	(2.01)	(2.38)
	CQ	2.35$^{p<.05}$	2.35$^{p<.05}$	2.35$^{p<.05}$	2.35$^{p<.05}$	2.35$^{p<.05}$	2.35$^{p<.05}$
		(0.930)	(0.930)	(0.930)	(0.930)	(0.930)	(0.930)
	VQ	10.8$^{p<.05}$	11.8$^{p<.05}$	13.1$^{p<.05}$	15.8^{NS}	18.7^{NS}	25.1^{NS}
		(5.33)	(5.87)	(6.56)	(8.09)	(9.81)	(13.8)
					20m		
FY04 vs pre	*1yr*	4.03	4.03	4.03			
		(0.401)	(0.401)	(0.401)			
	2yr	4.64	4.64	4.64	4.64		
		(0.452)	(0.452)	(0.452)	(0.452)		
	CL	4.82	4.82	4.82	4.82		
		(0.513)	(0.513)	(0.513)	(0.513)		
	VL	0.926^{NS}	0.672^{NS}	0.368^{NS}	-0.037^{NS}		
		(0.765)	(0.833)	(0.918)	(1.04)		
	CQ	0.971^{NS}	0.971^{NS}	0.971^{NS}	0.971^{NS}		
		(0.698)	(0.698)	(0.698)	(0.698)		
	VQ	3.41$^{p<.05}$	4.18$^{p<.05}$	4.97$^{p<.05}$	5.78$^{p<.05}$		
		(1.47)	(1.77)	(2.16)	(2.77)		

All models allow for intercept shifts in each period. *1yr* and *2yr* include no trends (comparing adjusted average quality scores for the one year or two years, respectively, around period changes). *CL* has a constant linear treand, *VL* allows variable linear trends, *CQ* assumes a constant quadratic trend, and *VQ* allows variable quadratic trends, as in the baseline analysis.

Units are percentage points of quality scores. Standard errors are robust to heteroskedasticity and clustering at the practice level. All models include practice fixed effects, and observations are weighted by the practice's average number of patients over the time period. Unless otherwise noted, all estimates are significant at $p < 0.001$. Models including linear or quadratic trends control for seasonality.

A.3 Appendix to chapter 3

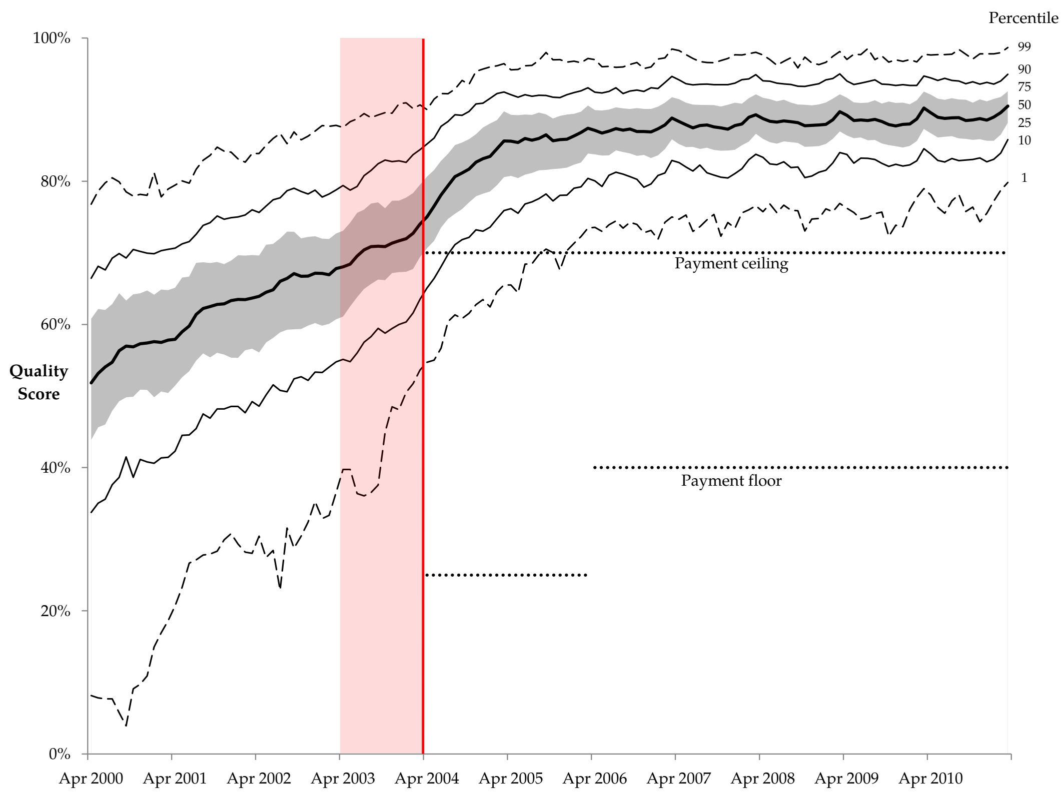


Figure A.3.1: Distribution of practice scores on indicator CHD 6: The percentage of patients with CHD, in whom the last blood pressure reading (in last 15 months) is 150/90 or less.

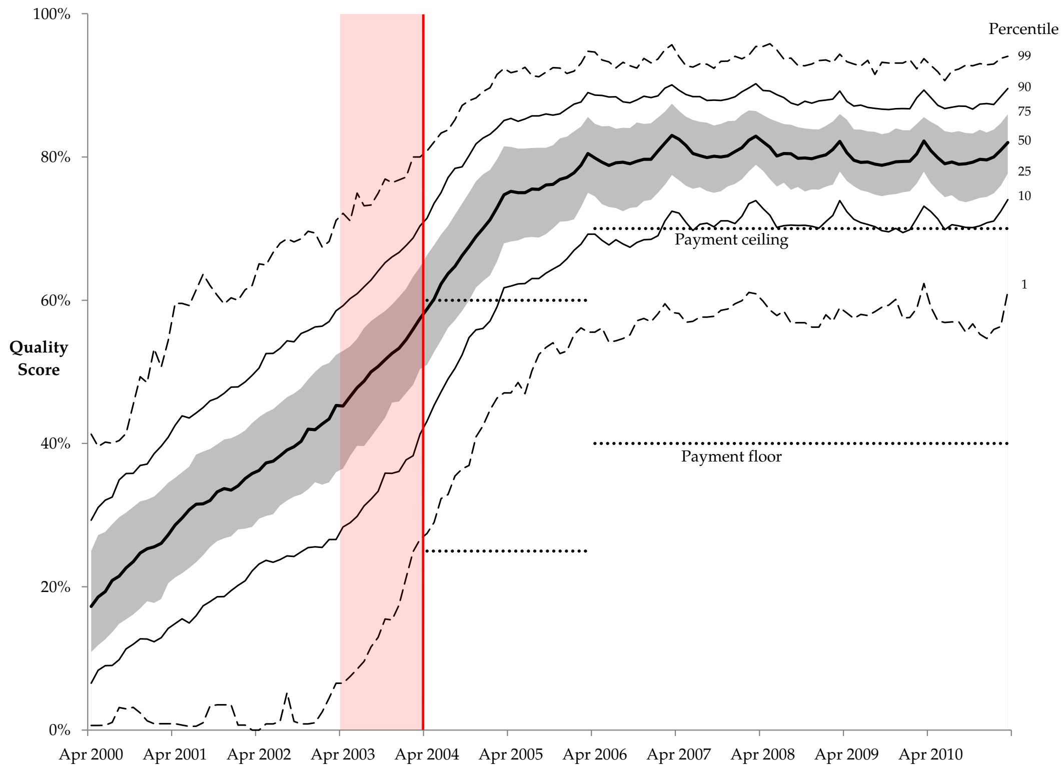


Figure A.3.2: Distribution of practice scores on indicator CHD 8: The percentage of patients with CHD whose last measured total cholesterol (in last 15 months) is 5 mmol/l or less.

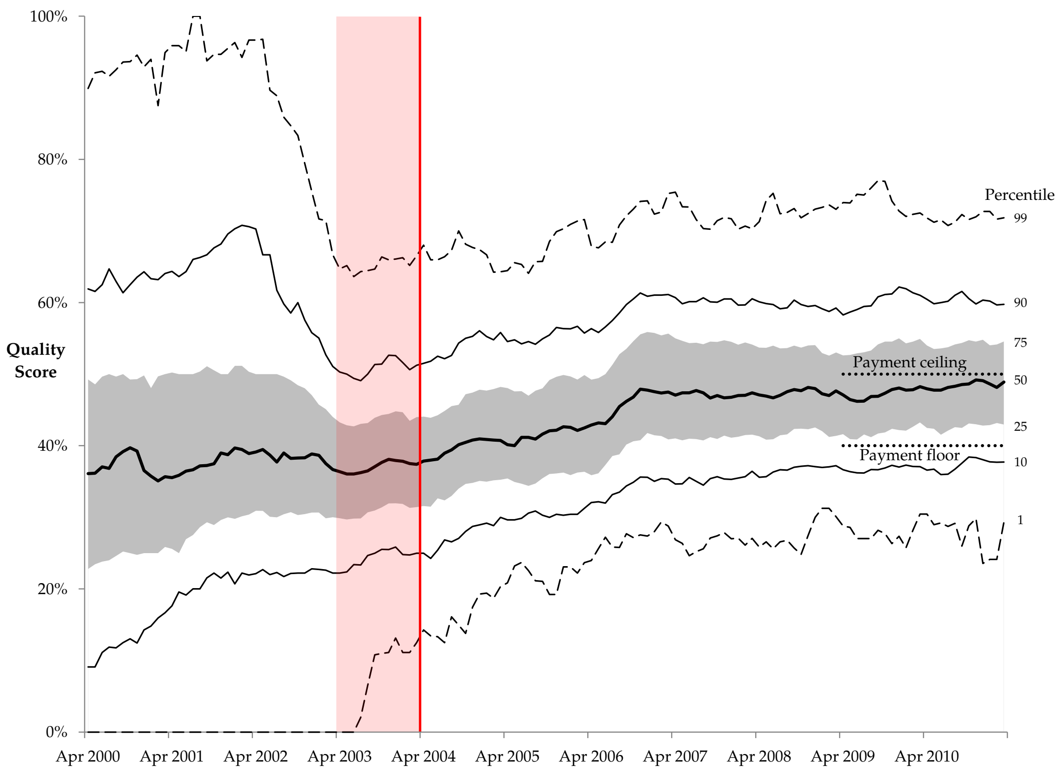


Figure A.3.3: Distribution of practice scores on indicator DM 23: The percentage of patients with diabates in whom the last HbA1c is 7 or less in last 15 months.

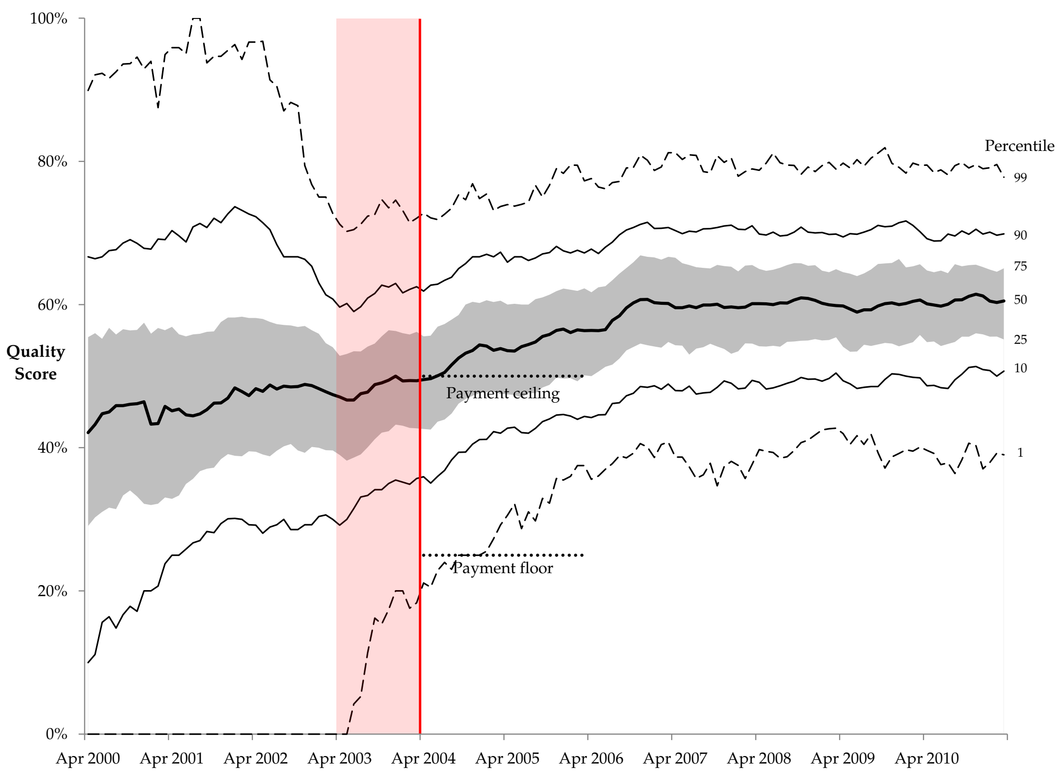


Figure A.3.4: Distribution of practice scores on indicator DM 6: The percentage of patients with diabates in whom the last HbA1c is 7.4 or less in last 15 months.

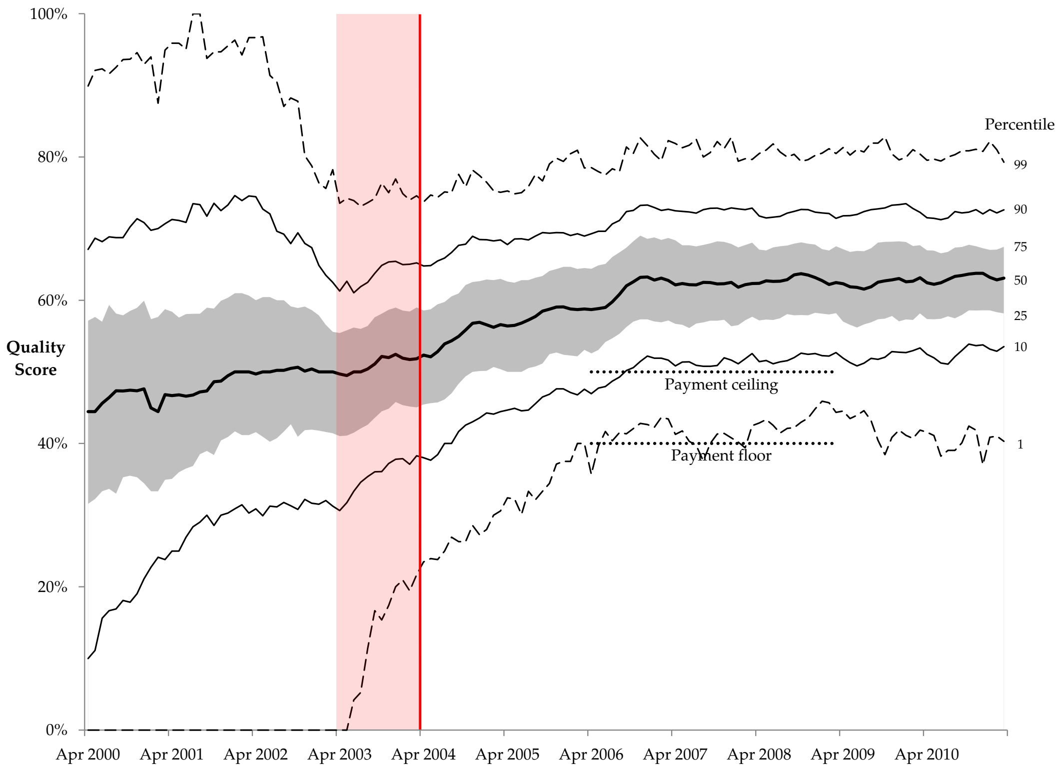


Figure A.3.5: Distribution of practice scores on indicator DM 20: The percentage of patients with diabates in whom the last HbA1c is 7.5 or less in last 15 months.

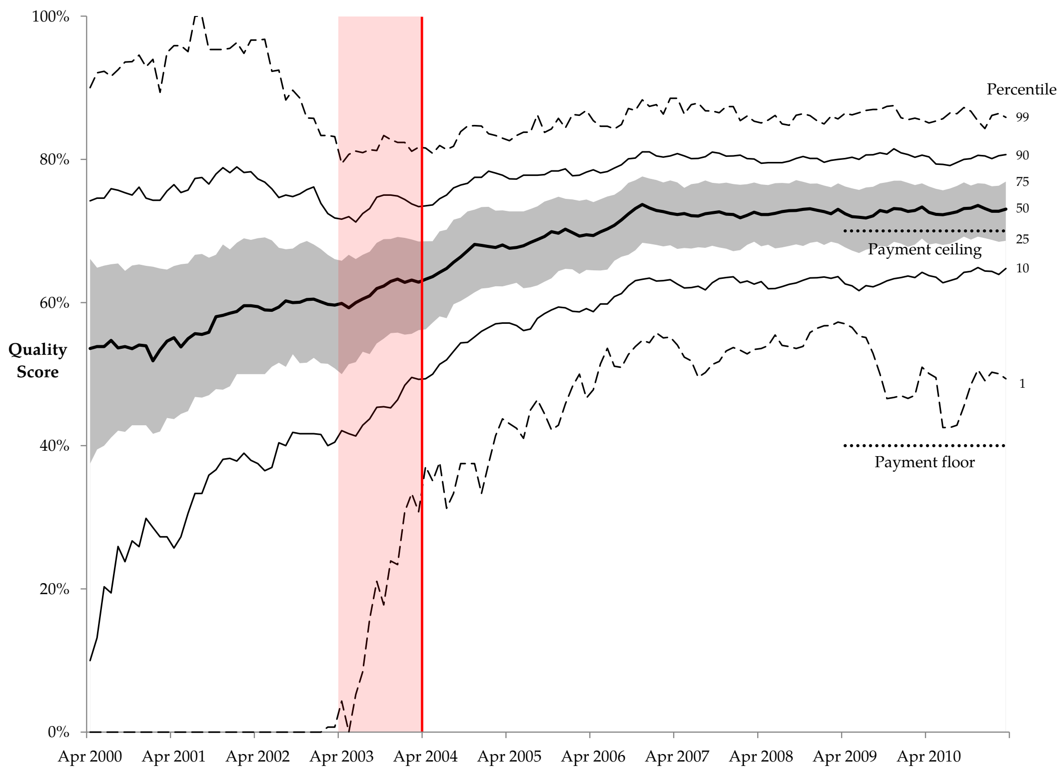


Figure A.3.6: Distribution of practice scores on indicator DM 24: The percentage of patients with diabates in whom the last HbA1c is 8 or less in last 15 months.

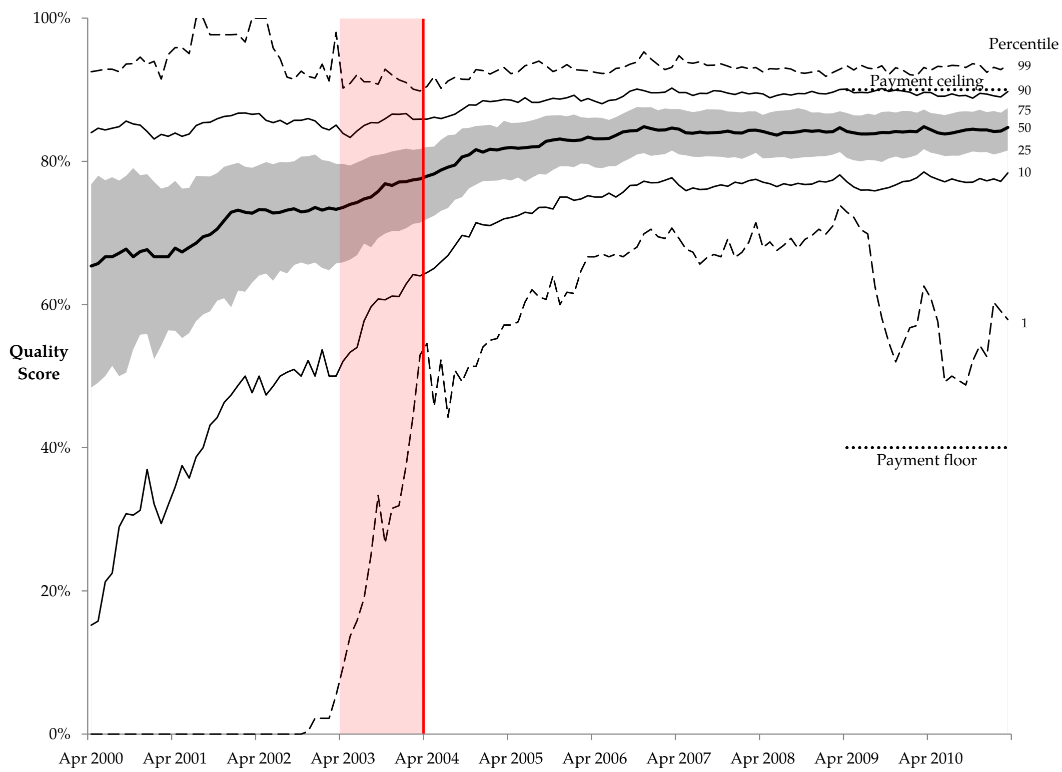


Figure A.3.7: Distribution of practice scores on indicator DM 25: The percentage of patients with diabates in whom the last HbA1c is 9 or less in last 15 months.

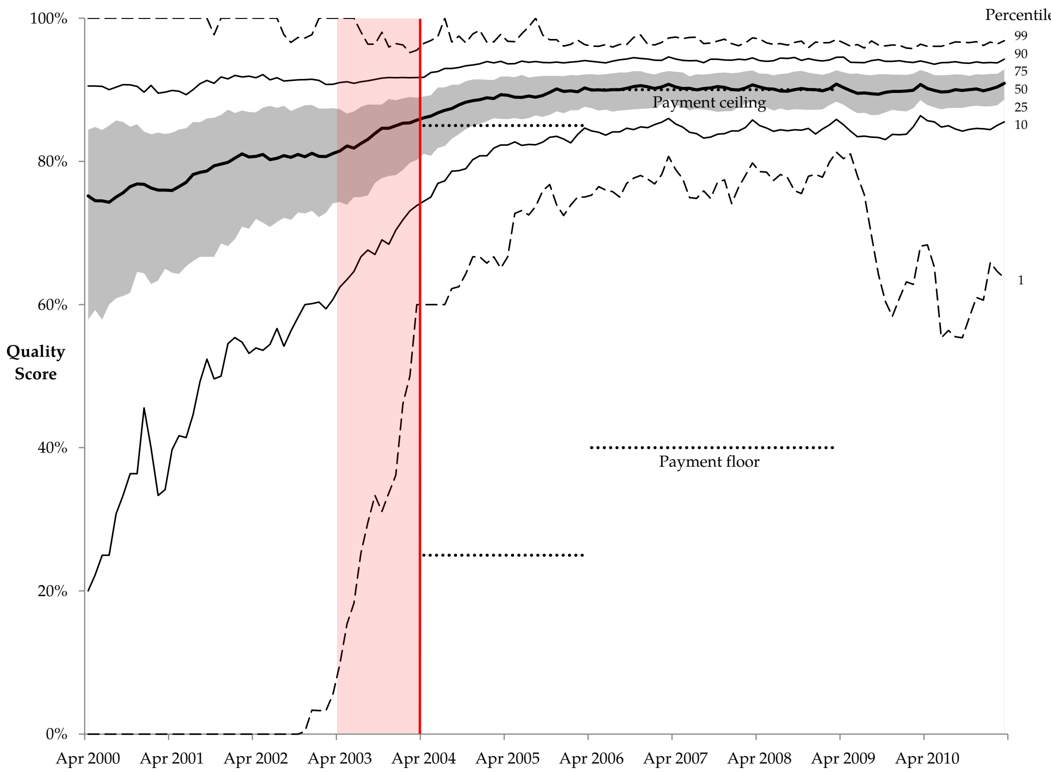


Figure A.3.8: Distribution of practice scores on indicator DM 7: The percentage of patients with diabates in whom the last HbA1c is 10 or less in last 15 months.

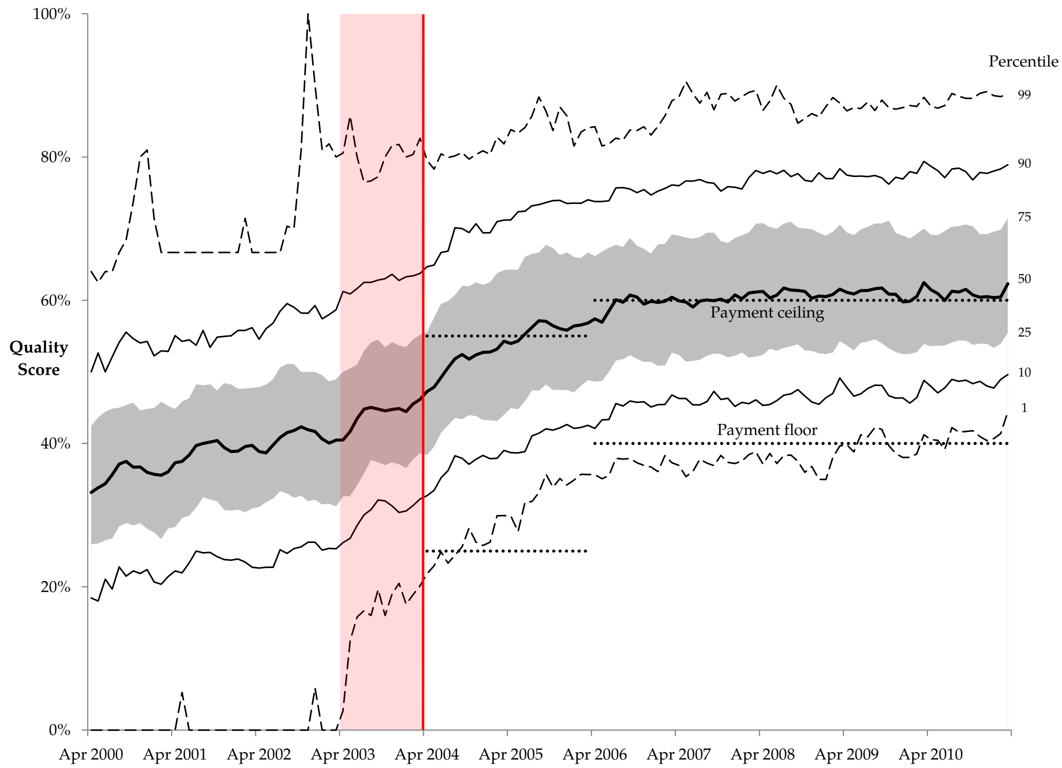


Figure A.3.9: Distribution of practice scores on indicator DM 12: The percentage of patients with diabetes in whom the last blood pressure (in last 15 months) is 145/85 or less.

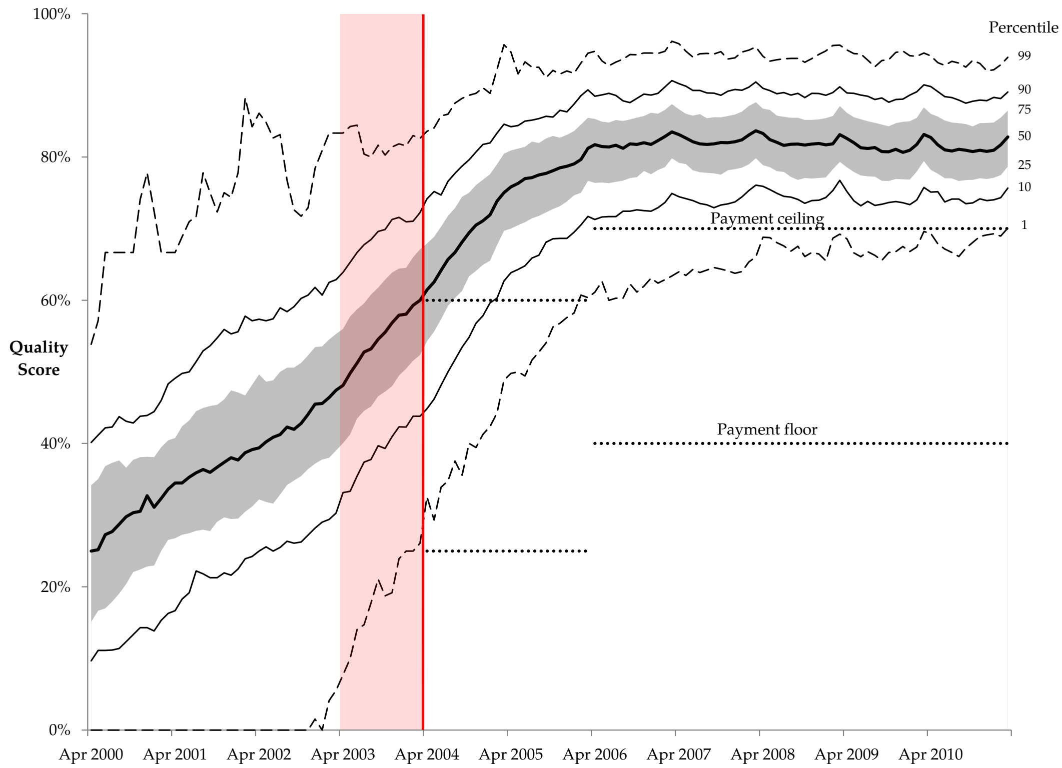


Figure A.3.10: Distribution of practice scores on indicator DM 17: The percentage of patients with diabetes whose last measured total cholesterol (in last 15 months) is 5 mmol/l or less.

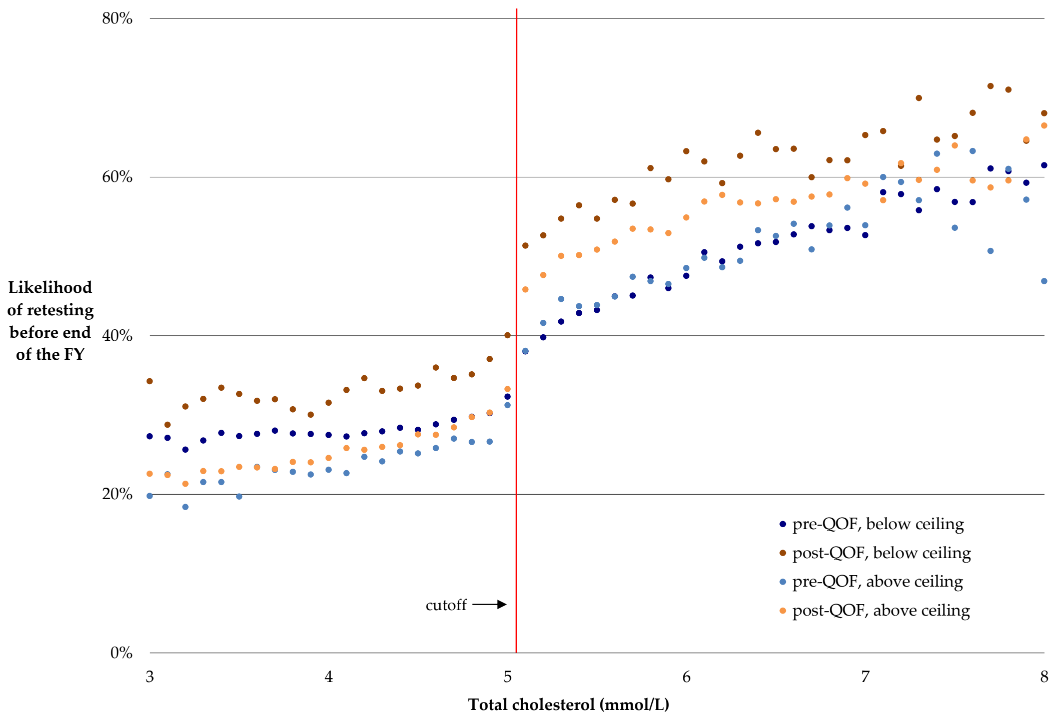


Figure A.3.11: Results from a linear probability model regression predicting the probability of cholesterol retesting before the end of the fiscal year. We control for patient fixed-effects and month-of-year indicators, and drop tests in the last month of the fiscal year. $N = 736,479$ cholesterol tests from $200,867$ unique patients with CHD and/or diabetes, but not with history of stroke or TIA.

Table A.3.1: Regression results for cholesterol retesting, comparing practices above and below payment ceilings

	β	SE
$\beta_1 \cdot postQOF$	0.0134^{*}	(0.00580)
$\beta_2 \cdot chol > 5$	0.167^{***}	(0.00415)
$\beta_3 \cdot postQOF \cdot chol > 5$	0.0627^{***}	(0.00699)
$\beta_4 \cdot over\_ceiling$	0.00159	(0.00569)
$\beta_5 \cdot postQOF \cdot over\_ceiling$	-0.0266^{***}	(0.00772)
$\beta_6 \cdot chol > 5 \cdot over\_ceiling$	0.0393^{***}	(0.00816)
$\beta_7 \cdot postQOF \cdot chol > 5 \cdot over\_ceiling$	-0.0218^{*}	(0.00979)

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

$N = 736,479$ cholesterol tests from $200,867$ unique patients with CHD and/or diabetes, but not with history of stroke or TIA, dropping tests in the last month of the fiscal year. The dependent variable in each regression is the likelihood of the patient having his or her cholesterol retested before the end of the fiscal year. A constant, indicator variables for months of the year, and patient fixed effects are included in the model and their coefficients are suppressed. Standard errors are robust to heteroskedasticity and clustering at the practice level.

Colophon

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 10 point TeXGyre Pagella, designed by members of the GUST e-foundry in the style of Palatino, by Hermann Zapf. Equations are set in 10 point Pazo Math, by Diego Puga. A template that can be used to format a PhD thesis with this look and feel is freely available online at https://github.com/suchow/.