Information Disclosure and the Equivalence of Prospective Payment and Cost Reimbursement

Ching-to Albert Ma
Department of Economics
Boston University
270 Bay State Road
Boston, MA 02215
USA
ma@bu.edu

Henry Y. Mak
Department of Economics
Indiana University-Purdue University Indianapolis
425 University Boulevard
Indianapolis, IN 46202
USA
makh@iupui.edu

May 2015

Abstract

A health care provider chooses unobservable service-quality and cost-reduction efforts. The efforts produce quality and cost efficiency. An insurer observes quality and cost, and chooses how to disclose this information to consumers. The insurer also decides how to pay the provider. In prospective payment, the insurer fully discloses quality, and sets a prospective payment price. In cost reimbursement, the insurer discloses a value index, a weighted average of quality and cost efficiency, and pays a margin above cost. The first-best quality and cost efforts can be implemented by prospective payment and by cost reimbursement. Cost reimbursemnt with value index eliminates dumping and cream skimming. Prospective payment with quality index eliminates cream skimming.

Keywords: information disclosure, prospective payment, cost reimbursement, quality, cost reduction, dumping, cream skimming

Acknowledgement: We have had many inspiring conversations with Mike Luca. For their comments, we thank Chiara Canta, Randy Ellis, Iris Kesternich, Ting Liu, Tom McGuire, Sara Machado, Inés Macho Stadler, Pau Olivella, Alex Poterack, Alyson Price, Jim Rebitzer, Sam Richardson, Tomàs Rodríguez Barraquer, Gilad Sorek, and seminar participants at Boston University, European University Institute, Frisch Centre at the University of Oslo, Harvard University, Lehigh University, University of Salerno, the 13th European Health Economic Workshop, and the 13th International Industrial Organization Conference. We thank coeditor Tom Gresik for discussing with us many details of this research, and an associate editor and three reviewers for their suggestions. The second author received financial and research support from the Max Weber Programme at European University Institute.

1 Introduction

The (provocative) title refers to prospective payment and cost reimbursement, the most common mechanisms for paying health care providers. In prospective payment, a provider receives a fixed price for delivering a medical service, irrespective of resources used. In cost reimbursement, a provider receives a revenue corresponding to resources used.[1] These two payment methods have been studied extensively and intensively in the past thirty years. The conventional wisdom is that prospective payment and cost reimbursement give rise to different quality and cost incentives. In this paper, we describe a model in which prospective payment and cost reimbursement can give rise to identical quality and cost incentives. This model differs from the conventional one only in how consumers learn about quality.

The canonical model is this. A health care provider chooses unobservable quality and cost-reduction efforts, and incurs disutilities in doing so. The efforts produce quality and reduce costs. A higher quality results in a higher variable cost and attracts more consumers, but a higher cost effort reduces the variable cost. An insurer wants to implement socially efficient quality and cost efforts.

Under prospective payment, the provider internalizes the production cost, so its cost-reduction incentive is aligned with social cost efficiency. An appropriate prospective payment level may then be chosen to align the provider's profit motive with social quality efficiency. Prospective payment kills two birds with one stone. Cost reimbursement works in a perverse way. Because all variable costs will be reimbursed, the provider lacks any incentive to expend cost effort. The quality incentive can still be implemented by paying the provider a margin above cost for services rendered. The provider raises quality to attract more consumers because of the profitable margin.

In the two payment systems, the common principle is demand response: higher quality raises demand, so a higher profit margin incentivizes quality effort. However, the provider internalizes costs under prospective payment, but does not do so under cost reimbursement.

[1] For our purpose, cost reimbursement is the same as conventional fee-for-service: a provider chooses medical services to supply, and receives a fee that amounts to the cost and a profit margin. Prospective payment may be supplemented by outlier compensations, local-market adjustments, etc. These variations are unimportant here.

A demand response requires consumers to know about quality. However, health care quality information can be difficult to obtain and interpret. Indeed, insurers, governments and sponsors increasingly have helped consumers find out about quality.[2] In this paper, we make an alternative assumption about information structure. We assume that consumers cannot observe quality directly, but the insurer can. The insurer can also observe costs. We set up an implementation problem; the insurer would like the provider to choose first-best quality and cost efforts, which are hidden actions, by information disclosure and payment incentives.

We prove two main results. First, first-best efforts can be implemented by prospective payment and full disclosure of quality, so we reaffirm a result of the canonical model. Second, and this is the surprise, first-best efforts can be implemented by cost reimbursement and partial disclosure of quality and cost. Partial information disclosure refers to a *value index*. A provider's unobservable efforts produce quality and cost efficiency (cost saving from a benchmark). For any quality and cost produced, the insurer constructs a weighted average and discloses this average—the value index—to consumers. We show that mixing quality and cost efficiency information can incentivize cost effort.

Why is there cost incentive under cost reimbursement when a value index about quality and cost is disclosed to consumers? Consumers only observe the value index, not quality, so they will draw inference about quality based on the value index. A given level of value index corresponds to some inferred quality level, which generates a demand. Consumers' belief about quality is based on the value index, not the actual quality effort. Hence, changing efforts that would maintain the index would leave demand (and revenue) unaffected. It follows that the provider must choose disutility-minimizing efforts to achieve an index.[3] Furthermore, the insurer can choose the index weight and profit margin to make the provider internalize the net social benefit of quality and cost efforts.

Starting with the basic model, we then consider more complex environments. In one extension, we consider dumping of high-cost consumers. Under prospective payment, the provider takes a loss when

[2] For a summary of empirical works on public reporting initiatives, see Dranove and Jin (2011).

[3] An "agency" explanation in line with the Mirrless-Holmstrom model goes as follows. An agent (the provider) chooses unobservable inputs (efforts) that produce two outputs (quality and cost efficiency). Consumer demand is based on one output (quality), but consumers observe nothing. The principal (the insurer) observes the two outputs, and (credibly) reports to consumers a weighted average. Belief on quality output depends only on the index. The agent's equilibrium efforts must minimize the disutility for achieving the index.

treating high-cost consumers whose costs are higher than the price, so will refuse to serve them. We show that dumping can be avoided under cost reimbursement, because cost variations will be absorbed by the insurer. Implementation of first-best efforts is possible under cost reimbursement, but not under prospective payment.

In another extension, we study cream skimming when health services have multiple qualities. Cream skimming refers to the overprovision of more profitable qualities and the underprovision of less profitable qualities. We illustrate how prospective payment and full disclosure create cream skimming incentives. We then show that under both prospective payment and cost reimbursement, the insurer can use partial disclosure to neutralize the provider's cream skimming incentives.[4]

It has not escaped our notice that our theory relies on the provider being unable to credibly disclose quality information. If a provider were able to do so, it could defeat the value-index manipulation. In practice, there does not seem to be any "danger" that any provider could fully disclose quality information. Otherwise, public agencies (such as *the Centers for Medicare and Medicaid Services*) and nonprofit organizations (such as *Consumer Reports* and *the National Committee for Quality Assurance*) would not have expended huge resources to make quality reports available to the general public. Furthermore, it is far from clear that a provider would honestly report quality information even when it was feasible to do so.

1.1 Literature

The literature on provider payment design is large. For surveys, see Newhouse (1996), McGuire (2000), and Leger (2008). Ma (1994) lays out the basic model of payment systems and their effects on health care quality and cost incentives. The general consensus is that cost reimbursement fails to achieve cost efficiency, and that prospective payment leads to perverse selection incentives such as dumping and cream skimming. Generally, neither cost reimbursement nor prospective payment achieves socially efficient outcomes.

We assume a demand response: consumers' demand for services reacts positively to quality, an assumption

[4] Prospective payment also encourages "fraudulent" upcoding. For example, *Medicare* uses the Diagnostic Related Group system to set prices. If an illness fits into more than one diagnosis (perhaps due to severity differences), a provider may choose to report the one with a higher price (Dafny, 2005).

commonly adopted in the literature: see for example, Rogerson (1994), Ma and McGuire (1997), Frank et al. (2000), Glazer and McGuire (2000), Brekke et al. (2006).[5] Recent papers empirically evaluate demand response to public reports. In commercial health-plan markets, both Beaulieu (2002) and Scanlon et al. (2002) show that consumers do avoid health plans with low ratings. Since 1999, *the Centers for Medicare and Medicaid Services* has launched quality-report initiatives for health plans, hospitals, physicians, and nursing homes (see www.cms.gov/QualityInitiativesGenInfo/). Dafny and Dranove (2008) find that the reports for Medicare health plans substantially affect enrollments.

Our paper is closely related to a small but growing literature on optimal public-report design. Glazer and McGuire (2006) propose a disclosure policy that achieves cross subsidies among *ex ante* heterogenous consumers to solve an adverse selection problem in a competitive market. Ma and Mak (2014a) characterize the optimal average-quality reports that mitigate monopoly price discrimination and quality distortion. The current paper contributes to the literature by simultaneously studying optimal payment and reporting policies in a hidden-action framework.

Information asymmetry has long been viewed as a source of inefficiency in the physician-patient interaction literature. For example, in both Dranove (1988) and Rochaix (1989), a physician utilizes his private information to induce patient demand for excessive treatments. By contrast, the insurer in our model holds back some information from consumers to induce cost-reduction effort.

Information disclosure has been extensively studied in the industrial organization literature. In Matthews and Postlewaite (1985) and Schlee (1996), product quality is unknown to the seller, consumers, or both. They show that quality information can harm consumers because of the seller's price response. Instead, we focus on how a trusted intermediary can utilize demand response to discipline a seller. In both Lizzeri (1999) and Albano and Lizzeri (2001), a profit-maximizing intermediary privately observes product quality. They show that the intermediary may underprovide quality information at the expense of market efficiency. However, the insurer in our model withholds information to achieve efficient quality and cost effort.

The rest of the paper is organized as follows. Section 2 presents the model. Section 3 sets up the

[5] One exception is Chalkley and Malcomson (1998). In their model, a capacity-constrained provider is motivated by altruism rather than demand response.

information structure, the extensive forms, and studies equilibria. We first study prospective payment, and then turn to cost reimbursement and value index. Section 4 presents three extensions. We allow for stochastic production of quality by means of a standard hidden-action model. Then we consider stochastic cost reduction and study dumping. Finally, we let the provider produce many qualities and study cream skimming. Section 5 draws some conclusions. The Appendix collects proofs, and an example is worked out in the Supplement.

2 Model

2.1 Consumers and a provider

A set of consumers is covered by an insurer. Health services are to be supplied by a provider. If consumers believe that health care quality is q, the quantity demanded is $D(q)$, which is strictly increasing and concave. The demand for health services also depends on copayments, deductibles, coinsurance rates, or their combinations. We let consumer cost-share parameters be given, so the demand function D already incorporates consumer cost shares. This makes for simpler notation because we are concerned with incentives for providers.[6] The social benefit from quality q is denoted by $B(q)$ which is strictly increasing and concave. In many applications B is consumer benefit from services, but we allow a more general interpretation so that externalities, equity, and any other such issues can be included.

A provider supplies health services to insured consumers. Its actions affect health care quality and cost efficiency. We call these actions *quality effort*, and *cost effort*, denoted by the nonnegative variables e and r, respectively. Quality and cost efforts are unobservable. Quality depends on effort e. In this and the next section, we assume deterministic quality production from effort e, so write quality q as a function of effort $q(e)$. We assume that q is strictly increasing and concave.[7] Later, in Subsection 4.1, we use a standard hidden-action model for stochastic quality production: we let effort e determine a distribution of possible

[6] We also abstract from strategic interaction among providers. This issue is addressed in Ma and Mak (2014b). There we show that when heterogeneous providers compete for consumers in a health care network, first-best implementation requires the insurer to coordinate disclosure, copayment, and provider payment policies.

[7] The inverse of the function q yields the effort that is used to achieve a quality. However, in this work, we only allow payments to be based on quantities.

qualities, as in Holmstrom (1979).

The unit cost for service is $C(e, r)$ given quality effort e and cost effort r. The function C is strictly increasing in e and strictly decreasing in r, and strictly convex. More effort on care quality requires a higher unit cost, but cost-reduction effort can reduce it. In addition, the provider incurs a fixed cost or disutility due to efforts, denoted by $\Lambda(e, r)$. The function Λ is strictly increasing and strictly convex. We assume that efforts are to be chosen from a (nonnegative) bounded set, and that equilibrium effort choices must be interior.[8] If the demand is $D(q(e))$, the provider incurs a total cost $D(q(e))C(e, r) + \Lambda(e, r)$.

2.2 Payment and information mechanisms

The quantity of services is observed *ex post* and payment can be based on it. The unit cost of services $C(e, r)$ is also observed *ex post*, and again payment can be based on it. Quality-cost effort disutilities are unobservable. We study the conventional payment systems: prospective payment and cost reimbursement, which currently still account for most providers' revenue.[9] We consider the use of information about quality and cost as an incentive instrument to supplement the conventional systems. The study of other systems such as pay-for-performance and valued-based purchases is left to other research.[10]

Under prospective payment, the provider receives a fixed price p per unit of delivered service. If the provider has satisfied a demand of $D(q(e))$, its revenue is $pD(q(e))$, and it bears the total cost $D(q(e)) \times C(e, r) + \Lambda(e, r)$. Under cost reimbursement, for each unit of delivered services the provider will be paid the variable cost $C(e, r)$ plus a margin m. If the provider has satisfied a demand $D(q(e))$, its revenue, net of variable cost, is $mD(q(e))$, and it only bears the disutility $\Lambda(e, r)$. Prospective payment p and the margin m are nonnegative. The provider's disutility due to effort, $\Lambda(e, r)$, cannot be observed and directly compensated

[8] In other words, we impose the common Inada conditions. Using subscripts to denote partial derivatives of the corresponding variables, we assume i) $C_1(e, r) \to 0$ and $\Lambda_1(e, r) \to 0$, as $e \to 0$; $C_2(e, r) \to -\infty$ and $\Lambda_2(e, r) \to 0$ as $r \to 0$, and ii) $C_1(e, r) \to \infty$ and $\Lambda_1(e, r) \to \infty$ as e approaches its upper bound; $C_2(e, r) \to 0$ and $\Lambda_2(e, r) \to \infty$ as r approaches its upper bound.

[9] In 2009, 79% of employees covered by employer-provided health plans received benefits under fee-for-service arrangements (Bureau of Labor Statistics, 2011). In 2012, 73% of Medicare enrollees were covered by fee-for-service plans (Centers for Medicare & Medicaid Services, 2013). The Centers use prospective payment to reimburse hospital services and a fixed fee schedule to reimburse physician services.

[10] In 2014, prospective payment and cost reimbursement accounted for 60% of commercial in-network payments (Catalyst for Payment Reform, 2014).

for. The provider may also receive a lump-sum payment, which can be positive or negative.

Our departure from the standard payment-design problem is on the information about quality. In the literature, consumers are assumed to observe quality. Here, consumers are unable to observe quality, and rely on the insurer to act as a trusted information intermediary. Although both quality and cost efforts are unobservable, the insurer can observe the provider's care quality q and variable cost $C(e, r)$. The insurer may disclose information fully, or choose to disclose an index, constructed as follows. First, we posit that there is a ceiling K so that the variable cost $C(e, r)$ is at most K. Given efforts, $K - C(e, r)$ is a measure of cost efficiency. We define a *value index* by $I(q, C; \theta) \equiv \theta q(e) + (1 - \theta)[K - C(e, r)]$, where $0 \leq \theta \leq 1$. After observing the provider's care quality $q(e)$ and variable cost $C(e, r)$, the insurer reports the value index to consumers.

If we set the weight of the value index θ to 1, then full quality information will be revealed to consumers. If θ is always set to 1, consumers observe the provider's quality choice and respond by demanding health care; this would be the standard model. The point of our paper, however, is that the weight should be set below 1 under cost reimbursement.

2.3 First best

In the first best, quality and cost efforts are contractible. The social welfare from the quality-cost effort pair (e, r) is

$$B(q(e)) - D(q(e))C(e, r) - \Lambda(e, r), \tag{1}$$

where B is social benefit. Let (e^*, r^*) be the quality-cost effort pair that maximizes social welfare in (1), which is assumed to be strictly quasi-concave in efforts. The following first-order conditions characterize the first best:

$$\begin{aligned} B'(q(e^*))q'(e^*) - D'(q(e^*))q'(e^*)C(e^*, r^*) - D(q(e^*))C_1(e^*, r^*) - \Lambda_1(e^*, r^*) &= 0 \qquad (2) \\ -D(q(e^*))C_2(e^*, r^*) - \Lambda_2(e^*, r^*) &= 0, \qquad (3) \end{aligned}$$

where we use the (numeral) subscript of a function to denote the corresponding partial derivative, and the superscript prime to denote derivatives. The first-order conditions have the standard interpretations. Raising quality effort increases social benefit, but it also raises demand, unit cost, and disutility. Raising cost effort reduces unit cost but raises disutility. The first-order conditions in (2) and (3) balance these effects.

3 Payment systems and implementation

3.1 Prospective payment and first best

We let the insurer be a public agency. The insurer's objective is to maximize a weighted sum of social net benefit and the provider's profit, with a lower weight on profit.[11] In prospective payment, the provider receives a price p per unit of service, and a transfer T. Suppose that the insurer fully discloses quality q (by setting $\theta = 1$ in the index $I(q, C; \theta)$). When the provider chooses quality and cost efforts, its payoff is

$$T + pD(q(e)) - D(q(e))C(e, r) - \Lambda(e, r). \tag{4}$$

The quality and cost efforts generate a social net benefit

$$B(q(e)) - pD(q(e)) - T, \tag{5}$$

which is the social benefit $B(q(e))$ less payments to the provider.

The insurer's objective is to choose the prospective price p and the transfer T to maximize

$$w[B(q(e)) - pD(q(e)) - T] + (1 - w)[T + pD(q(e)) - D(q(e))C(e, r) - \Lambda(e, r)], \tag{6}$$

where $.5 < w \leq 1$. The provider must make a nonnegative profit, so (4) must be nonnegative. Given that the welfare weight is larger on social net benefit, the optimal transfer T^* will make profit in (4) equal to zero. A choice of p implements the provider's best response in e and r to maximize profit (4). The following proposition is adapted from Ma (1994), and stated with its proof omitted:

Proposition 1 *By choosing* $p^* = \dfrac{B'(q(e^*))}{D'(q(e^*))}$ *and a suitable transfer* T^**, the insurer implements the first-best quality effort* e^* *and cost effort* r^**.*

[11] The transfer will be used to limit the provider's profits when the insurer's objective puts more weight on social net benefit. Otherwise, the transfer would be undefined. This is a common assumption; see, for example, the regulator's objective function (9) on p916 in Baron and Myerson (1982).

The intuition is well documented in the literature. Under prospective payment, the provider fully internalizes the social cost of quality and cost efforts. Its incentive on cost efficiency aligns with the insurer's. By setting the prospective price at the p^* in Proposition 1, the insurer makes the provider internalize the social benefit of quality as well. Any profit from the prospective payment is taxed away by the transfer, so the first best is implemented.

3.2 Cost reimbursement, value index, and first best

We study the following extensive-form game:

Stage 1 The insurer sets the transfer T, the margin m, and the weight θ in the value index, and commits to reimbursing the provider's variable cost.

Stage 2 The provider chooses unobservable quality and cost efforts, respectively, e and r.

Stage 3 The insurer observes the provider's quality q and the variable cost C, and reports the value index $I(q, C; \theta) \equiv \theta q + (1-\theta)[K - C]$ to consumers.

Stage 4 Consumers learn the level of value index I (but not the provider's quality, variable cost, or efforts), and decide on the quantity of services to obtain.

In this game, the insurer's strategy consists of the transfer T, the margin m and the weight θ. The provider's strategy consists of the quality and cost efforts, e and r (both being functions of the insurer's choices in Stage 1). Consumers do not observe the provider's quality effort, and form beliefs about it (as well as cost effort) based on the value index.[12] Given belief on effort, say $\widehat{e}$, demand will be given by $D(q(\widehat{e}))$. We solve for perfect-Bayesian equilibria under a belief restriction.

Suppose that in an equilibrium, the provider chooses quality-cost effort pair $(\widehat{e}, \widehat{r})$. The value index becomes $\widehat{I} \equiv \theta q(\widehat{e}) + (1-\theta)[K - C(\widehat{e}, \widehat{r})]$. Then in equilibrium, consumers must correctly infer from $\widehat{I}$ that quality is $q(\widehat{e})$, and their demand will be $D(q(\widehat{e}))$. What about indexes that are off the equilibrium

[12] Consumer belief of cost effort does not affect demand, but this belief is part of the description of a perfect-Bayesian equilibrium.

path? What should consumers believe when they observe an index different from $\widehat{I}$? We adopt the *wary belief* restriction by McAfee and Schwartz (1994, p221-222).[13] For our game, the restriction says that when consumers observe an index, they believe that the provider has chosen quality and cost efforts optimally to achieve that index. In effect, we draw no distinction between indexes that are on or off the equilibrium path.

Definition 1 (Wary Belief) *A quality-cost effort pair* $(\widetilde{e}, \widetilde{r})$ *is said to satisfy wary belief at index* $\widetilde{I}$ *if 1)* $\theta q(\widetilde{e}) + (1-\theta)[K - C(\widetilde{e}, \widetilde{r})] = \widetilde{I}$*, and 2)*

$$(\widetilde{e}, \widetilde{r}) = \underset{e,r}{\operatorname{argmax}}\, T + mD(q(\widetilde{e})) - \Lambda(e,r) \tag{7}$$

$$\textit{subject to} \quad \theta q(e) + (1-\theta)[K - C(e,r)] = \widetilde{I}. \tag{8}$$

For any index $\widetilde{I}$ and effort pair $(\widetilde{e}, \widetilde{r})$, wary belief requires that indeed the efforts can generate the index; this refers to the first condition in the definition. Next, wary belief requires that efforts maximize profit when consumers believe quality effort to be $\widetilde{e}$. Now, revenue is $T + mD(q(\widetilde{e}))$ given belief, but many quality-cost effort pairs with different disutilities can achieve $\widetilde{I}$. The profit from different quality-cost effort pairs are in (7). Hence, under wary belief $(\widetilde{e}, \widetilde{r})$ must maximize the provider's profit (7) given index $\widetilde{I}$ is to be achieved. Our key Lemma characterizes all quality-cost effort pairs at each index level that satisfy wary belief. (Proofs are in the Appendix.)

Lemma 1 *Under wary belief, for any index* I*, consumers believe that quality-cost effort pair* $(\widetilde{e}, \widetilde{r})$ *solves*

$$\min_{e,r} \Lambda(e,r)$$

$$\textit{subject to} \quad \theta q(e) + (1-\theta)[K - C(e,r)] = I. \tag{9}$$

Hence, for $0 < \theta < 1$*,* $(\widetilde{e}, \widetilde{r})$ *satisfies*

$$\frac{\Lambda_1(\widetilde{e}, \widetilde{r})}{\Lambda_2(\widetilde{e}, \widetilde{r})} = -\frac{\theta q'(\widetilde{e}) - (1-\theta)C_1(\widetilde{e}, \widetilde{r})}{(1-\theta)C_2(\widetilde{e}, \widetilde{r})}. \tag{10}$$

At each I *and* $0 < \theta < 1$*, there is a unique effort pair that satisfies wary belief. Furthermore,* $\widetilde{r} > 0$ *if and only if* $\theta < 1$*, and* $\widetilde{e} > 0$ *if and only if* $\theta > 0$*.*

[13] Wary belief is often adopted in the industrial organization literature. Recent papers that use the restriction include Arya and Mittendorf (2011), Inderst and Ottaviani (2012), Nocke and White (2007), and Rey and Verge (2004).

Lemma 1 states that under wary belief, quality and cost efforts must minimize their disutility for achieving any level of the value index. For a given index level I, consumers' belief about $q(\widetilde{e})$ is fixed, and so is the revenue $T+mD(q(\widetilde{e}))$. The maximization of (7) is the same as the minimization of $\Lambda(e,r)$. The condition in (10) gives the optimality condition for the constrained minimization of $\Lambda(e,r)$. The left-hand side of (10) is the ratio of the marginal disutilities and must be equal to the ratio of the marginal contributions of quality and cost efforts to achieve the index, given the quality weight θ.

For a given weight θ and a level of the index $\widehat{I}$, equilibrium efforts are unique. This follows from the convexity of Λ and the convexity of the constrained set. For strictly positive quality and cost efforts, the weight θ must be strictly between 0 and 1. The striking implication of Lemma 1 is that even when unit variable costs, $C(e,r)$, are completely reimbursed, the provider still has an incentive to exert cost effort. The key is that consumers infer quality from the value index. Cost effort contributes to the value index, so profit is maximized by a combination of quality and cost efforts.[14]

Lemma 1 stems from the provider maximizing profits. Even if consumers used an arbitrary rule to infer quality from index, say an increasing function $\Psi(I)$, the provider's revenue would remain unaffected by any deviations that would maintain the same index level. In equilibrium the provider must still choose efforts to minimize the disutility from achieving the index.

Equilibrium efforts can be illustrated in Figure 1. Because $\Lambda(e,r)$ is convex, its lower contour sets, $\{(e,r) : \Lambda(e,r) \leq \Lambda\}$, are convex, so in Figure 1 we show an iso-disutility line Λ_1 concave to the origin. Consider the constraint in Lemma 1. Because q is concave and C is convex, the upper contour sets, $\{(e,r) : \theta q(e) + (1-\theta)[K - C(e,r)] \geq I\}$, are convex. In Figure 1, the iso-index lines, at levels I_1 and I_2, $I_1 < I_2$, are the circular lines. A solution to the disutility minimization problem in Lemma 1 is the tangency point between the iso-index and iso-disutility lines. As the level of the value index changes, condition (10) defines a unique pair of efforts for every level of the value index I. In Figure 1, the dotted "expansion path"

[14] Wary belief coincides with the restriction that consumers cannot believe the provider choosing a weakly dominated strategy. Suppose the provider chooses an off-equilibrium index I'. Suppose that in the continuation game, some consumers decide to obtain services. The provider's payoff would be higher if it had chosen the quality-cost effort pair to minimize the disutility rather than any other pair. For any I', all effort pairs are weakly dominated by the disutility-minimizing pair.

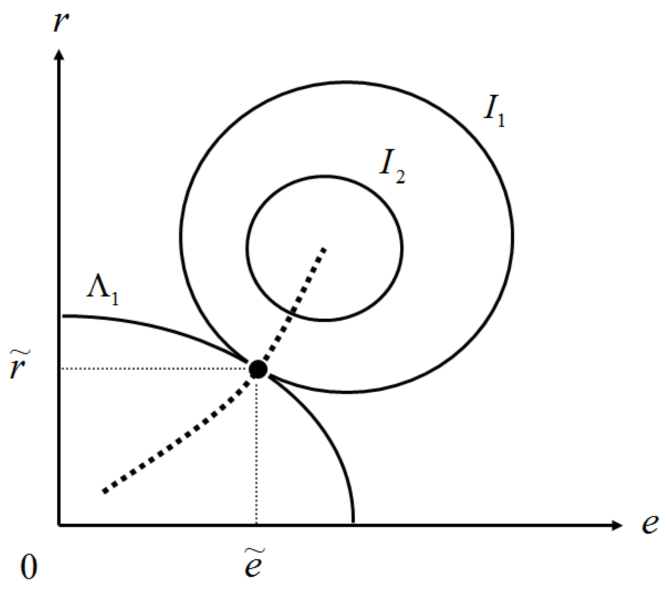

Figure 1: Disutility-minimizing quality and cost efforts

plots these quality-cost effort pairs. Changing the index weight θ corresponds to changing the entire map of the iso-index lines.

For any θ and I, let $e(I;\theta)$ and $r(I;\theta)$ be the unique solution of the disutility minimization program in Lemma 1. They are implicitly defined by (10) and the constraint (9). Furthermore, let $\overline{\Lambda}(I;\theta) \equiv \Lambda(e(I;\theta), r(I;\theta))$. It can be verified easily that $\overline{\Lambda}$ is strictly increasing and convex in I (and the proof is in the Appendix).

How does the provider choose equilibrium efforts in Stage 2? Given beliefs in Lemma 1, any equilibrium effort choice must be given by $e(I;\theta)$ and $r(I;\theta)$ for each I. Hence, we can equivalently let the provider choose an index level. For example, if the provider chooses to achieve the index level I_1 in Figure 1, the corresponding efforts must be $\widetilde{e}$ and $\widetilde{r}$. We rewrite the provider's profit as

$$T + mD(q(e(I;\theta))) - \overline{\Lambda}(I;\theta). \tag{11}$$

In the Appendix, we write down sufficient conditions for the profit in (11) to be strictly quasi-concave in I. These conditions say that equilibrium effort $e(I;\theta)$ is concave in the index I; so $e(I;\theta)$ cannot increase in I at an increasing rate. When (11) is strictly quasi-concave, we can relate the provider's equilibrium efforts to the margin and index weight by the first-order conditions of profit maximization.

Recall that first-best efforts are e^* and r^* in Subsection 2.3. Now define I^* and θ^* by

$$I^* = \theta^* q(e^*) + (1-\theta^*)[K - C(e^*, r^*)] \tag{12}$$

$$\frac{\Lambda_1(e^*, r^*)}{\Lambda_2(e^*, e^*)} = -\frac{\theta^* q(e^*) - (1-\theta^*)C_1(e^*, r^*)}{(1-\theta^*)C_2(e^*, e^*)} = \frac{C_1(e^*, r^*)}{C_2(e^*, r^*)} - \frac{\theta^*}{1-\theta^*}\frac{q'(e^*)}{C_2(e^*, r^*)}. \tag{13}$$

What is the rationale behind this construction of a particular pair of index and weight? From Lemma 1, any weight θ and I uniquely determine a pair of efforts given by (10). We have taken (10) and set θ to θ^* so that the first-best efforts satisfy (10), which we now rewrite as (13). Furthermore, if the value index happens to take the value of I^* in (12), equilibrium efforts will be first best. If the provider can be incentivized to choose I^*, first-best efforts will be equilibrium efforts. The payment margin that implements first-best efforts is

$$m^* = \frac{\overline{\Lambda}_1(I^*; \theta^*)}{D'(q^*)q'(e^*)e_1(I^*; \theta^*)}, \tag{14}$$

and we can now state our result.

Proposition 2 *Suppose that the profit function (11) is strictly quasi-concave in I. The insurer implements the first-best efforts (e^*, r^*) in the unique continuation equilibrium by setting the weight of the value index to θ^* in (13) and the cost margin to m^* in (14), and a suitable transfer T.*

Consumers infer quality from the value index. This inference stems from the key Lemma 1. The Lemma also indicates how the provider must choose efforts to attain any level of the index. By setting the weight at θ^*, first-best efforts are optimal at index level I^* (see (12) and (13) above). Any payment margin m incentivizes the provider to raise the index, and therefore both quality and cost efforts. The profit-maximizing index is one where the marginal profit $m^* D'(q)q'(e)e_1(I; \theta^*)$ is equal to marginal disutility $\overline{\Lambda}_1(I; \theta^*)$. The value of m^* ensures that the index I^* is the profit-maximizing index. The point of Proposition 2 is that information disclosure and payment policy can be coordinated for implementation of efficient cost and quality efforts—even when variable costs are fully reimbursed.

It is important that consumers rely on the value index to infer about quality. If a provider could credibly reveal its quality, it could avoid the constraint on the equilibrium mix of quality and cost effort due to the value index (Lemma 1). Cost information *per se* is not valuable to consumers. If the provider does not need to exert cost effort to convey quality information to consumers, the perverse cost effort property of cost reimbursement remains. The policy implication is perhaps quite obvious: public agencies should have a keen interest in information disclosure. A more radical policy would require public certification or regulation of any information disclosure.

It seems difficult for a firm to credibly disclose product qualities. The general consensus in the literature is that disclosure may involve another interested party, and a new set of incentive problems arises. (See again our discussion of that literature in Subsection 1.1.) Disclosure by a public agency or a trusted nonprofit organization may be more credible. Also, an insurer aiming to control cost in the long run should have an incentive to build a reputation of trustworthiness. That trust allows the implementation of the first best.

4 Extensions: stochastic quality, dumping, and cream skimming

In this section, we consider three extensions. First, we use a stochastic quality production model common in the principal-agent literature. Then we allow the provider to serve only profitable patients when costs are stochastic. Next, we revert to the deterministic production model but let health services have many qualities. This version allows us to consider cream skimming.

4.1 Stochastic quality and value index

In the previous section, the provider's quality effort produces quality in a deterministic fashion. We now consider stochastic quality production. We use the standard hidden-action model of Mirrlees (1976) and Holmstrom (1979). Quality effort e determines a distribution on quality q, a random variable defined on $\mathbb{R}_+$, the positive real numbers.[15] For any quality level q, the insurer cannot rule out any quality effort e. We write the density function of q as $f(q|e)$, and assume that the expected demand is strictly increasing and strictly concave in quality effort. That is, $\int_{\mathbb{R}_+} D(q)f(q|e)\mathrm{d}q$ is strictly increasing and concave in e.

We first provide the notation for the first best. The social welfare expression in (1) is rewritten as

[15] The full-support assumption is commonly made. There are two reasons. First, nonoverlapping supports in qualities as effort changes may allow the insurer to infer effort, so our assumption eliminates that kind of inference. Second, a full quality support serves as an approximation to any model with a bounded support; we can simply set the density to be arbitrarily close to zero.

$\int_{\mathbb{R}_+} [B(q) - D(q)C(e,r)] f(q|e)\mathrm{d}q - \Lambda(e,r)$. Here, the integral is the expected social benefit less variable costs while the remaining term is the effort disutility, and we assume that social welfare is strictly quasi-concave in efforts. For completeness, we write down the characterization of the first-best quality and cost efforts in this notation (but compare them with (2) and (3)):

$$\begin{aligned}\int_{\mathbb{R}_+} \left\{ [B(q) - D(q)C(e^*,r^*)] \left[\frac{\partial f(q|e^*)}{\partial e} \right] - D(q)C_1(e^*,r^*)f(q|e^*) \right\} \mathrm{d}q - \Lambda_1(e^*,r^*) &= 0 \\ \int_{\mathbb{R}_+} [-D(q)C_2(e^*,r^*)] f(q|e^*)\mathrm{d}q - \Lambda_2(e^*,r^*) &= 0.\end{aligned}$$

Stochastic quality from effort does not affect the performance of prospective payment in any way. The insurer pays a fixed price and reports any realized quality. In this notation, the prospective price implementing the first best in Proposition 1 is written as

$$p^* = \frac{\displaystyle\int_{\mathbb{R}_+} \left[B(q) \frac{\partial f(q|e^*)}{\partial e} \right] \mathrm{d}q}{\displaystyle\int_{\mathbb{R}_+} \left[D(q) \frac{\partial f(q|e^*)}{\partial e} \right] \mathrm{d}q},$$

and we omit the corresponding expression of the transfer.

Now we consider value index and cost reimbursement. The insurer observes the realized quality q and cost $C(e,r)$. Because q is stochastic, so is the index $I(q,C;\theta) \equiv \theta q + (1-\theta)[K - C]$. Accordingly our analysis has to proceed differently from the previous subsection. The extensive form is as in Subsection 3.2, except that in Stage 3, the insurer observes the realized quality according to the density chosen by the provider. Again, consumers know neither q nor $C(e,r)$. The level of the value index is all they observe.[16]

Consider an equilibrium in which the provider chooses effort pair $(\widehat{e},\widehat{r})$. In this equilibrium, consumers believe that unit cost is $C(\widehat{e},\widehat{r})$. Quality q will be drawn according to density $f(q|\widehat{e})$. When the level of value index is I, consumers believe that quality $\widehat{q}$ satisfies $\theta\widehat{q} + (1-\theta)[K - C(\widehat{e},\widehat{r})] = I$. From this, the inferred quality is

$$\widehat{q}(I|(\widehat{e},\widehat{r})) = \frac{I - (1-\theta)[K - C(\widehat{e},\widehat{r})]}{\theta}, \tag{15}$$

where we have emphasized that the inference rule $\widehat{q}(I|(\widehat{e},\widehat{r}))$ depends on the value index and equilibrium

[16] Quality can be any positive real number, and we assume that the value of $K - C(e,r)$ is always positive. So for any effort pair, the range of the index must be strictly positive. We specify that, should the value index have a level outside this range, consumers would believe that the quality is 0.

efforts.

Consider a quality level, say $\widetilde{q}$. Under equilibrium effort $(\widehat{e},\widehat{r})$, when $\widetilde{q}$ is realized, then $\widehat{q}(I|(\widehat{e},\widehat{r})) = \widetilde{q}$. Suppose that the provider deviates from $(\widehat{e},\widehat{r})$ to (e,r). The unit cost becomes $C(e,r)$, and this will be observed by the insurer but consumers continue to believe that the unit cost is $C(\widehat{e},\widehat{r})$. When the same quality $\widetilde{q}$ is realized, the index changes to $\widetilde{I} \equiv \theta\widetilde{q} + (1-\theta)[K - C(e,r)]$. Using the inference rule in (15), consumers now believe that the quality is

$$\begin{aligned}\widehat{q}(\widetilde{I}|(\widehat{e},\widehat{r})) &= \frac{\widetilde{I} - (1-\theta)[K - C(\widehat{e},\widehat{r})]}{\theta} \\ &= \frac{\theta\widetilde{q} + (1-\theta)[K - C(e,r)] - (1-\theta)[K - C(\widehat{e},\widehat{r})]}{\theta} \\ &= \widetilde{q} + \frac{1-\theta}{\theta}[C(\widehat{e},\widehat{r}) - C(e,r)] \neq \widetilde{q}. \qquad (16)\end{aligned}$$

The point is that if the provider reduces the unit cost, the quality perceived by consumers becomes higher. As an illustration, suppose that the realized quality $\widetilde{q}$ is 10, and $C(\widehat{e},\widehat{r})$ is also 10. Suppose that K is 20, so the index and inferred quality are both 10. By raising cost effort from $\widehat{r}$, the provider reduces the unit cost from 10, so the index increases from 10. Consumers, however, continue to believe that the cost is 10, so any increase in the index is mistakenly attributed to an increase of quality from 10.

In (16), the inferred quality $\widehat{q}(\widetilde{I}|(\widehat{e},\widehat{r}))$ is larger than the realized quality $\widetilde{q}$ if and only if $C(e,r) < C(\widehat{e},\widehat{r})$. This is the basic incentive for the provider to expend cost effort. More important, the insurer can influence this incentive by choosing θ: a smaller θ raises $\dfrac{1-\theta}{\theta}$ in (16). This implies a larger difference between the quality observed by the insurer and the quality inferred by consumers.

In equilibrium, consumers must not be misled, so the equilibrium effort $(\widehat{e},\widehat{r})$ must yield higher profit for the provider than any deviation. Suppose that the provider deviates from the equilibrium $(\widehat{e},\widehat{r})$. The expected payoff from another effort pair (e,r) is

$$m\int_{\mathbb{R}_+} D\left(q + \frac{1-\theta}{\theta}[C(\widehat{e},\widehat{r}) - C(e,r)]\right) f(q|e)\mathrm{d}q - \Lambda(e,r). \qquad (17)$$

In (17), we have used the inference rule (16) when the provider deviates from $(\widehat{e},\widehat{r})$ to express the demand; the integral is the expected revenue when the margin is set at m. Effort pair $(\widehat{e},\widehat{r})$ is an equilibrium if it maximizes (17).

Formally, an equilibrium effort pair $(\widehat{e}, \widehat{r})$ is a fixed point of the following provider's best-response-against-belief correspondence:

$$\phi(e,r) = \operatorname*{argmax}_{e',r'} \left\{ m \int_{\mathbb{R}_+} D\left(q + \frac{1-\theta}{\theta}[C(e,r) - C(e',r')]\right) f(q|e')\mathrm{d}q - \Lambda(e',r') \right\}. \tag{18}$$

By the Maximum Theorem, the correspondence $\phi(e,r)$ is upper-semi continuous when (17) is continuous in (e,r) and $(\widehat{e},\widehat{r})$. When (17) is strictly quasi-concave in (e,r) for any given $(\widehat{e},\widehat{r})$, the correspondence ϕ is single-valued, so it is actually a continuous function. We let ϕ be differentiable. Furthermore, we will assume that ϕ is a contraction map, so it has a unique fixed point. In the Appendix we write down sufficient conditions for (17) to be strictly quasi-concave in (e,r) for any given $(\widehat{e},\widehat{r})$, and for ϕ to be a contraction map.

The first-order derivatives of (17) with respect to efforts are (30) and (31) in the Appendix. We set these first-order derivatives to zero, and then set (e,r) in the first-order conditions to $(\widehat{e},\widehat{r})$ to get, respectively,

$$\int_{\mathbb{R}_+} m \left\{ D\left(q\right) \left[\frac{\partial f(q|\widehat{e})}{\partial e}\right] - D'\left(q\right) C_1(\widehat{e},\widehat{r}) \frac{1-\theta}{\theta} f(q|\widehat{e}) \right\} \mathrm{d}q - \Lambda_1(\widehat{e},\widehat{r}) \quad = \quad 0 \tag{19}$$

$$\int_{\mathbb{R}_+} -m \left[D'\left(q\right) C_2(\widehat{e},\widehat{r}) \frac{1-\theta}{\theta}\right] f(q|\widehat{e})\mathrm{d}q - \Lambda_2(\widehat{e},\widehat{r}) \quad = \quad 0. \tag{20}$$

For any margin m and weight θ, these two first-order conditions yield the unique equilibrium efforts.

Proposition 3 *Suppose that the profit function in (17) is strictly quasi-concave, and that the best-response ϕ is a contraction map. The insurer implements the first-best efforts (e^*, r^*) in the unique continuation equilibrium by setting the weight of the value index θ^* and margin m^* to satisfy (19) and (20) at $(\widehat{e},\widehat{r}) = (e^*, r^*)$ together with a suitable transfer T.*

To contrast with Proposition 2, we can combine (19) and (20), and set $(\widehat{e},\widehat{r})$ to (e^*, r^*) to get

$$\frac{\Lambda_1(e^*,r^*)}{\Lambda_2(e^*,e^*)} = \frac{C_1(e^*,r^*)}{C_2(e^*,r^*)} - \frac{\theta^*}{1-\theta^*}\frac{1}{C_2(e^*,r^*)}\frac{\displaystyle\int_{\mathbb{R}_+} D\left(q\right)\frac{\partial f(q|e^*)}{\partial e}\mathrm{d}q}{\displaystyle\int_{\mathbb{R}_+} D'\left(q\right) f(q|e^*)\mathrm{d}q}, \tag{21}$$

which can be interpreted similarly as (13) for the implementation of the first best under deterministic quality production.

The result here contrasts with the sufficient-statistic result in Holmstrom (1979). In the classical principal-agent model, the efficient way to motivate unobservable effort is to use payments based on signals that are sufficient statistics of the agent's action. Therefore, payments based on garbled informative signals are suboptimal. Here, the insurer purposefully garbles the information about quality with cost information, which is irrelevant to consumers. Garbled information leads to cost effort affecting demand through the value index.

4.2 Stochastic cost reduction and dumping

In this subsection, we discuss dumping under cost reimbursement and prospective payment. Here, we revert to the assumption that quality production is deterministic. We continue to assume that the insurer seeks to implement a given quality-cost effort pair. We now extend the model in Section 3 to include cost heterogeneity. Let variable cost c be random. Given that K has been defined as the cost ceiling, we let c vary on the closed support $[0, K]$. Let $g(c|e, r)$ denote the density of c, given effort pair (e, r). Now we let $C(e, r) \equiv \int_{0 \leq c \leq K} cg(c|e, r)\mathrm{d}c$ denote the average cost.

The use of a value index requires the insurer to obtain information about costs. We continue with the assumption that the provider cannot manipulate information. When costs are stochastic, the insurer may audit the provider to find out about the cost distribution after cost effort has been chosen. Alternatively, the insurer may sample patient cases to obtain cost estimates.[17] We assume that auditing or sampling are sufficiently accurate to estimate the average variable cost, so the average cost $C(e, r) \equiv \int_{0 \leq c \leq K} cg(c|e, r)\mathrm{d}c$ is used to construct the value index in cost reimbursement (see, for example, (9) in Lemma 1). The extensive form is the same as in Subsection 3.2 with two changes. First, in Stage 3, the insurer uses the average cost to construct the value index. Second, after Stage 4, the provider has an additional strategy of refusing to serve a consumer after the cost realization.

Dumping refers to a provider refusing to give service to high-cost consumers. Under cost reimbursement, realized costs are not the provider's responsibility; therefore, the provider has no incentive to turn away

[17] In the formal model, consumer demand is not determined until the value index is disclosed. However, we implicitly assume that the provider serves some consumers even before that. The insurer samples these cases to estimate the average cost. In practice, average-cost information from an earlier period may be used.

high-cost consumers. However, by expending cost effort r, the provider changes the average cost $C(e,r)$, and hence the value index. The incentive effect on cost effort remains the same as when costs are deterministic. Under prospective payment, the provider has to internalize all variable costs. If a consumer's cost turns out to be higher than the prospective price, the provider will turn away the consumer. The first best cannot be implemented under prospective payment, whereas cost reimbursement with value index can.

We have separately discussed stochastic quality production and stochastic cost reduction. Extending to the environment where quality production and cost reduction are both stochastic is straightforward. We now reinterpret $C(\widehat{e},\widehat{r})$ and $C(e,r)$ in the inference equation (16) as the average costs $\int_{0\le c\le K} cg(c|\widehat{e},\widehat{r})\mathrm{d}c$ and $\int_{0\le c\le K} cg(c|e,r)\mathrm{d}c$, respectively. The arguments for Proposition 3 remain valid.

4.3 Multiple qualities and cream skimming

Now we return to our basic model in Section 2, but let health services have two qualities, q_A and q_B. Let e_A and e_B be two corresponding quality efforts. For ease of exposition, we simply let $(q_A, q_B) = (e_A, e_B)$. We extend the notation for demand, social benefit, variable cost, and disutility in the obvious way: $D(q_A, q_B)$, $B(q_A, q_B)$, $C(q_A, q_B, r)$, $\Lambda(q_A, q_B, r)$. We also maintain the corresponding concavity and convexity assumptions.

The social welfare is now

$$B(q_A, q_B) - D(q_A, q_B)C(q_A, q_B, r) - \Lambda(q_A, q_B, r). \tag{22}$$

Let q_A^*, q_B^*, r^* be the first-best qualities and cost effort, those that maximize (22).[18] Under prospective payment with transfer T, price p, and complete quality-information disclosure, the provider's profit is

$$T + pD(q_A, q_B) - D(q_A, q_B)C(q_A, q_B, r) - \Lambda(q_A, q_B, r).$$

If the insurer discloses information of both q_A and q_B, a prospective price can be chosen to implement the

[18] They are characterized by the first-order conditions:

$$\begin{aligned} B_1(q_A^*, q_B^*) - D_1(q_A^*, q_B^*)C(q_A^*, q_B^*, r^*) - D(q_A^*, q_B^*)C_1(q_A^*, q_B^*, r^*) - \Lambda_1(q_A^*, q_B^*, r^*) &= 0 \\ B_2(q_A^*, q_B^*) - D_2(q_A^*, q_B^*)C(q_A^*, q_B^*, r^*) - D(q_A^*, q_B^*)C_2(q_A^*, q_B^*, r^*) - \Lambda_2(q_A^*, q_B^*, r^*) &= 0 \\ -D(q_A^*, q_B^*)C_3(q_A^*, q_B^*, r^*) - \Lambda_3(q_A^*, q_B^*, r^*) &= 0, \end{aligned}$$

which have the usual interpretations.

first best if and only if

$$\frac{B_1(q_A^*, q_B^*)}{D_1(q_A^*, q_B^*)} = \frac{B_2(q_A^*, q_B^*)}{D_2(q_A^*, q_B^*)} \tag{23}$$

(which is also the prospective price). This result is obtained by comparing the first-order conditions for the first best (as in Footnote 18) and for the provider's profit maximization (as in Proposition 1).

With a single quality, a single prospective price implements the first best, as in Proposition 1, but with multiple qualities, a single prospective price generally fails. The provider internalizes cost under prospective payment. However, each quality's marginal contribution to the provider's revenue is generally different from its marginal contribution to social benefit. Condition (23) imposes the equality of these marginal contributions. To see this, rearrange (23) to

$$\frac{B_1(q_A^*, q_B^*)}{B_2(q_A^*, q_B^*)} = \frac{pD_1(q_A^*, q_B^*)}{pD_2(q_A^*, q_B^*)}, \tag{24}$$

which says that the marginal rates of substitution between the two qualities have to be identical in the social benefit function and the revenue function. When (24) fails to hold, the provider will engage in cream skimming by exploiting the differential demand responses from different qualities.

To prevent cream skimming, the misalignment between the provider's and the social tradeoff between qualities must be resolved. Under prospective payment, the insurer can correct this misalignment by disclosing a quality index, rather than full information about the qualities. Suppose that the service qualities are q_A and q_B. Construct the quality index $J(q_A, q_B; \phi) \equiv \phi q_A + (1-\phi) q_B$, where $0 \le \phi \le 1$. The insurer announces this quality index. When consumers observe $J(q_A, q_B; \phi)$, they draw inferences about the unobservable qualities q_A and q_B.

Analogous to Lemma 1, the equilibrium inference must be qualities $\widehat{q}_A$ and $\widehat{q}_B$ which solve

$$\max_{q_A, q_B, r} T + pD(\widehat{q}_A, \widehat{q}_B) - D(\widehat{q}_A, \widehat{q}_B) C(q_A, q_B, r) - \Lambda(q_A, q_B, r)$$
$$\text{subject to } \phi q_A + (1-\phi) q_B = \widehat{J} = \phi \widehat{q}_A + (1-\phi)\widehat{q}_B. \tag{25}$$

Any choice of qualities that achieve the quality index level $\widehat{J}$ will yield the same inference. The provider optimally chooses those quality efforts that maximize profit, given the quality index. A suitable choice of the index weight ϕ therefore can implement the first-best marginal rate of substitution between the two quality efforts, as in (24). The insurer next chooses a prospective price. Given that the provider internalizes the total cost, a quality index and a prospective payment are sufficient to implement the first best.

Cost reimbursement with value index can perform exactly the same. Here, the insurer constructs a value index: $I(q_A, q_B, C; \theta_A, \theta_B) \equiv \theta_A q_A + \theta_B q_B + (1 - \theta_A - \theta_B)[K - C]$, where the weights, θ_A and θ_B, are positive and $\theta_A + \theta_B \leq 1$. Under cost reimbursement, equilibrium qualities and cost effort must minimize the disutility. Any equilibrium $\widehat{q}_A$, $\widehat{q}_B$ and $\widehat{r}$ solve

$$\begin{aligned} &\max_{q_A, q_B, e_2} T + mD(\widehat{q}_A, \widehat{q}_B) - \Lambda(q_A, q_B, r) \\ \text{subject to} \quad &\theta_A q_A + \theta_B q_B + (1 - \theta_A - \theta_B)[K - C(q_A, q_B, r)] \\ &= \widehat{I} = \theta_A \widehat{q}_A + \theta_B \widehat{q}_B + (1 - \theta_A - \theta_B)[K - C(\widehat{q}_A, \widehat{q}_B, \widehat{r})]. \end{aligned} \tag{26}$$

Using the value-index weights, the insurer controls how the provider trades off between each quality and the cost effort, analogous to Lemma 1. Finally, using the margin, the insurer implements the first best, as in Proposition 2.

5 Conclusion

Prospective payment and cost reimbursement are common payment mechanisms to providers for health care services. In the past thirty years, many theoretical and empirical studies have pointed out the different quality and cost incentives of the two payment systems. In this paper, we have shown how an insurer, by optimally choosing the content of public report, can make the two payment systems implement identical quality and cost incentives. Our results are robust to environments where the provider's productions of quality and cost reduction are stochastic, and where health services have many qualities. Furthermore, cost reimbursement may perform better than prospective payment when provider dumping of expensive consumers is possible. Quality and value indexes may eliminate cream-skimming incentives.

The main point here is that information can act as an incentive strategy. Given that health service quality is difficult for consumers to know about, it is incumbent upon insurers and regulators to inform consumers. The usual approach is a sort of "empowering" consumers with as much information as common consumer cognition allows. Here, we question this approach. Information disclosure affects a provider's incentive to invest in quality and cost efforts, and should be considered along with payment mechanisms.

We have assumed that the insurer can make a lump-sum transfer to the provider. This is consistent with the vast majority of the literature on provider payment design. Two recent papers study optimal provider payment systems when lump-sum transfer is not allowed. Mougeot and Naegelen (2005) show that the first-best quality and cost efforts are not attainable without transfer. They then characterize the constrained-optimal prospective price and margin. Miraldo et al. (2011) further characterize the constrained-optimal prospective price list when providers have different cost types. In our model, the first best may not be achieved when transfer is not allowed; a single prospective price or margin cannot handle both distribution and incentive problems. Yet, value-index reporting will continue to induce cost-reduction effort under cost reimbursement.

As the health care market evolves, payment systems have tended to become complicated. Pay-for-performance incentive design is now discussed often in policy and theoretical research; see, for example, works by Eggleston (2005), Kaarboe and Siciliani (2011), McClellan (2011), and Richardson (2011). Our paper calls for a more fundamental approach. Any reward system must be based on available information. A central issue, as we have shown here, is how the insurer may strategically disclose information. Furthermore, information and financial instruments should be chosen simultaneously to align incentives.

Appendix

Proof of Lemma 1

For any given I and belief $\widetilde{e}$, e and r maximize $T + mD(q(\widetilde{e})) - \Lambda(e,r)$ subject to $\theta q(e) + (1-\theta)[K - C(e,r)] = I$ if and only if e and r minimize $\Lambda(e,r)$ subject to $\theta q(e) + (1-\theta)[K - C(e,r)] = I$, which is the constrained minimization program in the Lemma. Minimizing $\Lambda(e,r)$ subject to $\theta q(e)+(1-\theta)[K-C(e,r)] = I$, we obtain the first-order condition (10). Uniqueness follows from the strict convexity of Λ and C, and the strict concavity of q.

Finally, if $\theta = 1$, the constraint becomes $q(e) = I$. Because Λ is increasing, the disutility-minimizing cost effort must be zero. If $\theta = 0$, the constraint becomes $K - C(e,r) = I$. Because both Λ and C are increasing in e, the disutility-minimizing quality effort must be zero.

Sufficient condition for the profit function (11) to be strictly quasi-concave in I

We first show that $\overline{\Lambda}(I;\theta) \equiv \Lambda(e(I;\theta), r(I;\theta))$ is convex in I. Omit the variable θ in the proof. Recall that $(e(I), r(I)) = \text{argmin}_{e,r}\, \Lambda(e,r)$ subject to $\theta q(e) + (1-\theta)[K - C(e,r)] = I$. We modify this constrained minimization program to $\min_{e,r} \Lambda(e,r)$ subject to $G(e,r) \geq I$, where $G(e,r) \equiv \theta q(e) + (1-\theta)[K - C(e,r)]$. Clearly, G is concave because q is concave and C is convex. The relaxation of the constraint to the weak inequality is of no consequence, because at any solution the constraint binds.

Consider two indexes I_1 and I_2, and $I = \alpha I_1 + (1-\alpha)I_2$, for some $0 < \alpha < 1$. Then $G(e(I_1), r(I_1)) = I_1$ and $G(e(I_2), r(I_2)) = I_2$. By the concavity of G, we have $G(\alpha e(I_1) + (1-\alpha)e(I_2), \alpha r(I_1) + (1-\alpha)r(I_2)) > \alpha G(e(I_1), r(I_1)) + (1-\alpha)G(e(I_2), r(I_2)) = \alpha I_1 + (1-\alpha)I_2 = I$. Hence the effort pair $(\alpha e(I_1) + (1-\alpha)e(I_2), \alpha r(I_1) + (1-\alpha)r(I_2))$ achieves the index I. Therefore, $\overline{\Lambda}(I;\theta) \equiv \Lambda(e(I), r(I)) \leq \Lambda(\alpha e(I_1) + (1-\alpha)e(I_2), \alpha r(I_1) + (1-\alpha)r(I_2)) < \alpha\Lambda(e(I_1), r(I_1)) + (1-\alpha)\Lambda(e(I_2), r(I_2)) \equiv \alpha\overline{\Lambda}(I_1;\theta) + (1-\alpha)\overline{\Lambda}(I_2;\theta)$. We conclude that $\overline{\Lambda}(I;\theta)$ is convex in I.

Next, we provide a sufficient condition for $D(q(e(I;\theta)))$ to be strictly concave in I. This and the convexity of $\overline{\Lambda}(I;\theta)$ guarantee that the profit function (11) is strictly concave, and hence quasi-concave in I. The second-order derivative of $D(q(e(I)))$ with respect to I is

$$D'' \left[q' e_1(I;\theta)\right]^2 + D' q'' \left[e_1(I;\theta)\right]^2 + D' q' e_{11}(I;\theta), \tag{27}$$

where we have suppressed the arguments in D, q, and their derivatives. Because D and q are increasing and concave, (27) is negative if $e_{11}(I;\theta) < 0$. Therefore, if $e_{11}(I;\theta) < 0$, $D(q(e(I)))$ is concave. We can use the first-order conditions from Lemma 1, apply the implicit function theorem to find the derivatives of e and r in terms of I. Then we can find the second-order derivatives of e and r in terms of D, q, and Λ. The manipulations are tedious but straightforward, and omitted.

Proof of Proposition 2

By construction and Lemma 1, at $\theta = \theta^*$, if the provider chooses $I = I^*$, the provider chooses efforts e^* and r^*. Now at $m = m^*$, the derivative of (11) with respect to I is

$$m^* D'(q) q'(e) e_1(I;\theta^*) - \overline{\Lambda}_1(I;\theta^*),$$

which vanishes at $I = I^*$. Because the profit function is assumed to be strictly quasi-concave in I, I^* is the unique maximizer of (11). The value of the transfer T is chosen such that $T + m^* D(q(e^*)) - \Lambda(e^*, r^*) = 0$, so the provider makes a zero profit.

Sufficient condition for the profit function (17) strictly quasi-concave in e and r

Let $\sigma(e,r) \equiv q + \dfrac{1-\theta}{\theta}[C(\widehat{e},\widehat{r}) - C(e,r)]$, we can rewrite the profit function in (17) as

$$L(e,r) \equiv m \int_{\mathbb{R}_+} D\left(\sigma(e,r)\right) f(q|e) \mathrm{d}q - \Lambda(e,r).$$

$L(e,r)$ is strictly quasi-concave in e and r if at any (e,r), $(e,r) \neq (\widehat{e},\widehat{r})$,

$$2L_1 L_2 L_{12} - L_1^2 L_{22} - L_2^2 L_{11} > 0 \tag{28}$$

(Chiang and Wainwright (2005, p370)), where subscripts denote partial and cross-partial derivatives, and where

$$\begin{aligned}
L_1 &= m\int_{\mathbb{R}_+}(Df' + D'\sigma_1 f)\mathrm{d}q - \Lambda_1 \\
L_2 &= m\int_{\mathbb{R}_+} D'\sigma_2 f\mathrm{d}q - \Lambda_2 \\
L_{11} &= m\int_{\mathbb{R}_+}(D''\sigma_1^2 f + D'\sigma_{11} f + 2D'\sigma_1 f' + Df'')\mathrm{d}q - \Lambda_{11} \\
L_{22} &= m\int_{\mathbb{R}_+}(D''\sigma_2^2 f + D'\sigma_{22} f)\mathrm{d}q - \Lambda_{22} \\
L_{12} &= m\int_{\mathbb{R}_+}(D''\sigma_1\sigma_2 f + D'\sigma_{12} f + D'\sigma_2 f')\mathrm{d}q - \Lambda_{12}.
\end{aligned}$$

Sufficient condition for the correspondence (18) to be a contraction map

Let (18) be differentiable. The correspondence is a contraction map if

$$\left\| \begin{array}{cc} \phi_{11}(e,r) & \phi_{12}(e,r) \\ \phi_{12}(e,r) & \phi_{22}(e,r) \end{array} \right\| < \lambda < 1 \tag{29}$$

at every nonnegative (e, r) (Hasselblatt and Katok (2003, p38)), where subscripts denote partial and cross-partial derivatives, and where $\| \;\; \|$ denotes the norm of the matrix.

Proof of Proposition 3

The partial derivatives of (17) with respect to e and r are

$$\int_{\mathbb{R}_+} m \left\{ \begin{array}{c} D\left(q + \dfrac{1-\theta}{\theta}[C(\widehat{e},\widehat{r}) - C(e,r)]\right)\left[\dfrac{\partial f(q|e)}{\partial e}\right] \\ -D'\left(q + \dfrac{1-\theta}{\theta}[C(\widehat{e},\widehat{r}) - C(e,r)]\right) C_1(e,r)\dfrac{1-\theta}{\theta} f(q|e) \end{array} \right\} \mathrm{d}q - \Lambda_1(e,r) \tag{30}$$

$$\int_{\mathbb{R}_+} - m\left[D'\left(q + \frac{1-\theta}{\theta}[C(\widehat{e},\widehat{r}) - C(e,r)]\right) C_2(e,r)\frac{1-\theta}{\theta}\right] f(q|e)\mathrm{d}q - \Lambda_2(e,r). \tag{31}$$

Then we set (e, r) to $(\widehat{e}, \widehat{r})$ and the two derivatives to zero. These equations then yield the equilibrium conditions (19) and (20).

Given conditions (28) and (29) above, the profit function (17) is strictly quasi-concave and the correspondence (18) is a contraction map, hence the equilibrium conditions characterize the unique pair of (e, r) that maximizes profit. To implement the first best, set $(\widehat{e}, \widehat{r})$ in (19) and (20) to (e^*, r^*). The values of θ^* and m^* are, respectively, given by (21) and

$$m^*\int_{\mathbb{R}_+}\left\{D(q)\left[\frac{\partial f(q|e^*)}{\partial e}\right]\right\}\mathrm{d}q - \Lambda_1(e^*, r^*) - \Lambda_2(e^*, r^*)\left[-\frac{C_1(e^*, r^*)}{C_2(e^*, r^*)}\right] = 0.$$

Finally, the value of T is again chosen so that $T + m^* \int_{\mathbb{R}_+} D(q) f(q|e^*) \mathrm{d}q - \Lambda(e^*, r^*) = 0$.

References

Albano GL, Lizzeri A. Strategic certification and provision of quality. International Economic Review 2001; 42; 267-283.

Arya A, Mittendorf B. Disclosure standards for vertical contracts. RAND Journal of Economics 2011; 42; 595–617.

Baron DP, Myerson RB. Regulating a monopolist with unknown costs. Econometrica 1982; 50; 911-930.

Beaulieu ND. Quality information and consumer health plan choices. Journal of Health Economics 2002; 21; 43-63.

Brekke K, Nuscheler R, Rune Straume O. Quality and location choices under price regulation. Journal of Economics and Management Strategy 2006; 15; 207-227.

Bureau of Labor Statistics. Selected medical benefits: A report from the Department of Labor to the Department of Health and Human Services 2011.

Catalyst for Payment Reform. National scorecard on payment reform 2014.

Centers for Medicare and Medicaid Services. The Medicare and Medicaid statistical supplement 2013; Chapter 2.

Chalkley M, Malcomson J. Contracting for health services when patient demand does not reflect quality. Journal of Health Economics 1998; 17; 1-19.

Chiang A, Wainwright K. Fundamental methods of mathematical economics. McGraw-Hill: Columbus; 2005.

Dafny L. How do hospitals respond to price changes? American Economic Review 2005; 95; 1525-1547.

Dafny L, Dranove D. Do report cards tell consumers anything they don't already know? The case of Medicare HMOs. RAND Journal of Economics 2008; 39; 790-82.

Dranove D. Demand inducement and the physician/patient relationship. Economic Inquiry 1988; 26; 281-298.

Dranove D, Jin GZ. Quality disclosure and certification: Theory and practice. Journal of Economic Literature 2011; 48; 935-963.

Eggleston K. Multitasking and mixed systems for provider payment. Journal of Health Economics 2005; 24; 211-223.

Frank R, Glazer J, McGuire TG. Measuring adverse selection in managed health care. Journal of Health Economics 2000; 19; 829-854.

Glazer J, McGuire TG. Optimal risk adjustment in markets with adverse selection: An application to managed care. American Economic Review 2000; 90; 1055-1071.

Glazer J, McGuire TG. Optimal quality reporting in markets for health plans. Journal of Health Economics 2006; 25; 295-310.

Hasselblatt B, Katok A. A first course in dynamics with a panorama of recent developments. Cambridge University Press: Cambridge; 2003.

Holmstrom B. Moral hazard and observability. Bell Journal of Economics 1979; 10; 74-91.

Inderst R, Ottaviani M. Competition through commissions and kickbacks. American Economic Review 2012; 102; 780-809.

Kaarboe O, Siciliani L. Multi-tasking, quality and pay for performance. Health Economics 2011; 20; 225-238.

Leger PT. Physician payment mechanisms. In: Lu M, Jonsson E (Eds), Financing health care: New ideas for a changing society. Wiley-VCH; 2008. 149-176.

Lizzeri A. Information revelation and certification intermediaries. RAND Journal of Economics 1999; 30; 214-231.

Ma CA. Health care payment systems: Cost and quality incentives. Journal of Economics & Management Strategy 1994; 3; 93-112.

Ma CA, Mak H. Public report, price, and quality. Journal of Economics & Management Strategy 2014a; 23; 443-464.

Ma CA, Mak H. Tiered and value-based health care networks. Mimeo, Department of Economics, Boston University, 2014b.

Ma CA, McGuire TG. Optimal health insurance and provider payments. American Economic Review 1997; 87; 685-689.

Matthews S, Postlewaite A. Quality testing and disclosure. RAND Journal of Economics 1985; 16; 328-340.

McAfee RP, Schwartz M. Opportunism in multilateral vertical contracting: Nondiscrimination, exclusivity, and uniformity. American Economic Review 1994; 84; 210-230.

McClellan M. Reforming payments to healthcare providers: The key to slowing healthcare cost growth while improving quality? Journal of Economic Perspectives 2011; 25; 69-92.

McGuire TG. Physician agency. In: Culyer AJ, Newhouse JP (Eds), Handbook of health economics, vol.1. North-Holland: Amsterdam; 2000. 461-536.

Miraldo M, Siciliani L, Street A. Price adjustment in the hospital sector. Journal of Health Economics 2011; 30; 829-854.

Mirrlees, JA. The optimal structure of incentives and authority within an organization. Bell Journal of Economics 1976; 7; 105-131.

Mougeot M, Naegelen F. Hospital price regulation and expenditure cap policy. Journal of Health Economics 2005; 24; 55-72.

Newhouse JP. Reimbursing health plans and health providers efficiency in production versus selection. Journal of Economic Literature 1996; 34; 1236-1263.

Nocke V, White L. Do vertical mergers facilitate upstream collusion? American Economic Review 2007; 97; 1321-1339.

Rey P, Verge T. Bilateral control with vertical contracts. RAND Journal of Economics 2004; 35; 728-746.

Richardson S. Integrating pay-for-performance into health care payment systems. Mimeo, Department of Health Care Policy, Harvard University, 2011.

Rochaix L. Information asymmetry and search in the market for physicians' services. Journal of Health Economics 1989; 8; 53-84.

Rogerson WP. Choice of treatment intensities by a nonprofit hospital under prospective pricing. Journal of Economics & Management Strategy 1994; 3; 7–51.

Scanlon D, Chernew M, McLaughlin C, Solon G. The impact of health plan report cards on managed care enrollment. Journal of Health Economics 2002; 21; 19-41.

Schlee E. The value of information about product quality. RAND Journal of Economics 1996; 27; 803-815.

Supplement: Example

We describe an example. Let $\Lambda(e,r) \equiv \gamma(e+r)$, where γ is increasing and convex.[19] Also, let $C(e,r)$ be $c(e-r)$, where c is increasing and convex, and $q = e$. Quality and cost efforts are perfect substitutes. Consider the disutility minimization in Lemma 1. The constraint in (9) is now

$$\theta e + (1-\theta)[K - c(e-r)] = I. \tag{32}$$

Moreover, we have $\Lambda_1(e,r)/\Lambda_2(e,r) = 1$, $C_1(e-r) = c'(e-r)$, and $C_2(e-r) = -c'(e-r)$. The first-order condition in (10) gives

$$1 = -\frac{\theta - (1-\theta)c'(e-r)}{-(1-\theta)c'(e-r)} \qquad \text{or} \qquad 2(1-\theta)c'(e-r) - \theta = 0. \tag{33}$$

Totally differentiate (32) and (33) with respect to e, r, and I, we have

$$e_1(I;\theta) = r_1(I;\theta) = \frac{-c''(e-r)}{-c''(e-r)(\theta-(1-\theta)c'(e-r)) - c''(e-r)(1-\theta)c'(e-r)} = \frac{1}{\theta}. \tag{34}$$

These derivatives implicitly define $e(I;\theta)$ and $r(I;\theta)$ as the unique solution of the disutility minimization for any θ, $0<\theta<1$, and I, $0<I$.

Using (34), we can write the first-order derivative of the profit function in (11) as

$$\begin{aligned} & mD'(e(I;\theta))e_1(I;\theta) - \overline{\gamma}_1(I;\theta) \\ = \;& mD'(e)e_1(I;\theta) - \gamma'(e+r)(e_1(I;\theta)+r_1(I;\theta)) \\ = \;& \frac{1}{\theta}\left[mD'(e) - 2\gamma'(e+r)\right]. \end{aligned}$$

Because D is concave and γ is convex, the profit function is strictly quasi-concave.

To implement (e^*, r^*), set θ^* and m^* to satisfy

$$\begin{aligned} \theta^* &= \frac{2c'(e^*-r^*)}{1+2c'(e^*-r^*)} \\ m^* &= 2\frac{\gamma'(e^*+r^*)}{D'(e^*)}. \end{aligned}$$

Given θ^* and m^*, the provider will choose to achieve $I^* = \theta^* e^* + (1-\theta^*)[K - c(e^*-r^*)]$ in the unique equilibrium.

[19] This is the disutility function used in Ma (1994).