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.

Keywords: information disclosure, hidden action, prospective payment, cost reimbursement, quality, cost reduction

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, Tomás Rodríguez Barraquer, and seminar participants at Boston University, European University Institute, Frisch Centre at the University of Oslo, Harvard University, Lehigh University, University of Salerno, and the 13th European Health Economic Workshop. We thank Tom Gresik for discussing with us many details of this research, and an associate editor and two 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. 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 by doing so. The efforts produce quality and cost efficiency. A higher quality results in a higher marginal cost and attracts more consumers, but a higher cost effort reduces the marginal 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 marginal 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.

---

1For our purpose, cost reimbursement is the same as conventional fee-for-service: a provider chooses medical services to supply, and receives a fee for each chosen service. This fee reflects the cost of the service and allows a profit margin. There are variations in prospective payment; it may be supplemented by outlier compensations, local-market adjustments, etc. These variations are unimportant for this paper.
The notion of a demand response requires consumers to know about quality. Most researchers, however, would agree that health care quality information can be difficult to obtain and interpret. Indeed, insurers, governments and sponsors increasingly have helped consumers find out about quality. 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. The insurer would like the provider to choose first-best quality and cost efforts, which are hidden actions to produce quality and cost. Information disclosure and payments together set up an implementation problem.

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 (the 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 be used to 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 infer about quality based on the value index. A given level of value index corresponds to some inferred quality level, generates a demand, and, hence, profits. Consumers’ belief about quality is based on the value index, not the actual quality effort. Hence, changing efforts that would not change the index would leave demand (and revenue) unaffected. It follows that the provider must choose those efforts to achieve an index with minimal disutility. Furthermore, the insurer can choose the index weight and profit margin to make the provider

---

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

3 We consider both deterministic and stochastic production of quality and cost from efforts. In each case, levels of outputs (quality and cost) would not allow direct inference of inputs (quality and cost efforts). Our stochastic production model is a standard Mirrless-Holmstrom model.

4 An “agency” explanation 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.
internalize the net social benefit of quality and cost efforts.

Our results are much more than a theoretical curiosity. Prospective payment has selection consequences. First, the provider takes a loss when treating high-cost consumers. Second, the provider earns more profit by attracting low-cost consumers. In the literature, dumping and cream skimming under prospective payment have been well recognized (see, for example, Newhouse, 1996). Third, prospective payment encourages fraudulent upcoding. For hospitals, prospective payment is implemented by the Diagnostic Related Group system: after a treatment episode, the provider reports the consumer’s primary diagnosis for payment. Differences in payment induce the provider to misreport a consumer’s diagnosis to get a higher price (Dafny, 2005). For physicians, the Medicare fee schedule encourages seeking a higher price by lying about the actual treatment intensity (Brunt, 2011).

The current theoretical and policy debates have been heavily against cost-based payments. However, cost reimbursement has none of the problems of dumping, cream skimming, and upcoding, simply because under cost reimbursement, consumer cost heterogeneity is of no concern to the provider. Cost reimbursement avoids a host of selection issues.

It has not escaped our notice that our theory relies on the provider being unable to disclose credibly quality information. If a provider was 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 Center 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 on quality reports to the general public. Furthermore, it is far from clear that a provider would report honestly 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.


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 heterogeneous 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,

\[5\text{One exception is Chalkley and Malcomson (1998). In their model, a capacity-constrained provider is motivated by altruism rather than demand response.}\]
the insurer in our model withholds information to achieve efficient quality and cost effort.

The plan of the paper is as follows. Section 2 presents the model. Section 3 sets up the information structure, the extensive forms, and studies equilibria. In Section 3, production of quality and cost efficiency by efforts is assumed to be deterministic. Section 4 uses a stochastic production function. Section 5 considers three robustness issues. We show the implementation of the first best when consumers may misinterpret the value index, and when the provider chooses many qualities to cream-skim consumers. Then we show that cost reimbursement outperforms prospective payment when a provider can practice dumping. Finally, Section 6 draws some conclusions.\footnote{In the Supplement, we have provided two detailed examples.}

2 The 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.\footnote{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.} 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 \textit{quality effort}, and \textit{cost effort}, denoted by the nonnegative variables \( e_1 \) and \( e_2 \), respectively. Quality and cost efforts are unobservable. Quality depends on effort \( e_1 \). In this and the next section, we assume deterministic quality production from effort \( e_1 \), so write quality \( q \) as a function
of effort $q(e_1)$. We assume that the function $q$ is not invertible, so multiple efforts can produce a quality level. However, we will assume that $q$ is strictly quasi-concave (so it can have the shape of an inverted U). In Section 4, we use a standard hidden-action model for stochastic quality production: we let effort $e_1$ determine a distribution of possible qualities, as in Holmstrom (1979).

The unit cost for service is $C(e_1, e_2)$ given quality effort $e_1$ and cost effort $e_2$. The function $C$ is strictly increasing in $e_1$ and strictly decreasing in $e_2$, 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 $\Lambda(e_1, e_2)$. The function $\Lambda$ 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. If the demand is $D(q(e_1))$, the provider incurs a total cost $D(q(e_1))C(e_1, e_2) + \Lambda(e_1, e_2)$.

### 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 $C(e_1, e_2)$ 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. 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 another research.

---

8 Given our assumptions on $q$ and $C$, a pair of quality and cost could be achieved by multiple quality-cost effort pairs. Therefore, it is impossible to infer exactly which quality and cost efforts have been taken without studying the provider’s incentives.

9 The model can be extended to incorporate cost heterogeneity. Dumping and cream skimming for the current model have been addressed by Ma (1994). More discussions are in Subsections 5.2 and 5.3.

10 In other words, we impose the common Inada conditions. Using subscripts to denote partial derivatives of the corresponding variables, we assume i) $C_1(e_1, e_2) \to 0$ and $A_1(e_1, e_2) \to 0$, as $e_1 \to 0$; $C_2(e_1, e_2) \to -\infty$ and $A_2(e_1, e_2) \to 0$ as $e_2 \to 0$, and ii) $C_1(e_1, e_2) \to \infty$ and $A_1(e_1, e_2) \to \infty$ as $e_1$ approaches its upper bound; $C_2(e_1, e_2) \to 0$ and $A_2(e_1, e_2) \to \infty$ as $e_2$ approaches its upper bound.

11 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 are 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.

12 In 2013, only 11% of commercial in-network payments were tied to quality or cost efficiency (Catalyst for Payment Reform, 2013).
Under prospective payment, the provider receives a fixed price $p$ per unit of delivered service. If the provider has satisfied a demand of $D(q(e_1))$, its revenue is $pD(q(e_1))$, and it bears the total cost $D(q(e_1)) \times C(e_1, e_2) + \Lambda(e_1, e_2)$. Under cost reimbursement, for each unit of delivered services the provider will be paid the variable cost $C(e_1, e_2)$ plus a margin $m$. If the provider has satisfied a demand $D(q(e_1))$, its revenue, net of variable cost is $mD(q(e_1))$, but it only bears the disutility $\Lambda(e_1, e_2)$. Prospective payment $p$ and the margin $m$ are nonnegative. The provider’s disutility due to effort, $\Lambda(e_1, e_2)$, cannot be observed and directly compensated 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 $C(e_1, e_2)$. 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 $C(e_1, e_2)$ is at most $K$. Given efforts, $K - C(e_1, e_2)$ is a measure of cost efficiency. We define a value index by $I(q, C; \theta) = \theta q(e_1) + (1 - \theta)[K - C(e_1, e_2)]$, where $0 \leq \theta \leq 1$. After observing the provider’s care quality $q(e_1)$ and variable cost $C(e_1, e_2)$, the insurer reports the value index to consumers.

If we set the weight of the value index $\theta$ to 1, then full quality information will be revealed to consumers. If $\theta$ 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 different from 1 under cost reimbursement.

### 2.3 The first best

In the first best, quality and cost efforts are contractible. The social welfare from the quality-cost effort pair $(e_1, e_2)$ is

$$B(q(e_1)) - D(q(e_1))C(e_1, e_2) - \Lambda(e_1, e_2),$$

(1)
which is social benefit less total cost. Let $(e_1^*, e_2^*)$ be the quality-cost effort pair that maximizes social welfare in (1). The following first-order conditions characterize the first best:

\begin{align*}
B'(q(e_1^*))q'(e_1^*) - D'(q(e_1^*))q'(e_1^*)C(e_1^*, e_2^*) - D(q(e_1^*))C_1(e_1^*, e_2^*) - \Lambda_1(e_1^*, e_2^*) &= 0 \quad (2) \\
-D(q(e_1^*))C_2(e_1^*, e_2^*) - \Lambda_2(e_1^*, e_2^*) &= 0, \quad (3)
\end{align*}

where we use the subscript of a function to denote the corresponding partial derivative, and the superscript prime to denote derivatives. The first-order conditions have the usual marginal 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 either operate in a competitive market or 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.\(^\text{13}\)

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

\begin{equation}
T + pD(q(e_1)) - D(q(e_1))C(e_1, e_2) - \Lambda(e_1, e_2). \quad (4)
\end{equation}

The quality and cost efforts generate a social net benefit

\begin{equation}
B(q(e_1)) - pD(q(e_1)) - T, \quad (5)
\end{equation}

which is the social benefit $B(q(e_1))$ less payments to the provider.

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

\begin{equation}
w[B(q(e_1)) - pD(q(e_1)) - T] + (1 - w)[T + pD(q(e_1)) - D(q(e_1))C(e_1, e_2) - \Lambda(e_1, e_2)], \quad (6)
\end{equation}

\(^{13}\)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 p. 916 in Baron and Myerson (1982).
where \(0.5 < w \leq 1\). 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 \(e_1\) and \(e_2\) to maximize profit (4). The following proposition is adapted from Ma (1994), and stated with its proof omitted:

**Proposition 1**: By choosing \(p^* = \frac{B'(q(e^*_1))}{D'(q(e^*_1))}\) and a suitable transfer \(T^*\), the insurer implements the first-best quality effort \(e^*_1\) and cost effort \(e^*_2\).

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 perfect-Bayesian equilibria of the following extensive-form game:

**Stage 1**: The insurer sets the transfer \(T\), the margin \(m\), and the weight \(\theta\) in the value index. The insurer also commits to reimbursing the provider’s operating cost.

**Stage 2**: The provider chooses unobservable quality and cost efforts, respectively \(e_1\) and \(e_2\).

**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), form beliefs about quality and cost efforts, and decide on the quantity of services to obtain.

Consumers do not observe the provider’s quality and cost efforts, and form beliefs about them based on the value index. Suppose that in an equilibrium, the provider chooses quality-cost effort pair \((\hat{e}_1, \hat{e}_2)\). The value index becomes \(\hat{I} \equiv \theta q(\hat{e}_1) + (1 - \theta)[K - C(\hat{e}_1, \hat{e}_2)]\). Then in equilibrium, consumers must correctly infer from \(\hat{I}\) that quality is \(q(\hat{e}_1)\), and their demand will be \(D(q(\hat{e}_1))\).
There are many ways that the provider can achieve this particular level of the value index $\hat{I}$. In fact, any combination of quality and cost efforts, $e_1$ and $e_2$ satisfying $\theta q(e_1) + (1 - \theta)[K - C(e_1, e_2)] = \hat{I}$ can do so. If the provider deviates from efforts $(\hat{e}_1, \hat{e}_2)$ to $(e_1, e_2)$, consumers’ beliefs about quality remain unchanged at $q(\hat{e}_1)$ (because they observe the same value index level $\hat{I}$). However the provider’s payoff after the deviation, is

$$T + mD(q(\hat{e}_1)) - \Lambda(e_1, e_2).$$

In a perfect-Bayesian equilibrium, such a deviation must not yield a higher payoff, so any equilibrium efforts $(\hat{e}_1, \hat{e}_2)$ must maximize (7) subject to $\theta q(e_1) + (1 - \theta)[K - C(e_1, e_2)] = \hat{I}$.

**Lemma 1 :** Equilibrium quality and cost efforts $(\hat{e}_1, \hat{e}_2)$ must solve

$$\min_{(e_1, e_2)} \Lambda(e_1, e_2)$$

subject to $\theta q(e_1) + (1 - \theta)[K - C(e_1, e_2)] = \hat{I} \equiv \theta q(\hat{e}_1) + (1 - \theta)[K - C(\hat{e}_1, \hat{e}_2)]$.

Hence, $(\hat{e}_1, \hat{e}_2)$ satisfies

$$\frac{\Lambda_1(\hat{e}_1, \hat{e}_2)}{\Lambda_2(\hat{e}_1, \hat{e}_2)} = -\frac{\theta q'(\hat{e}_1) - (1 - \theta)C_1(\hat{e}_1, \hat{e}_2)}{(1 - \theta)C_2(\hat{e}_1, \hat{e}_2)}.$$  

(8)

For any out-of-equilibrium level of the index $\hat{I}$, $\hat{I} \neq \hat{I}$, consumers’ beliefs $(\hat{e}_1, \hat{e}_2)$ must also satisfy (8) if they believe that the provider would never choose a dominated strategy.

Lemma 1 says that equilibrium quality and cost efforts must minimize their disutility for achieving any level of the value index. For a given index level $\hat{I}$, consumers’ belief about $q(\hat{e}_1)$ is constant, and so is the revenue $T + mD(q(\hat{e}_1))$. The maximization of (7) is the same as the minimization of $\Lambda(e_1, e_2)$. The condition in (8) gives the optimality condition for the constrained minimization of $\Lambda(e_1, e_2)$. The left-hand side of (8) 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-cost weight $\theta$. For a given weight $\theta$ and the equilibrium level of the index $\hat{I}$, equilibrium efforts are unique. This follows from the concavity of $q$, convexity of $\Lambda$, and the convexity of the constrained set.

Lemma 1 also specifies consumer beliefs off the equilibrium path. It says that for any value index $\hat{I}$, not just the one chosen by the provider in equilibrium, consumers believe that quality and cost efforts have
been chosen to minimize the disutility. This is a weak belief restriction. The provider’s strategy of choosing quality and cost efforts that do not minimize disutility for any value index is strictly dominated by one that does. Given the beliefs, for any $I(q, C; \theta)$, condition (8) defines a function of $e_2$ in $e_1$, say $\tilde{e}_2(e_1; \theta)$.

Even when unit variable costs, $C(e_1, e_2)$, 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. Given an index, consumers realize that a provider chooses efforts to maximize its payoff, so their beliefs about efforts follow Lemma 1.

The equilibrium efforts can be illustrated in Figure 1. Consider the objective function $\Lambda(e_1, e_2)$ in Lemma 1. Because $\Lambda$ is convex, its lower contour sets, $\{(e_1, e_2) : \Lambda(e_1, e_2) \leq \lambda\}$, are convex, so in Figure 1 we show an iso-disutility line that is concave to the origin. Consider the constraint in Lemma 1. Because $q$ is concave and $C$ is convex, the upper contour sets of the iso-index line, $\{(e_1, e_2) : q(e_1) + (1 - \theta)[K - C(e_1, e_2)] \geq \mathcal{T}\}$, are convex. In Figure 1, the iso-index lines, at levels $\mathcal{T}_1$ and $\mathcal{T}_2$, $\mathcal{T}_1 < \mathcal{T}_2$, are the circular lines. An equilibrium is the tangency point between the iso-index and iso-disutility lines. As the level of the value index changes, condition (8) continues to define a unique pair of efforts for every level of the value index $I$. In Figure 1, the dotted “expansion path” plots these quality-cost effort pairs. The expansion path in fact is the function $\tilde{e}_2(e_1; \theta)$. Changing the index weight $\theta$ corresponds to changing the entire map of the iso-index lines, and the entire function $\tilde{e}_2(e_1; \theta)$ may rotate, shift, or do both.

To implement the first best, first set $\theta$ to $\theta^*$ where

$$
\frac{\Lambda_1(e_1^*, e_2^*)}{\Lambda_2(e_1^*, e_1^*)} = -\frac{\theta^* q(e_1^*) - (1 - \theta^*)C_1(e_1^*, e_2^*)}{(1 - \theta^*)C_2(e_1^*, e_1^*)}.
$$

This guarantees that the first best is a potential equilibrium. Because the right-hand side of (9) is monotone increasing in $\theta^*$, there is a unique $\theta^*$ that satisfies it. Next, given $\theta = \theta^*$, the provider’s profit is $mD(q(e_1)) - \Lambda(e_1, e_2)$. We replace $e_2$ by $\tilde{e}_2(e_1; \theta^*)$ to obtain the reduced-form profit function

$$
mD(q(e_1)) - \Lambda(e_1, \tilde{e}_2(e_1; \theta^*)).
$$

We assume that for any $m$, the function (10) is quasi-concave in $e_1$. The derivative of (10) with respect to $e_1$ is

$$
mD'(q(e_1))q'(e_1) - \Lambda_1(e_1, \tilde{e}_2(e_1; \theta^*)) - \Lambda_2(e_1, \tilde{e}_2(e_1; \theta^*)) \times \tilde{e}_2'(e_1; \theta^*).
$$
Finally, we set $m$ at $m^*$ where:

\[ m^* D'(q(e_1^*))q'(e_1^*) - \Lambda_1(e_1^*, \tilde{e}_2(e_1^*; \theta^*)) - \Lambda_2(e_1^*, \tilde{e}_2(e_1^*; \theta^*)) \times \tilde{e}_2'(e_1^*; \theta^*) = 0. \]  

(11)

**Proposition 2**: Suppose that the profit function (10) is quasi-concave, then the insurer implements the first-best efforts $(e_1^*, e_2^*)$ by setting the weight of the value index $\theta^*$ given by (9), the cost margin $m^*$ given by (11), and a suitable transfer $T$.

Proposition 2 can be explained as follows. By Lemma 1, the insurer can choose a weight so that the first-best quality and cost efforts minimize disutility $\Lambda(e_1, e_2)$. This is the weight $\theta^*$ defined by (9) (and expressed explicitly in (25) in the Appendix). The weight $\theta^*$ ensures that the first best is on the dotted expansion path in Figure 2, so it is a candidate for an equilibrium.

Next, the margin $m$ is to be chosen so that the first best indeed is the provider’s equilibrium choice. Respecting the constraint (8) in Lemma 1, the provider chooses $e_1$ and $e_2$ to maximize profit. Consider the iso-profit line obtained by setting $T + mD(q(e_1)) - \Lambda(e_1, e_2)$ to a constant. This implicitly defines a function $e_2$ in terms of $e_1$. The derivative of the iso-profit line is $[mD'(q(e_1))q'(e_1) - \Lambda_1(e_1, e_2)]/\Lambda_2(e_1, e_2)$. 

![Figure 1: Disutility-minimizing quality and cost efforts](image-url)
The derivative is positive for low $e_1$'s, where marginal profit from quality effort is positive and profit is kept constant by raising cost effort, but turns negative at high $e_1$'s where marginal profit from quality effort is negative. Hence, the function assumes the shape of an inverted U, like the solid curve in Figure 2. Points below the iso-profit line yield higher profits to the provider. A higher value of $m$ shifts this function upward and also makes the iso-profit line steeper. At $m = m^*$, the tangency point between the iso-profit line and the expansion path occurs at the first best $(e_1^*, e_2^*)$.

The optimal margin $m^*$ in (11) illustrates how information and payment policies have to be coordinated. Consumers are only interested in quality. Hence the provider’s marginal profit $mD'(q(e_1))q'(e_1)$ is increasing in $e_1$ only. However, the value index forces the provider to exert both quality and cost efforts. When the provider raises quality effort by one unit, cost minimization requires the adjustment of cost effort by $c_2^*(e_1; \theta^*)$ as well. In equilibrium, consumers correctly infer quality from the value index and respond accordingly. The optimal $m^*$ is set such that the provider’s marginal profit $m^*D'(q(e_1^*))q'(e_1^*)$ equates the sum of marginal disutilities $\Lambda_1(e_1^*, e_2^*) + \Lambda_2(e_1^*, e_2^*) \times c_2^*(e_1^*; \theta^*)$. 

Figure 2: Implementing the first-best quality and cost efforts
Socially optimal information disclosure requires the choice of appropriate weights for quality and cost. These weights must take into account the provider’s strategic response. A combination of value-index weight and the margin over cost reimbursement implements the first best. This is the point of Proposition 2.

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 disclose credibly 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 credibly. 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 Stochastic quality and value index

In the previous section, the provider’s quality effort produces quality in a deterministic fashion (although efforts cannot be recovered from quality and cost). We now consider stochastic quality production by effort. We use the standard hidden-action model of Mirrlees (1976) and Holmstrom (1979). Quality effort \( e_1 \) determines a distribution on quality \( q \), a random variable defined on \( \mathbb{R}_+ \), the positive real numbers.\(^{14}\) Given any quality level \( q \), the insurer cannot rule out any quality effort level \( e_1 \). We write the density function of \( q \) as \( f(q|e_1) \), and assume that the expected demand is strictly increasing and strictly concave in quality effort. That is, \( \int_{\mathbb{R}_+} D(q)f(q|e_1)\,dq \) is strictly increasing and concave in \( e_1 \).

\(^{14}\)The full-support assumption in 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.
We first provide the notation for the first best. The social welfare expression in (1) can now be written as
\[ \int_{\mathbb{R}^+} [B(q) - D(q)C(e_1, e_2)] f(q|e_1) dq - \Lambda(e_1, e_2). \]  
(12)
Here, the integral is the expected social benefit less variable costs while the remaining term is the effort disutility.

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)):
\[ \int_{\mathbb{R}^+} \left\{ [B(q) - D(q)C(e_1^*, e_2^*)] \frac{\partial f(q|e_1^*)}{\partial e_1} \right\} dq - \Lambda_1(e_1^*, e_2^*) = 0 \]
\[ \int_{\mathbb{R}^+} [-D(q)C_2(e_1^*, e_2^*)] f(q|e_1^*) dq - \Lambda_2(e_1^*, e_2^*) = 0. \]
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
\[ p^* = \frac{\int_{\mathbb{R}^+} B(q) \frac{\partial f(q|e_1^*)}{\partial e_1} dq}{\int_{\mathbb{R}^+} D(q) \frac{\partial f(q|e_1^*)}{\partial e_1} dq} \]
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 \( C(e_1, e_2) \). Because \( q \) is stochastic, so is the index \( I(q, C; \theta) \equiv \theta q + (1 - \theta)[K - C] \). 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 \( C(e_1, e_2) \). The level of the value index is all they observe.\(^{15}\)

Consider an equilibrium in which the provider chooses effort pair \((\hat{e}_1, \hat{e}_2)\). In this equilibrium, consumers believe that unit cost is \( C(\hat{e}_1, \hat{e}_2) \). Quality \( q \) will be drawn according to density \( f(q|\hat{e}_1) \). When the level of value index is \( I \), consumers believe that quality \( \hat{q} \) satisfies \( \theta \hat{q} + (1 - \theta)[K - C(\hat{e}_1, \hat{e}_2)] = I \). From this, the

\(^{15}\)Quality can be any real number, and the value of \( K - C(e_1, e_2) \) is assumed to be always positive. So for some effort pairs, the range of the index may be strictly positive. We specify that should the value index have a level outside this range, consumers would believe that the quality was 0.
inferred quality is
\[ \hat{q}(I|(\hat{e}_1, \hat{e}_2)) = \frac{I - (1 - \theta)[K - C(\hat{e}_1, \hat{e}_2)]}{\theta}, \]  
(13)

where we have emphasized that the inference rule \( \hat{q} \) depends on the value index and equilibrium efforts.

Consider a quality level, say \( \tilde{q} \). Under equilibrium effort \((\hat{e}_1, \hat{e}_2)\), when \( \tilde{q} \) is realized, then \( \hat{q}(I|(\hat{e}_1, \hat{e}_2)) = \tilde{q} \).

Suppose that the provider deviates from \((\hat{e}_1, \hat{e}_2)\) to \((e_1, e_2)\). The unit cost becomes \( C(e_1, e_2) \), and this will be observed by the insurer (although consumers continue to believe that unit cost is \( C(\hat{e}_1, \hat{e}_2) \)). When the same quality \( \tilde{q} \) is realized, the index changes to \( \tilde{I} \equiv \theta \tilde{q} + (1 - \theta)[K - C(e_1, e_2)] \). Because consumers continue to believe that efforts are \((\hat{e}_1, \hat{e}_2)\), they use the inference rule in (13), and believe that the quality is
\[ \hat{q}(\tilde{I} | (\hat{e}_1, \hat{e}_2)) = \frac{\tilde{I} - (1 - \theta)[K - C(\hat{e}_1, \hat{e}_2)]}{\theta} = \frac{\theta \tilde{q} + (1 - \theta)[K - C(e_1, e_2)] - (1 - \theta)[K - C(\hat{e}_1, \hat{e}_2)]}{\theta} = \tilde{q} + \frac{1 - \theta}{\theta} [C(\hat{e}_1, \hat{e}_2) - C(e_1, e_2)] \neq \tilde{q}. \]  
(14)

The point is that if the provider reduces the unit cost, the quality perceived by consumers becomes higher. In (14), the inferred quality \( \hat{q}(\tilde{I} | (\hat{e}_1, \hat{e}_2)) \) is larger than the realized quality \( \tilde{q} \) if and only if \( C(\hat{e}_1, \hat{e}_2) - C(e_1, e_2) > 0 \). This is the basic incentive for the provider to expend cost effort. More important, the insurer can influence this incentive by choosing \( \theta \): a smaller \( \theta \) raises \( \frac{1 - \theta}{\theta} \) in (14). This implies a larger difference between the quality observed by the insurer and the quality inferred by consumers.

In an equilibrium, consumers must not be misled, so the equilibrium effort \((\hat{e}_1, \hat{e}_2)\) must be one which the provider finds unprofitable to deviate from.\(^{16}\) Suppose that the provider deviates from the equilibrium \((\hat{e}_1, \hat{e}_2)\). The expected payoff from another effort pair \((e_1, e_2)\) is
\[ m \int_{\mathbb{R}_+} D \left( q + \frac{1 - \theta}{\theta} [C(\hat{e}_1, \hat{e}_2) - C(e_1, e_2)] \right) f(q|e_1) dq - \Lambda(e_1, e_2). \]  
(15)

In (15), we have used the inference rule (14) when the provider deviates from \((\hat{e}_1, \hat{e}_2)\) to express the demand; the integral is the expected revenue when the margin is set at \( m \). We assume that the expected payoff (15) is quasi-concave. Effort pair \((\hat{e}_1, \hat{e}_2)\) is an equilibrium if it maximizes (15).

\(^{16}\)Because the value index is stochastic, it is not meaningful to analyze disutility-minimizing effort pairs given an index level, as in the previous section.
We obtain the first-order conditions with respect to efforts (these are equations (29) and (30) in the appendix). To derive an equilibrium, we then set \((\tilde{e}_1, \tilde{e}_2)\) in the first-order conditions to \((\tilde{e}_1, \tilde{e}_2)\) to get, respectively,

\[
\int_{\mathbb{R}^+} m \left\{ D(q) \left[ \frac{\partial f(q|\tilde{e}_1)}{\partial \tilde{e}_1} \right] - D'(q) C_1(\tilde{e}_1, \tilde{e}_2) \frac{1 - \frac{\theta}{\theta}}{\theta} f(q|\tilde{e}_1) \right\} dq - \Lambda_1(\tilde{e}_1, \tilde{e}_2) = 0 \quad (16)
\]

\[
\int_{\mathbb{R}^+} -m \left[ D'(q) C_2(\tilde{e}_1, \tilde{e}_2) \frac{1 - \frac{\theta}{\theta}}{\theta} \right] f(q|\tilde{e}_1) dq - \Lambda_2(\tilde{e}_1, \tilde{e}_2) = 0. \quad (17)
\]

**Proposition 3**: Suppose that the profit function (15) is quasi-concave, then the insurer implements the first-best efforts \((e_1^*, e_2^*)\) by setting the weight of the value index \(\theta^*\) and margin \(m^*\) to satisfy (16) and (17) at \((\hat{e}_1, \hat{e}_2) = (e_1^*, e_2^*)\) together with a suitable transfer \(T\).

The result here contrasts with the sufficient-statistics 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 that are based on garbling of informative signals are suboptimal. Here, the insurer purposefully garbles the information about quality with cost information, which is irrelevant to consumers. The value index alters demand response in quality. However, the value index can be improved by cost effort. The insurer optimally mixes quality and cost information, so cost effort can indirectly affect demand. Garbling of information acts as an incentive instrument.

5 Consumer rationality, multiple qualities, and selection

In this section we discuss a number of issues, and will use both the deterministic and stochastic quality effort models.

5.1 Consumer rationality

It may appear that Lemma 1 relies on consumers being fully rational. In fact, Lemma 1 stems from the firm maximizing profits. Consider an arbitrary inference rule (such as consumers naively believing that quality is always 50% of the value index). If the value index takes a value of \(I\), assume that consumers believe that the quality is \(\Psi(I)\), where \(\Psi\) is an increasing and differentiable function. In the deterministic quality model, the
provider’s choices can be separated into disutility-minimization and profit-maximization components. We show that the same choice of value index weight as in the rational consumer case will be used. However, a different margin is necessary for the first-best implementation.

Under cost reimbursement, given a margin \( m \) and an index level \( I \), the provider’s profit is \( mD(\Psi(I)) - \Lambda(e_1, e_2) \). Equilibrium quality and cost efforts \((\hat{e}_1, \hat{e}_2)\) must solve

\[
\begin{align*}
\max_{e_1, e_2} & \quad mD(\Psi(I)) - \Lambda(e_1, e_2) \\
\text{subject to} & \quad \theta q(e_1) + (1 - \theta)[K - C(e_1, e_2)] = \hat{I} \equiv \theta q(\hat{e}_1) + (1 - \theta)[K - C(\hat{e}_1, \hat{e}_2)].
\end{align*}
\]

Because the index level \( \hat{I} \) is fixed in this constrained maximization program, the first-order conditions with respect to \( e_1 \) and \( e_2 \) are:

\[
\begin{align*}
-\Lambda_1(e_1, e_2) - \lambda[\theta q'(e_1) - (1 - \theta)C_1(e_1, e_2)] &= 0 \\
-\Lambda_2(e_1, e_2) + \lambda(1 - \theta)C_2(e_1, e_2) &= 0
\end{align*}
\]

where \( \lambda \) is the multiplier. Setting \((e_1, e_2)\) to \((\hat{e}_1, \hat{e}_2)\), we obtain

\[
\frac{\Lambda_1(\hat{e}_1, \hat{e}_2)}{\Lambda_2(\hat{e}_1, \hat{e}_2)} = \frac{\theta q'(\hat{e}_1) - (1 - \theta)C_1(\hat{e}_1, \hat{e}_2)}{(1 - \theta)C_2(\hat{e}_1, \hat{e}_2)},
\]

which is (8) in Lemma 1.

As long as the provider minimizes the disutility due to quality and cost efforts, the value-index weight determines how the provider must trade off quality against cost efforts, given any consumer inference. The value index therefore incentivizes the provider to reduce cost. The same function \( \bar{e}_2(e_1; \theta) \) (from (8)) defines a function of possible equilibrium effort pairs.

We can follow a similar procedure laid out before the presentation of Proposition 2. For the implementation of the first best \((e_1^*, e_2^*)\), the insurer sets \( \theta \) at \( \theta^* \) in (9). Given the inference rule \( \Psi(I) \), demand will be \( D(\Psi(I)) \), where \( I \equiv \theta q(e_1) + (1 - \theta)[K - C(e_1, \bar{e}_2(e_1; \theta^*))]. \) Next, for a given margin \( m \), profit is

\[
mD(\Psi(I)) - \Lambda(e_1, \bar{e}_2(e_1; \theta^*)). \tag{18}
\]

For the implementation of \((e_1^*, e_2^*)\), it is sufficient that the provider finds it optimal to choose \( e_1^* \) to maximize profit (18). Earlier we have assumed that (10) is quasi-concave, so if the inference rule \( \Psi \) is
increasing and concave, then (18) is quasi-concave. For the implementation of the first best, the insurer chooses \( m = m^* \) at which the first-order derivative of (18) vanishes at \( e_1^* \). This margin will be different from the one in Proposition 2.

We now turn to the stochastic quality model. By the same notation, consumers use the rule \( \Psi(I) \) for quality when they observe index \( I \). Given the index weight \( \theta \), If the provider chooses \((e_1, e_2)\), \( q \) will be drawn according to \( f(q|e_1) \), and \( I \equiv \theta q + (1 - \theta)[K - C(e_1, e_2)] \). The demand at realized quality \( q \) is \( D(\Psi(\theta q + (1 - \theta)[K - C(e_1, e_2)]) \). For a given margin \( m \), expected payoff becomes

\[
m \int_{\mathbb{R}_+} D(\Psi(\theta q + (1 - \theta)[K - C(e_1, e_2)]) f(q|e_1) dq - \Lambda(e_1, e_2). \tag{19}
\]

The key difference from the model in Section 4 is that consumers’ inferences here are incorrect even when the provider chooses the equilibrium efforts.\(^{17}\) In any case, if the insurer can find weight \( \theta^* \) and margin \( m^* \) which makes \( e_1^* \) and \( e_2^* \) maximize (19), then the first best is implemented.\(^{18}\)

5.2 Multiple qualities and cream skimming

Health services are often multidimensional. Suppose now there are two service qualities, \( q_A \) and \( q_B \). Let \( e_{1A} \) and \( e_{1B} \) be two efforts. To save on notation and for ease of exposition, we simply let \( (q_A, q_B) = (e_{1A}, e_{1B}) \). We extend the notation for demand, social benefit, variable cost, and disutility in the obvious way: \( D(q_A, q_B), B(q_A, q_B), C(q_A, q_B, e_2), \Lambda(q_A, q_B, e_2) \). We also maintain the corresponding concavity and convexity assumptions.

The social welfare is now

\[
B(q_A, q_B) - D(q_A, q_B)C(q_A, q_B, e_2) - \Lambda(q_A, q_B, e_2) \tag{20}
\]

\(^{17}\)Since our focus is on provider’s incentives, we define the first best in terms of the provider’s choice of quality and cost efforts. When consumers’ inferences are incorrect, the insurer can adjust demand-side incentives such as copayment and co-insurance rates to correct over-consumption or under-consumption.

\(^{18}\)At this level of generality for the inference rule \( \Psi \), we cannot write down a condition for this.
Let \( q_A^*, q_B^*, e_2^* \) be the first-best qualities and cost effort, those that maximize (20). Under prospective payment with transfer \( T \), price \( p \), and complete quality-information disclosure, the provider’s profit is

\[
T + pD(q_A, q_B) - D(q_A, q_B)C(q_A, q_B, e_2) - \Lambda(q_A, q_B, e_2).
\]

If the insurer discloses information of both \( q_A \) and \( q_B \), a prospective price can be chosen to implement the first best if and only if

\[
\frac{B_1(q_A^*, q_B^*)}{D_1(q_A^*, q_B^*)} = \frac{B_2(q_A^*, q_B^*)}{D_2(q_A^*, q_B^*)},
\]

(which is also the prospective price). This result is obtained by comparing the first-order conditions for the first best (as in footnote 19) and for the provider’s profit maximization (as in Proposition 1).

With a single quality, a single prospective price is sufficient for the first best, as in Proposition 1. With multiple qualities, a single prospective price is insufficient generally. 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 (21) imposes the equality of these marginal contributions. To see this, rearrange (21) to

\[
\frac{B_1(q_A^*, q_B^*)}{B_2(q_A^*, q_B^*)} = \frac{pD_1(q_A^*, q_B^*)}{pD_2(q_A^*, q_B^*)},
\]

which says that the marginal rates of substitution between the two qualities have to be identical in the social benefit function and in the marginal revenue function.

The insurer can still implement the first best if it discloses 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 \( 0 \leq \phi \leq 1 \). 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 \( q_B \).

---

\footnote{They are characterized by the first-order conditions:

\[
B_1(q_A^*, q_B^*) - D_1(q_A^*, q_B^*)C_1(q_A^*, q_B^*, e_2^*) - D(q_A^*, q_B^*)C_1(q_A^*, q_B^*, e_2^*) - \Lambda_1(q_A^*, q_B^*, e_2^*) = 0,
\]

\[
B_2(q_A^*, q_B^*) - D_2(q_A^*, q_B^*)C_2(q_A^*, q_B^*, e_2^*) - D(q_A^*, q_B^*)C_2(q_A^*, q_B^*, e_2^*) - \Lambda_2(q_A^*, q_B^*, e_2^*) = 0,
\]

\[
-D(q_A^*, q_B^*)C_3(q_A^*, q_B^*, e_2^*) - \Lambda_3(q_A^*, q_B^*, e_2^*) = 0,
\]

which have the usual interpretations.}
Analogous to Lemma 1, the equilibrium inference must be qualities $\hat{q}_A$ and $\hat{q}_B$ which solve

$$\max_{\hat{q}_A, \hat{q}_B, e_2} T + pD(\hat{q}_A, \hat{q}_B) - D(\hat{q}_A, \hat{q}_B)C(q_A, q_B, e_2) - \Lambda(q_A, q_B, e_2)$$

subject to $\phi q_A + (1 - \phi) q_B = \hat{J} \equiv \phi \hat{q}_A + (1 - \phi) \hat{q}_B$. \hspace{1cm} (23)

Any choice of qualities that achieve the quality index level $\hat{J}$ 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 $\phi$ therefore can implement the first-best marginal rate of substitution between the two quality efforts, as in (22). 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, $\theta_A$ and $\theta_B$, are positive and $\theta_A + \theta_B \leq 1$. Under cost reimbursement, equilibrium qualities and cost effort must minimize the disutility. Any equilibrium $\hat{q}_A, \hat{q}_B$ and $\hat{e}_2$ solve

$$\max_{\hat{q}_A, \hat{q}_B, e_2} T + mD(\hat{q}_A, \hat{q}_B) - \Lambda(q_A, q_B, e_2)$$

subject to $\theta_A q_A + \theta_B q_B + (1 - \theta_A - \theta_B)[K - C(q_A, q_B, e_2)]$

$$= \hat{I} \equiv \theta_A \hat{q}_A + \theta_B \hat{q}_B + (1 - \theta_A - \theta_B)[K - C(\hat{q}_A, \hat{q}_B, \hat{e}_2)].$$

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.

The multiple-quality model can be easily modified to study cream skimming. Here, cream skimming refers to the use of different quality levels to discriminate against consumers. Suppose that there are two types of consumers, types $A$ and $B$. Let $q_A$ and $q_B$ be the respective (unidimensional) quality levels the provider chooses for the type-$A$ and type-$B$ consumers. The service unit costs for type-$A$ and type-$B$ consumers are $C(q_A)$ and $\beta C(q_B)$, respectively, where $1 < \beta$, so type-$B$ consumers are less profitable. Let the corresponding demands from the two types be $D(q_A)$ and $D(q_B)$.

Under prospective payment, a uniform quality $q = q_A = q_B$ is implementable only if the insurer can
observe $q_A$ and $q_B$ separately, and construct a quality index as in (23). But under cost reimbursement, the unit costs $C(q_A)$ and $\beta C(q_B)$ are fully reimbursed. The provider would not have an incentive to cream-skim consumers even if the insurer cannot perfectly differentiate $q_A$ and $q_B$.

5.3 Dumping

In this subsection, we discuss dumping under cost reimbursement and prospective payment. We continue to assume that the insurer seeks to implement a given quality-cost effort pair. We first extend the model of deterministic quality 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 $g(c|e_1, e_2)$ denote the density of $c$, given effort pair $(e_1, e_2)$. Now we let $C(e_1, e_2) \equiv \int_{0 \leq c \leq K} c g(c|e_1, e_2)dc$ denote the average cost. This is what is observed by the insurer for the construction of the value index in cost reimbursement. The extensive form is the same as in Subsection 3.2 (except that the average, not the actual, cost is reported).

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 high-cost consumers. However, by expending cost effort $e_2$, the provider changes the average cost $C(e_1, e_2)$, 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. Cost reimbursement with value index performs better.

6 Conclusion

Prospective payment and cost reimbursement are common payment mechanisms 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, by optimally choosing the content of public report, how an insurer can make the two payment systems implement identical quality and cost incentives. The insurer samples or audits patient cases to obtain cost estimates. We assume that sampling or auditing are sufficiently accurate to estimate the average cost. 
incentives. Our results are robust to stochastic quality production, consumer misinterpreting the value index, and multiple qualities. Because prospective payment is known to create dumping, cream skimming, and up-coding incentives, our result is particularly relevant when patient selection and provider misreporting problems are serious.

The main point here can be interpreted as using information 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). This 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: Suppose that equilibrium quality and cost efforts \((\tilde{e}_1, \tilde{e}_2)\) achieve the level of value index \(\tilde{I} = \theta q(\tilde{e}_1) + (1 - \theta)[K - C(\tilde{e}_1, \tilde{e}_2)]\). Consumers believe, in equilibrium, that the provider’s quality effort is \(\tilde{e}_1\), so the demand is \(D(q(\tilde{e}_1))\). Now suppose that the provider deviates to any other pair of efforts such that the same value index level is achieved. That is, suppose the provider deviates to \(e_1\) and \(e_2\) where \(\theta q(e_1) + (1 - \theta)[K - C(e_1, e_2)] = \tilde{I}\), then consumers must continue to believe that the quality effort is \(\tilde{e}_1\). The provider’s profit is \(T + mD(q(\tilde{e}_1)) - \Lambda(e_1, e_2)\). By definition of an equilibrium, \(T + mD(q(\tilde{e}_1)) - \Lambda(e_1, e_2) \leq T + mD(q(\tilde{e}_1)) - \Lambda(\tilde{e}_1, \tilde{e}_2)\). Because \(\Lambda\) is strictly convex, the inequality is strict if and only if \((e_1, e_2) \neq (\tilde{e}_1, \tilde{e}_2)\). Maximizing (7) subject to \(\theta q(e_1) + (1 - \theta)[K - C(e_1, e_2)] = \tilde{I}\), we obtain the first-order condition (8).

Now suppose the reported index level is \(\bar{I}, \bar{I} \neq \tilde{I}\). Let \((\bar{e}_1, \bar{e}_2)\) be the consumers’ beliefs about efforts for this level \(\bar{I}\), so the demand is \(D(q(\bar{e}_1))\). Let \((e_1^\dagger, e_2^\dagger)\) maximize profit \(T + mD(q(\bar{e}_1)) - \Lambda(e_1, e_2)\) subject to \(\theta q(e_1) + (1 - \theta)[K - C(e_1, e_2)] = \bar{I}\). For any belief, any effort pair \((e_1, e_2)\) is dominated by \((e_1^\dagger, e_2^\dagger)\).

If consumers never believe that the provider would choose a dominated strategy, we must have \((\bar{e}_1, \bar{e}_2) = (e_1^\dagger, e_2^\dagger)\). Therefore, the first-order condition (8) also defines the unique profit-maximizing pair of efforts at any out-of-equilibrium index level \(\bar{I}\).

Proof of Proposition 2: First, by Lemma 1, to ensure that the first-best quality and cost efforts \((e_1^*, e_2^*)\) can be an equilibrium choice by the provider, the value of the weight must satisfy (9). Solving this equation for \(\theta^*\) yields

\[
\theta^* = \frac{\frac{\Lambda_1(e_1^*, e_2^*) - C_1(e_1^*, e_2^*)}{\Lambda_2(e_1^*, e_2^*)} - \frac{C_1(e_1^*, e_2^*)}{C_2(e_1^*, e_2^*)}}{\frac{1}{C_2(e_1^*, e_2^*)} - \frac{C_1(e_1^*, e_2^*)}{C_2(e_1^*, e_2^*)}} \quad (25)
\]

where \(0 < \theta^* < 1\) because \(C_2\) is negative and \(\Lambda_1, \Lambda_2, C_1\) are positive. For the rest of the proof, \(\theta\) is set at this value.

Second, again from Lemma 1, we use (8) to define implicitly \(e_2\) as a function of \(e_1\) at \(\theta^*\). Upon simplifi-
For any given $e_2 = \bar{e}_2(e_1; \theta^*)$, with
\[
e_2'(e_1; \theta^*) = \frac{\Lambda_1(e_1^*, e_2^*) - \Lambda_2(e_1^*, e_2^*) + C_1(e_1^*, e_2^*) - C_1(e_1^*, e_2^*)}{\Lambda_1(e_1^*, e_2^*)} - \frac{C_2(e_1^*, e_2^*) - C_2(e_1^*, e_2^*)}{\Lambda_2(e_1^*, e_2^*)}.
\] (26)

For any given $m$, the provider's objective can now be regarded as a choice of $e_1$ that maximizes
\[
T + mD(q(e_1)) - \Lambda(e_1, (\bar{e}_2(e_1; \theta^*))).
\] (27)

By assumption, the profit function (27) is quasi-concave, so a unique $e_1$ maximizes profit at any given $m > 0$.

The first-order condition is
\[
m = \frac{\Lambda_1(e_1, \bar{e}_2(e_1; \theta^*)) + \Lambda_2(e_1, \bar{e}_2(e_1; \theta^*)) \times \bar{e}_2'(e_1; \theta^*)}{D'(q(e_1))q'(e_1)}.
\] (28)

Condition (28) characterizes the relationship between the profit-maximizing $e_1$ and $m$ for any $e_1$ in the nonnegative bounded set. The value of $m^*$ in the Proposition is the solution for $m$ in (28) at $e_1 = e_1^*$ and $\theta = \theta^*$. Given $m^*$ and $\theta^*$, the provider's profit is maximized at $(e_1^*, e_2^*)$.

Finally, the value of the transfer $T$ is chosen such that $T + m^*D(q(e_1^*)) - \Lambda(e_1^*, e_2^*) = 0$.

**Proof of Proposition 3:** Differentiate (15) with respect to $e_1$ and $e_2$, we have
\[
\int_{\mathbb{R}_+} \left\{ \begin{array}{l}
D \left( q + \frac{1 - \theta}{\theta} [C(\bar{e}_1, \bar{e}_2) - C(e_1, e_2)] \right) \left[ \frac{\partial f(q(e_1))}{\partial e_1} \right] \\
- D' \left( q + \frac{1 - \theta}{\theta} [C(\bar{e}_1, \bar{e}_2) - C(e_1, e_2)] \right) C_1(e_1, e_2) \frac{1 - \theta}{\theta} f(q(e_1))
\end{array} \right\} dq - \Lambda_1(e_1, e_2) = 0
\] (29)

\[
\int_{\mathbb{R}_+} -m \left[ D' \left( q + \frac{1 - \theta}{\theta} [C(\bar{e}_1, \bar{e}_2) - C(e_1, e_2)] \right) C_2(e_1, e_2) \frac{1 - \theta}{\theta} f(q(e_1)) \right] dq - \Lambda_2(e_1, e_2) = 0.
\] (30)

Set the $(e_1, e_2)$ in the two first-order derivatives to $(\bar{e}_1, \bar{e}_2)$, we obtain the equilibrium conditions (16) and (17).

By assumption, the profit function (15) is quasi-concave, hence a unique pair of $(e_1, e_2)$ maximizes profit at any given pair of nonnegative $\theta$ and $m$. The values of $\theta^*$ and $m^*$ are the solutions of $\theta$ and $m$ in (16) and (17) at $(e_1^*, e_2^*)$. Solving for the two equilibrium conditions, we have
\[
\theta^* = \frac{\Lambda_1(e_1^*, e_2^*) - \Lambda_2(e_1^*, e_2^*) + C_1(e_1^*, e_2^*)}{\Lambda_2(e_1^*, e_2^*)} \times \frac{1}{\int_{\mathbb{R}_+} \left\{ D(q) \left[ \frac{\partial f(q(e_1))}{\partial e_1} \right] \right\} dq}
\]
and
\[ m^* \int_{\mathbb{R}_+} \left\{ D(q) \left[ \frac{\partial f(q|e_1^*)}{\partial e_1} \right] \right\} dq - \Lambda_1(e_1^*, e_2^*) - \Lambda_2(e_1^*, e_2^*) \left[ -\frac{C_1(e_1^*, e_2^*)}{C_2(e_1^*, e_2^*)} \right] = 0. \]

Moreover, \( 0 < \theta^* < 1 \) and \( 0 < m^* \) because \( C_2 \) is negative, \( \Lambda_1, \Lambda_2, C_1 \) are positive, and \( \int_{\mathbb{R}_+} [D(q)f(q|e_1)]dq \) is strictly increasing and concave in \( e_1 \).

Given \( m^* \) and \( \theta^* \), the provider’s expected profit is maximized at \((e_1^*, e_2^*)\), and \( T \) is again chosen such that \( T + m^* \int_{\mathbb{R}_+} [D(q)f(q|e_1)]dq - \Lambda(e_1^*, e_2^*) = 0 \).
References


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


Brunt CS. CPT fee differentials and visit upcoding under Medicare Part B. Health Economics 2011; 20; 831-841.


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.


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.


Supplement: Examples

We describe two examples. In the first, let $\Lambda(e_1, e_2) \equiv \gamma(e_1 + e_2)$, where $\gamma$ is increasing and convex. Also, let $C(e_1, e_2) = c(e_1 - e_2)$, where $c$ is increasing and convex, and $q = e_1$. Quality and cost efforts are perfect substitutes. We have $\frac{\Lambda_1(e_1, e_2)}{\Lambda_2(e_1, e_2)} = 1$, $C_1(e_1 - e_2) = c'(e_1 - e_2)$, and $C_2(e_1 - e_2) = -c'(e_1 - e_2)$. The first-order conditions from disutility minimization in (8) give

$$1 = -\frac{\theta - (1 - \theta)c'(e_1 - e_2)}{(1 - \theta)c'(e_1 - e_2)} \quad \text{or} \quad 2(1 - \theta)c'(e_1 - e_2) = \theta,$$

which implicitly defines the expansion path $\bar{e}_2(e_1; \theta)$. Totally differentiating $2(1 - \theta)c'(e_1 - e_2) = \theta$, we obtain $c''(e_1 - e_2)d(e_1 - e_2) = 0$, so $\bar{e}_2(e_1; \theta) = 1$. The expansion path has a unit slope, and $\theta$ shifts this 45-degree line.

For a margin $m$, the provider’s profit is $\pi(e_1) = mD(e_1) - \gamma(e_1 + \bar{e}_2(e_1; \theta))$. Hence, its derivatives are

$$\frac{d\pi(e_1)}{de_1} = mD'(e_1) - \gamma'(e_1 + \bar{e}_2(e_1; \theta))(1 + \bar{e}_2'(e_1; \theta)) = mD'(e_1) - \gamma'(e_1 + \bar{e}_2(e_1; \theta))2$$

$$\frac{d^2\pi(e_1)}{de_1^2} = mD''(e_1) - \gamma''(e_1 + \bar{e}_2(e_1; \theta))4 < 0,$$

so the profit function is concave. For effort pair $(e_1^*, e_2^*)$, set $\theta$ at $\theta^*$ to satisfy $c'(e_1^* - e_2^*) = \frac{\theta^*}{1 - \theta^*}$, and $m$ at $m^*$ to satisfy $m^*D'(e_1^*) = \gamma'(e_1^* + e_2^*)2$. The weight $\theta^*$ and margin $m^*$ implement $(e_1^*, e_2^*)$.

In the second example, we keep $\Lambda(e_1, e_2) \equiv \gamma(e_1 + e_2)$ and $q = e_1$, but let $C(e_1, e_2)$ be $c(e_1) + d(e_2)$ with $c$ increasing, $d$ decreasing, and both $c$ and $d$ convex. We have $C_1(e_1, e_2) = c'(e_1)$ and $C_2(e_1, e_2) = d'(e_2)$. The first-order condition from disutility minimization in (8) give

$$1 = -\frac{\theta - (1 - \theta)c'(e_1)}{(1 - \theta)d'(e_2)} \quad \text{or} \quad (1 - \theta)[d'(e_2) - c'(e_1)] = \theta,$$

which implicitly defines the expansion path $\bar{e}_2(e_1; \theta)$. From total differentiation, we obtain $\bar{e}_2'(e_1; \theta) = c''(e_1)/d''(e_2) > 0$.

---

21 This is the disutility function used in Ma (1994).
The provider’s profit \( \pi(e_1) \equiv mD(e_1) - \gamma(e_1 + \bar{e}_2(e_1; \theta)) \) has derivatives

\[
\frac{d\pi(e_1)}{de_1} = mD'(e_1) - \gamma'(e_1 + \bar{e}_2(e_1; \theta))(1 + \bar{e}_2'(e_1; \theta))
\]

\[
\frac{d^2\pi(e_1)}{de_1^2} = mD''(e_1) - \gamma''(e_1 + \bar{e}_2(e_1; \theta))(1 + \bar{e}_2'(e_1; \theta))^2 - \gamma'(e_1 + \bar{e}_2(e_1; \theta))\bar{e}_2''(e_1; \theta).
\]

The second-order derivative of \( \pi(e_1) \) is negative if \( \bar{e}_2''(e_1; \theta) \leq 0 \). Using \( \bar{e}_2'(e_1; \theta) = c''(e_1)/d''(e_2) \), we have

\[
\bar{e}_2''(e_1; \theta) = \frac{d''(e_2)c'''(e_1) - c''(e_1)d'''(e_2)c''(e_1)/d''(e_2)}{[d''(e_2)]^2},
\]

which gives suitable conditions on the the third-order derivatives of \( c \) and \( d \) for \( \bar{e}_2''(e_1; \theta) \leq 0 \), and, hence, for the concavity of \( \pi(e_1) \equiv mD(e_1) - \gamma(e_1 + \bar{e}_2(e_1; \theta)) \). When concavity of \( \pi(e_1) \) holds, we can find weight \( \theta^* \) and margin \( m^* \) to implement effort pair \( (e_1^*, e_2^*) \).