

Integrating Pay-for-Performance into Health Care Payment Systems

Samuel S. Richardson*

January 13, 2012

Abstract

Health care payers have implemented pay-for-performance contracts without explicit characterization of the underlying quality problem or an understanding of the interaction between pay-for-performance and existing payment mechanisms. In this paper, inefficient quality of health care services stems from difficulty consumers have in observing some aspects of quality. Traditional health care payment systems result in under-provision of those attributes that are poorly observed by consumers, but not those that are well-observed. Within a model of provider competition on two dimensions of quality, I show that the efficient pay-for-performance contract corrects the market failure within the existing payment system, and rewards those dimensions of quality that are under-supplied in the existing system. These will be dimensions that are poorly observed, not necessarily those that are most important for improving patient health. Policymakers often worry about multitasking problems arising from pay-for-performance. I argue that physician allocation of effort to various tasks is inefficient without pay-for-performance, and the optimal design of the pay-for-performance contract can mitigate multitasking.

Keywords: Pay-for-performance, health care quality, multitasking, mechanism design.

JEL Classification Numbers: I18.

*PhD Candidate, Harvard University PhD Program in Health Policy. E-mail: *richard5@fas.harvard.edu*. I am grateful to Tom McGuire, Meredith Rosenthal, Amitabh Chandra, Christopher Avery, Julian Jamison, and Martin Anderson for advice and suggestions. I am also grateful to seminar participants at the Harvard PhD in Health Policy Research Seminar for helpful comments.

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 [Seddon et al., 2001, McGlynn et al., 2003]. 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]. It would be difficult to argue that consumers do not observe their satisfaction with the health care they consume, so patient-centered measures of quality should not be the focus of pay-for-performance contracts, assuming consumers are able to choose their providers.

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

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

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) 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

²We 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 benefits across consumers, so facing informed consumers, the providers would choose efficient quality levels.

³Note 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, with near-universal adoption of electronic medical records among primary care practices in

Once we go beyond one-dimensional 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.

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

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].

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 [Rosenthal et al., 2005, Mullen et al., 2010]. 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

⁴It 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.

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 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.

⁵Here 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].

⁶In 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 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.

2. Profit-maximizing⁷ 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 $c(q_{1j}, q_{2j})$, 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 $c(q_{1j}, q_{2j})$ 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 receives utility $u(q_{1b}, q_{2b}, i) = q_{1b} + q_{2b} - t \cdot (1 - i)$. 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 $\mu_j(q_{1j}, q_{2j}, q_{1,-j}, q_{2,-j})$. Profit to provider j is:

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

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

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

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.

⁷Most 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.

4.1 Demand

First, I derive demand for provider j as a function of $(q_{1a}, q_{2a}, q_{1b}, q_{2b})$. 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 $(\mathbf{1}[\text{logical expression}]$ is an indicator function that takes the value 1 if *logical expression* is true and 0 otherwise):

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

Assuming that absent additional information, consumers believe the two providers choose the same quality vectors,⁸ this reduces to:

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

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

⁸Note 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.

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

4.2 Provider profit-maximization

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

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

First-order conditions for a maximum are given by:

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

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

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) must hold with equality in equilibrium.⁹ 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 1. *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 each quality dimension: $\frac{\partial \mu}{\partial q_k} = \frac{\lambda_k}{2t}$. Substituting into our first-order conditions yields:

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

⁹Later 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 $c_2(x, 0) = 0$.

Second-order conditions are given by:

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

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.

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:

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

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.

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:

$$\begin{aligned} \max_{p_b, p_c} SW &= q_1 + q_2 - c(q_1, q_2) \\ \text{s.t. } FOC_1 &: (1 - p_c) \cdot c_1 = \frac{\lambda_1}{t} [p_b - (1 - p_c) c] \\ FOC_2 &: (1 - p_c) \cdot c_2 = \frac{\lambda_2}{t} [p_b - (1 - p_c) c] \\ IR &: p_b \geq (1 - p_c) c \end{aligned}$$

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; however, bundled payment is not able 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.

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:

$$\begin{aligned} \max_{p_1} SW &= q_1 + q_2 - c(q_1, q_2) \\ \text{s.t. } FOC_1 &: c_1 - p_1 = \frac{\lambda_1}{t} [p_1 q_1 - c] \\ FOC_2 &: c_2 = \frac{\lambda_2}{t} [p_1 q_1 - c] \\ IR &: p_1 q_1 \geq c \end{aligned}$$

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 .¹⁰

Compared to the results in subsection 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 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 .

¹⁰Note 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^* .

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.¹¹ This can be modeled as bundled payment, plus lower p_1 than was the case in the traditional Medicare payment model.¹²

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:

$$\begin{aligned} \max_{p_b, p_1} SW &= q_1 + q_2 - c(q_1, q_2) \\ \text{s.t. } FOC_1 &: c_1 - p_1 = \frac{\lambda_1}{t} (p_b + p_1 q_1 - c) \\ FOC_2 &: c_2 = \frac{\lambda_2}{t} (p_b + p_1 q_1 - c) \\ IR &: p_b + p_1 q_1 \geq c \end{aligned}$$

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

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

¹¹This 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 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.

¹²Note 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.

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 4.4). 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 $p_1 < 1$: the marginal payment for quality is strictly less than the marginal benefit to patients. 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.

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 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.¹³ 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.

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 ζ sick types and $(1 - \zeta)$ healthy types; $0 < \zeta < 1$. 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 $c(q_1, q_2)$, and the cost per healthy patient is $c(q_1, 0)$. 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.¹⁴

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/skimming 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

¹³When 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.

¹⁴In 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.

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 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 [2002].

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).

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).¹⁵

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.¹⁶ This follows from the equation for the provider's marginal profit associated with changes in quality dimension k :

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

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 p_b towards negative infinity.

In section 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

¹⁵In this light, it is curious that medical education is subsidized in most developed countries, rather than being taxed.

¹⁶Setting $\frac{1-p_1}{1-p_2} = \frac{\lambda_1}{\lambda_2}$ 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.

(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.

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.

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.

The insights from my model can be applied beyond multitasking problems arising from failures of demand response to quality. In section 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”.

References

- S. M. Campbell, D. Reeves, E. Kontopantelis, B. Sibbald, and M. Roland. Effects of pay for performance on the quality of primary care in England. *N Engl J Med*, 361(4):368–78, 2009.
- T. Doran, C. Fullwood, H. Gravelle, D. Reeves, E. Kontopantelis, U. Hiroeh, and M. Roland. Pay-for-performance programs in family practices in the United Kingdom. *New England Journal of Medicine*, 355(4):375–84, 2006.
- T. Doran, E. Kontopantelis, J. M. Valderas, S. Campbell, M. Roland, C. Salisbury, and D. 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:d3590, 2011.
- K. Eggleston. Multitasking and mixed systems for provider payment. *Journal of Health Economics*, 24(1):211–23, 2005.
- R. P. Ellis. Creaming, skimping and dumping: provider competition on the intensive and extensive margins. *Journal of Health Economics*, 17(5):537–555, 1998.
- J. Glazer and T. G. McGuire. Setting health plan premiums to ensure efficient quality in health care: Minimum variance optimal risk adjustment. *Journal of Public Economics*, 84:153–173, 2002.
- B. R. Golden and F. A. Sloan. Physician pay for performance. In F. A. Sloan and H. Kasper, editors, *Incentives and Choice in Health Care*, pages 289–317. MIT Press, Cambridge, MA, 2008.
- B. Holmstrom and P. Milgrom. Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization*, 7:24–52, 1991.
- D. 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.
- O. Kaarboe and L. Siciliani. Multi-tasking, quality and pay for performance. *Health Economics*, 20(2):225–238, 2011.

- C-t. A. Ma. Health care payment systems: Cost and quality incentives. *Journal of Economics & Management Strategy*, 3(1):93–112, 1994.
- C-t. A. Ma and T. G. McGuire. Optimal health insurance and provider payment. *American Economic Review*, 87(4):685–704, 1997.
- M. McClellan, A. N. McKethan, J. L. Lewis, J. Roski, and E. S. Fisher. A national strategy to put accountable care into practice. *Health Affairs (Millwood)*, 29(5):982–90, 2010.
- E. A. McGlynn, S. M. Asch, J. Adams, J. Keeseey, J. Hicks, A. DeCristofaro, and E. A. Kerr. The quality of health care delivered to adults in the United States. *New England Journal of Medicine*, 348(26):2635–45, 2003.
- T. G. McGuire. Physician agency. In A.J. Culyer and J.P. Newhouse, editors, *Handbook of Health Economics*, pages 461–536. Elsevier, North-Holland, 2000.
- K. J. Mullen, R. G. Frank, and M. 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.
- M. Roland. Linking physicians’ pay to the quality of care: A major experiment in the United Kingdom. *New England Journal of Medicine*, 351(14):1448–1454, 2004.
- M. B. Rosenthal and R. G. Frank. What is the empirical basis for paying for quality in health care? *Med Care Res Rev*, 63(2):135–57, 2006.
- M. B. Rosenthal, R. G. Frank, Z. Li, and A. M. Epstein. Early experience with pay-for-performance: From concept to practice. *JAMA*, 294(14):1788–1793, 2005.
- 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.
- B. Serumaga, D. Ross-Degnan, A. J. Avery, R. A. Elliott, S. R. Majumdar, F. Zhang, and S. B. Soumerai. Effect of pay for performance on the management and outcomes of hypertension in the United Kingdom: interrupted time series study. *BMJ*, 342:d108, 2011.
- A. M. Spence. Monopoly, quality, and regulation. *The Bell Journal of Economics*, 6(2):417–429, 1975.
- A. Tay. Assessing competition in hospital care markets: The importance of accounting for quality differentiation. *The RAND Journal of Economics*, 34(4):786–814, 2003.

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

B. 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.

Appendix: Proofs

Proposition 1. *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 : $A\pi_j(q_{1j}, q_{2j}) = p_1q_{1j} + p_2q_{2j} + p_b - (1 - p_c) \cdot c(q_{1j}, q_{2j})$. 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 $(q_{1,-j}, q_{2,-j})$. 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 $(q_{1a}, q_{2a}) \neq (q_{1b}, q_{2b})$, $DA\pi_a(q_{1a}, q_{2a}) \cdot \mathbf{v} \neq DA\pi_b(q_{1b}, q_{2b}) \cdot \mathbf{v}$.

Define $D_+F(\mathbf{x}) \cdot \mathbf{v} = \lim_{h \rightarrow 0^+} \frac{F(\mathbf{x}+h\mathbf{v})-F(\mathbf{x})}{h}$ and $D_-F(\mathbf{x}) \cdot \mathbf{v} = \lim_{h \rightarrow 0^-} \frac{F(\mathbf{x}+h\mathbf{v})-F(\mathbf{x})}{h}$: 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 $\frac{\partial_- \mu_a}{\partial q_{ka}} = \frac{\partial_+ \mu_b}{\partial q_{kb}}$, $D_+\mu_a(q_{1a}, q_{2a}; q_{1b}q_{2b}) \cdot \mathbf{v} = D_-\mu_b(q_{1b}, q_{2b}; q_{1a}q_{2a}) \cdot \mathbf{v} \equiv D_+\mu_a$ and $D_-\mu_a(q_{1a}, q_{2a}; q_{1b}q_{2b}) \cdot \mathbf{v} = D_+\mu_b(q_{1b}, q_{2b}; q_{1a}q_{2a}) \cdot \mathbf{v} \equiv D_-\mu_a$.

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

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

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

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

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

$$\mu_a \cdot [DA\pi_a(q_{1a}, q_{2a}) \cdot \mathbf{v}] + A\pi_a(D\mu \cdot \mathbf{v}) = 0$$

$$\mu_b \cdot [DA\pi_b(q_{1b}, q_{2b}) \cdot \mathbf{v}] + A\pi_b(D\mu \cdot \mathbf{v}) = 0$$

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

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