Health Care Payment Systems: Cost and Quality Incentives

Ching-to Albert Ma
Department of Economics
Boston University
Boston, MA 02215

This paper compares the cost and quality incentive effects of cost reimbursement and prospective payment systems in the health industry. When a provider cannot refuse patients who require high treatment costs or discriminate patients by qualities, optimally designed prospective payments can implement the efficient quality and cost reduction efforts, but cost reimbursement cannot induce any cost incentive. When the provider can refuse expensive patients, implementation of the first best requires a piecewise linear reimbursement rule that can be interpreted as a mixture of pure prospective payment and pure cost reimbursement. Under appropriate conditions, prospective payment can implement the first best even when the provider can use qualities to discriminate patients.

1. Introduction

The debate about the effectiveness of prospective payment versus cost reimbursement payment has attracted a lot of attention in the discussion of health care policies in the United States. It is widely believed that the government's accommodating cost reimbursement policies in the 1960s and 1970s are largely responsible for the exorbitant growth in the share of health care expenditures in the gross national product (GNP). The change in Medicare's reimbursement policies 10 years ago was an attempt to avoid the supposedly perverse incentives under cost reimbursement policies. In effect, Medicare's

A preliminary version of this paper was presented at the Industrial Organization of Health Care Conference at the Osgood Hill Conference Center, North Andover, Massachusetts, September 19–21, 1993. I thank Thomas Lyon, Thomas McGuire, Michael Riordan, Daniel Spulber, a coeditor, and a referee for their helpful comments. Financial support from the Management Science Group, Veterans Administration, Bedford, Massachusetts, is gratefully acknowledged.

Journal of Economics & Management Strategy, Volume 3, Number 1, Spring 1994, 93–112

prospective payment system, which essentially pays a fixed "price" per discharge, with the price being determined by the patient's discharge diagnosis (diagnostic related group, or DRG), shifts cost responsibilities to health care providers. It is hoped that by making providers bear their "production" costs, incentives for cost reductions will be introduced. This paper provides a theoretical framework to evaluate the comparative effectiveness of prospective, fixed-price policies versus retrospective, cost-based reimbursement policies.

The contribution of this paper is to draw attention to the different *incentive trade-offs* between the two methods of reimbursement. We emphasize the *interaction* of cost and quality incentives under prospective payment and cost reimbursement systems and investigate whether the counterstrategies available to the regulated health care provider can neutralize or further enhance the intended effects of prospective payments. Thus, we adopt a *multitask* agency approach in the paper; see Holmstrom and Milgrom (1991) for a general model. The health care provider is portrayed as an agent with abilities to change its performance in various dimensions. It can use efforts to increase product or care quality; to reduce costs; to screen and, hence, be able to refuse excessively costly patients (dumping); and to use quality to attract low-cost patients (quality discrimination or cream skimming). Given a payment system, the agent determines the mix of efforts on cost reduction and quality enhancement activities to maximize net profit.

In the first version of our model, a hospital, the regulated health care provider, can only allocate its efforts between cost reduction and quality enhancement. Obviously under cost reimbursement, the hospital is not motivated to perform any cost reduction, and the tradeoff between cost and quality incentives is extreme—only quality enhancement incentives will be present. Nevertheless, the optimal cost reimbursement policy can induce a constrained efficient quality effort—constrained in the sense that cost effort is set at minimum. Contrary to the common fear that prospective payments may encourage cost reduction excessively, we show that the prospective payments system can achieve the efficient allocation of efforts between cost and quality. The intuition is actually simple to state. Given that the hospital fully internalizes treatment (production) costs, the prospective-payment price level can be set to make the hospital fully internalize the benefit of quality as well. So if a hospital can only choose between cost reduction and quality enhancement, prospective payment unambiguously represents an improvement over cost reimbursement.[1]

The previous result serves as a benchmark for the evaluation of the prospective- and cost-reimbursement policy regimes when the hospital's strategic ability is made more complex. In the next version of the model, the hospital is allowed to refuse patients if it cannot recover their costs of treatment through reimbursements. For simplicity, we adopt the assumption that the hospital can predict a potential patient's cost exactly.[2] While earlier researchers have focused on the harm of hospital refusing expensive patients per se, our emphasis is on its effect on cost incentives. Because the hospital never refuses a patient under cost reimbursement, the policy can again guarantee a constrained efficient level of quality. Under prospective payment, the hospital will refuse expensive payments; forcing the hospital to internalize costs fully is infeasible. Depending on the structural relationship between effort and cost reduction and quality enhancement outcomes, the distortion induced by dumping may be more or less severe. Therefore, prospective payment need not always achieve a better overall incentive than cost reimbursement. We also show that a combination of prospective payment and cost reimbursement can achieve the efficient mix of efforts.

We next investigate the performance of the policy regimes when the hospital can use different quality levels to discriminate patients whose treatment costs vary systematically. This cream-skimming phenomenon may arise because the classifications in the prospective payment system fail to distinguish systematic cost differences among patients with the same diagnosis, or the hospital obtains superior information about patient illness severities and can treat some patients with lower costs. Again, the welfare comparison of the two policy regimes is generally ambiguous. But under some general conditions on consumer demands, we show that quality discrimination actually remedies the deficiency in the prospective payment classification DRGs. With cream skimming the regulator sometimes can implement the efficient mix of incentives corresponding to the refined DRG classification that can distinguish systematic cost variations. Thus, in these situations, prospective payment must dominate cost reimbursement.

[1] To the extent that hospital costs may not be observed easily, prospective payment appears even more attractive, since it eliminates costs observation. In this paper, we are not directly concerned with administrative and auditing costs; throughout we assume that hospital costs can be verified ex post.

[2] It has been argued that through observing patient characteristics, a hospital can predict costs for 10–20% of its patients; see Dranove (1984).

After the introduction of prospective payments, a number of papers have discussed its limitations and incentive problems.[3] Ellis and McGuire (1986) question whether the prospective payment system alone can adequately align differences in preferences between doctors and patients; they show that a combination of prospective payment and (partial) cost reimbursement will be necessary when the doctor is an imperfect agent for the patient. Dranove (1987) analyzes the dumping and specialization incentives under prospective payment. Allen and Gertler (1991) demonstrate that when patient heterogeneity exists within a DRG, some type of patients must receive inefficient quality under prospective payment.

Although economists were rather quick to point out the inadequacies of the prospective payment system, few analyzed the balance between cost reduction and quality enhancement efforts allocations. Nevertheless, Shleifer (1985) shows that prospective payments can induce the efficient cost decisions by hospital. A crucial difference between the Shleifer paper and the present one is the inclusion of quality decisions in our analysis. Pope (1989) studies an imperfectly competitive market in which hospitals use quality to compete for patients. In that paper, the hospital can enhance quality and reduce "managerial slack." While Pope shows that competition may reduce managerial slack but result in excessive quality, the consequences of dumping and quality discrimination are not considered. We do not consider competition, but the quality and cost efforts in this paper are similar to the hospital's available strategies in Pope (1989). Ellis (1993) studies a duopoly model in which hospitals can employ dumping as well as care intensity (quality) to compete for and discriminate patients. Although dumping, creaming, and skimping are considered, Ellis (1993) does not allow cost reduction efforts.

The change from cost reimbursements to prospective payments in the health sector markedly resembles the recent change from rate of return regulation to price-cap regulation in the telecommunications industry and in public utilities. A group of papers in the literature studies incentive issues related to such changes in these industries; see Brennan (1989); Cabral and Riordan (1989); Clemenz (1991); Armstrong, Rees, and Vickers (1991); and the references they cite. These papers emphasize the incentive effects on input choices and cost reduction under different regulatory regimes. Institutional factors that are particularly important in the health care market, such as insur-

[3] See Newhouse (1983, 1984) and Pauly (1984), and Russell (1989) for a summary of empirical assessments.

and consumer response to product qualities, so far have not been incorporated.

Researchers in the cost-plus versus price-cap debate mostly agree that the former regime gives the wrong incentives for the regulated firm to choose input mix efficiently, or it fails to motivate cost reduction activities. Much of the (normative) research is on the design of the optimal rate of decrease of the price cap or the optimal duration during which the price cap is maintained. The analysis here not only focuses on the effect of price regulation on cost reduction but emphasizes the incentive tradeoff between cost reduction and quality enhancement.

The paper is organized as follows: The model is described in Section 2. The basic efficiency result is derived in the following section; we show that when cost reduction and quality enhancement are the only activities for the hospital, prospective payment implements the efficient allocation. Then in Sections 4 and 5, we investigate the incentive tradeoffs when the hospital can dump expensive patients, and when the hospital can use quality to discriminate patients with systematic cost differences. The last section draws some conclusions.

2. The Basic Model

We consider the regulation of a single firm in the health care industry. While we will use the term *hospital,* the firm under question can be any general or specific health care provider. The analysis concentrates on the supply side; we assume that consumers are fully insured and ignore the possible incentive effects of demand-side cost sharing. From the full-insurance assumption, it follows that consumers' choice of a health care provider depends on the quality of care. Accordingly, consumer demands at the hospital are an increasing function of its quality of care.

Central to our analysis is the assumption that a hospital manager can expend "effort" for various activities. For example, by reducing slack in the system or making treatment procedures more effective, the hospital can lower its treatment costs or enhance care quality. The variables t_1 and t_2 denote efforts a hospital manager can direct to quality enhancement and cost reduction dimensions, respectively; minimum effort levels are normalized at zero. These efforts impose a total disutility of $\gamma(t_1 + t_2)$ to the hospital manager. The function γ is increasing and convex.

A higher care quality may require higher effort input by the hospital, higher unit cost, or both. By a slight abuse of notation, we also use t_1 to denote the quality of care at the hospital.[4] The hospital has a constant return-to-scale treatment technology. The unit cost of treating a patient is $c(t_1, t_2)$ when the hospital chooses quality of care t_1 and cost reduction effort t_2. With subscript i denoting the partial derivative of a function with respect to that its ith argument, we assume $c_1 \geq 0$ and $c_2 < 0$. Furthermore, we assume that c is convex. Also, let the increasing and concave function $\mu(t_1)$ be the hospital's demand when it provides services at quality t_1. We assume that the total cost $c(t_1, t_2)\mu(t_1)$ is a convex function.

We will consider two reimbursement methods. Under cost reimbursement, the hospital's actual production costs are completely paid for; in addition, the hospital may be paid a nonnegative margin, m, so that if the hospital's cost of treating a patient is c, it receives a total reimbursement of $c + m$. The margin may be used to motivate the hospital to supply effort on quality enhancement. Under prospective payment, the hospital receives a fixed payment per discharge, p, independent of the total costs of treatment. The hospital is assumed to maximize revenue less treatment costs and the disutility of efforts, or its net profit. Hence, the hospital's net profit functions under cost reimbursement and prospective payment are respectively:

$$(c(t_1, t_2) + m)\mu(t_1) - c(t_1, t_2)\mu(t_1) - \gamma(t_1 + t_2) = m\mu(t_1) - \gamma(t_1 + t_2) \quad (1)$$

$$p\mu(t_1) - c(t_1, t_2)\mu(t_1) - \gamma(t_1 + t_2). \quad (2)$$

Observe that since both profit functions are concave, the hospital's profit-maximizing choices of t_1 and t_2 can be fully characterized by the first order conditions. Finally, the reservation net profit for the hospital is set at zero.

The gross benefit produced by the hospital generally depends on the number of consumers using the hospital and the quality of care they receive. But the hospital's demand is an increasing function of its quality; therefore we simply write the gross benefit as a function of quality alone, $W(t_1)$. We assume that W is increasing and concave. Notice that we use a general gross benefit function. In many applications, benefits are taken to be consumer surplus. The function W includes that as a special case but is sufficiently general to incorporate other policy issues that are thought to be important in health care. For instance, because of moral hazard, and other institutional or infor-

[4] More generally, let $q + t_1$ denote the total quality of care and $c(q + t_1, t_2)$ the corresponding unit cost when the hospital chooses base quality q, efforts t_1 and t_2. Then, redefining $q + t_1$ as a new variable, we eliminate q.

mational distortions, the appropriate benefit measure may be bigger or smaller than consumer surplus.

A regulator or a payer uses a payment policy to reimburse the hospital for its costs of treating patients. This regulator/payer may be a public agency or a private insurance company. In any case, the regulator's preferences are given by the difference between consumer benefit and the total cost of production:

$$W(t_1) - c(t_1, t_2)\mu(t_1) - \gamma(t_1 + t_2). \tag{3}$$

When $W(t_1)$ represents gross consumer surplus, then expression (3) measures the sum of net consumer surplus and producer surplus.

Generally the regulator's preferences may be defined as a weighted sum of consumer welfare and net profits (or hospital utility). That is, if R represents the hospital's revenue, then the regulator's preferences can be written as $\alpha\{W(t_1) - R\} + (1 - \alpha)\{R - c(t_1, t_2)\mu(t_1) - \gamma(t_1 + t_2)\}$. Our formulation in expression (3) assumes equal weights on consumer welfare and profits.

If the regulator is a public agency, the assumption that weights on consumer welfare and profits are equal is plausible, although a weaker assumption may be that the weight on net consumer benefit is at least as large as that on profits ($\alpha \geq .5$). If the regulator is a private insurance company seeking to maximize profits in a competitive environment, it may only care about net consumer surplus so that the weight on profits becomes zero ($\alpha = 1$).

Nevertheless, even under the general specification, expression (3) can be justified when the payment to the hospital includes a lump-sum transfer. Then the lump-sum transfer will be chosen by the regulator/payer to extract net profits; that is, $R = c(t_1, t_2)\mu(t_1) + \gamma(t_1 + t_2)$. As a result expression (3) reemerges. Although it does not appear that the current payment systems explicitly specify transfers, in practice hospitals often receive direct subsidies in various forms. Thus, our implicit assumption that transfers can be used need not be implausible. In any case, our definition in expression (3) allows us to ignore distribution issues and concentrate on efficiency.

As a benchmark, the efforts that maximize the regulator's objective function (3) are called the efficient allocation of efforts.[5] From concavity, the efficient allocation of efforts on quality enhancement and cost reduction, t_1^* and t_2^*, will be given by the following first order

[5] Notice that due to insurance, consumers are not responsible for their (ex post) care expenditure, and, hence, the efficient allocation of efforts refers to a *second best*.

conditions:

$$W'(t_1^*) - c_1(t_1^*, t_2^*)\mu(t_1^*) - c(t_1^*, t_2^*)\mu'(t_1^*) = \gamma'(t_1^* + t_2^*) \tag{4}$$

$$-c_2(t_1^*, t_2^*)\mu(t_1^*) = \gamma'(t_1^* + t_2^*). \tag{5}$$

3. Efficiency and Payment Systems

In this section, we compare the relative efficiency of the cost reimbursement and prospective payment systems when quality enhancement and cost reduction are the hospital's only activities. It is obvious that by completely paying for the hospital's costs, the cost reimbursement system eliminates any cost reduction incentives. Moreover, other researchers have suggested that quality of care may be excessive under this system. On the other hand, the prospective payment system makes the hospital fully responsible for costs. It has been argued that this policy may provide too much incentive for cost reduction and too little for quality improvement. Our formal model allows us to evaluate these propositions.

We begin by investigating the hospital's behavior under cost reimbursement. Clearly, from expression (1), the hospital will take a minimum effort for cost reduction. Given the margin m, the hospital will choose t_1 to maximize expression (1). The corresponding first order condition becomes:

$$m\mu'(t_1) = \gamma'(t_1). \tag{6}$$

Because eq. (6) defines a one-to-one relationship between m and t_1, a given value of t_1 can be implemented by a choice of m satisfying eq. (6). The equilibrium m and t_1 will be those that maximize the regulator's objective (3) with the constraints that cost reduction effort t_2 be at its minimum, and that m implements t_1.

Now let $t_1^\dagger$ denote the maximizer of the regulator's preferences in expression (3) when t_2 is set at its minimum (zero). Then $t_1^\dagger$ is given by

$$W'(t_1^\dagger) - c_1(t_1^\dagger, 0)\mu(t_1^\dagger) - c(t_1^\dagger, 0)\mu'(t_1^\dagger) = \gamma'(t_1^\dagger).$$

Then this "constrained" efficient[6] level of quality, $t_1^\dagger$, can be implemented simply by setting

$$m = \frac{\gamma'(t_1^\dagger)}{\mu'(t_1^\dagger)} = \frac{W'(t_1^\dagger) - c_1(t_1^\dagger, 0)\mu(t_1^\dagger) - c(t_1^\dagger, 0)\mu'(t_1^\dagger)}{\mu'(t_1^\dagger)}. \tag{7}$$

[6] "Constrained" is in the sense that cost reduction effort is set at its minimum.

Profits are given by

$$\int_0^{t_1^\dagger} \left[\frac{W'(t_1^\dagger) - c_1(t_1^\dagger, 0)\mu(t_1^\dagger) - c(t_1^\dagger, 0)\mu'(t_1^\dagger)}{\mu'(t_1^\dagger)} \mu'(x) - \gamma'(x) \right] dx.$$

From the definition of $t_1^\dagger$, the integrand is equal to zero at $x = t_1^\dagger$. Hence, the concavity of expression (3) and μ implies that the integrand is nonnegative for $x < t_1^\dagger$. Thus, the hospital earns nonnegative profits. In summary, under cost reimbursement, the equilibrium margin will be given by eq. (6), and the equilibrium quality effort level is $t_1^\dagger$.

Under a general condition, we can compare the optimal quality effort under cost reimbursement $t_1^\dagger$ and the efficient quality effort t_1^*. Consider the "marginal social cost" of quality: $c_1(t_1, t_2)\mu(t_1) + c(t_1, t_2)\mu'(t_1) + \gamma'(t_1 + t_2)$. Suppose that this is increasing in t_2, then we must have $t_1^\dagger > t_1^*$. When the marginal social cost of quality is increasing in cost reduction effort, the optimal cost reimbursement policy leads to excessive quality.[7] Intuitively, if a fall in t_2 reduces the marginal social cost of providing quality effort, then the equilibrium quality effort must be above the efficient level when t_2 is at its minimum.

Despite its extreme incentive properties, the cost reimbursement system can achieve efficiency in one dimension of the overall incentive trade-offs. In summary, cost reimbursement can be a valuable policy alternative when the efficient trade-off of incentives in different dimensions cannot be maintained by other payment systems, as we see in the next two sections.[8]

Now we turn to the incentive properties of prospective payments. Given the payment parameter p, the hospital's behavior can be described by its choice of t_1 and t_2 that maximize profits (2). The associated first order conditions with respect to t_1 and t_2 are

$$p\mu'(t_1) - c_1(t_1, t_2)\mu(t_1) - c(t_1, t_2)\mu'(t_1) = \gamma'(t_1 + t_2) \tag{8}$$

$$-c_2(t_1, t_2)\mu(t_1) = \gamma'(t_1 + t_2). \tag{9}$$

[7] Consider the first order condition (4): $W'(t_1^*) - c_1(t_1^*, t_2^*)\mu(t_1^*) - c(t_1^*, t_2^*)\mu'(t_1^*) - \gamma'(t_1^* + t_2^*) = 0$. If t_2^* is reduced to 0, then the expression on the left-hand side becomes strictly positive. Thus, the first order derivative of expression (3) at $t_1 = t_1^*$ and $t_2 = 0$ is strictly positive. Therefore, $t_1^* < t_1^\dagger$.

[8] We have assumed throughout that patients' benefits do not affect the hospital's objective directly. Because most hospitals are nonprofit organizations, other researchers have modified the provider's objective to incorporate patients' benefits. In these models, when compared with the efficient level, hospital expenditure typically will be excessive. In our model, this will mean that quality effort will be provided even under cost reimbursement with a zero margin.

By forcing the hospital to bear costs, the prospective payment system also makes the hospital internalize the costs of treatment. Clearly, the first order conditions with respect to cost reduction efforts (t_2) for the hospital's profit maximization and for the regulator's preference maximization (respectively, eqs. [5] and [9]) are identical! Since the private cost reduction incentive is already aligned with that in the regulator's objective, the efficient allocation of quality and cost efforts, t_1^* and t_2^*, can be implemented by setting the price p appropriately. More precisely, simply set the prospective payment equal to

$$p = \frac{W'(t_1^*)}{\mu'(t_1^*)}. \tag{10}$$

Given this value of p, eqs. (8) and (9) yield a solution (t_1^*, t_2^*).

Moreover, the hospital earns nonnegative profits when p is given by eq. (10). Profits are given by

$$\int_0^{t_1^*} \left[\frac{W'(t_1^*)}{\mu'(t_1^*)} \mu'(x) - c_1(x, t_2^*)\mu(x) - c(x, t_2^*)\mu'(x) - \gamma'(x + t_2^*)\right] dx.$$

The integrand is equal to zero when $x = t_1^*$. Thus, the concavity of expression (3) and μ implies that the integrand is nonnegative for $x < t_1^*$, and profits are nonnegative. Hence, the equilibrium efforts under prospective payment are t_1^* and t_2^*. Since a prospective payment system can induce the efficient mix of quality enhancement and cost reduction efforts, the regulator gains from changing the optimal cost reimbursement regime to the optimal prospective payment regime.[9]

This result appears to contrast with that in the usual monopoly regulation model (Spence, 1975). Here we achieve the efficient allocation by setting a single price for the hospital, whereas a nonlinear price schedule will be necessary. Our result lies in our special assumption that demand is price-inelastic due to insurance. In the usual model, the regulator is concerned with both the size of the market

[9] When profits are weighed less than consumer surplus, optimal cost reimbursement and prospective payment policies must limit profits at the expense of distorting the hospital's efforts from the respective efficient levels. Thus, the optimal prospective payment will not implement t_1^* and t_2^*. Nor will the optimal reimbursement margin implement $t_1^\dagger$. Although we have been unable to establish analytical results, we conjecture that the equilibrium efforts will become lower. This is because the prospective payment and reimbursement margin in eqs. (7) and (10) allow the hospital to earn positive profits. Nevertheless, since prospective payment is superior to cost reimbursement when profits and consumer surplus are weighed equally, this ranking is preserved when profits are not significantly weighed less than consumer surplus. It is unclear how the ranking is affected if the weight on profits is significantly less.

served by the monopolist and product quality. Hence, a simple price policy will be ineffective in motivating the monopolist to serve the efficient number of consumers *and* to supply the efficient quality level. If the insurance aspect of health care is incorporated into our model in a more complicated way, consumers' demand function will depend on the hospital's quality of care, insurance premiums, and co-insurance rates or deductibles. Then the role for nonlinear prices in aligning the payer's and the hospital's objectives will become important.

4. Dumping and Cost Reduction Efforts

In the last section, we demonstrated the attractiveness of the prospective payment system. When the hospital fully internalizes treatment cost, the pricing policy can be set to promote quality enhancement effort. Crucial to the previous result is the assumption that forcing the hospital to internalize cost is feasible. But in practice, this may not be achieved easily. In fact ever since the prospective system was introduced, economists have worried that the problem of dumping, in which excessively costly patients are refused treatments by the hospital, would be an important drawback.

In this section we show that dumping generally adversely affects the performance of the cost reduction incentives under prospective payments. This harm is a result of the hospital's ability to refuse high-cost patients and, therefore, to avoid incentives to reduce those costs. We focus on the cost reduction incentive effect of dumping and only implicitly include other consequences that have been emphasized by other researchers. Generally, dumping can restrict patients' access to their preferred medical facilities, result in inappropriate care, and create higher and even excessive demands at county or public hospitals, which must take all patients. Although these other costs of dumping may be significant, we have chosen in this work to capture only those components that are attributable to cost and quality incentives.

For simplicity, in this section and the next, we assume that higher care quality only requires higher effort; thus, the unit cost of treatment is independent of the quality of care. Extending the model in Section 2, we assume that patient severity is distributed randomly. Accordingly, the hospital's cost of treating a patient is distributed according to some (cumulative) distribution, F, with support $[0, \hat{c}]$. We assume that F is a function of t_2, the hospital's effort in cost reduction activities. Thus, $F(c; t_2)$ represents the percentage of patients who can be treated with cost c or lower if the hospital uses effort t_2 to reduce costs. Let $\bar{c}(t_2)$ denote the average cost of treatment if the hospital chooses effort t_2:

$$\bar{c}(t_2) = \int_0^{\hat{c}} c \, dF(c; t_2).$$

We assume that the function $\bar{c}(t_2)$ is decreasing in t_2. This may be because a higher cost reduction effort leads to a better cost distribution in the sense of first order stochastic dominance.[10]

Under cost reimbursement, the hospital's treatment costs are fully paid for by the regulator. Clearly, the hospital does not exert any effort to reduce cost. Nor does it refuse to admit patients. Hence, the quality decision will be identical to the corresponding quality choice under cost reimbursement in Section 3. The constrained efficient quality effort $t_1^\dagger$ will be implemented; see eqs. (6) and (7).

Next, we consider prospective payment. For the efficiency benchmark, one can simply replace the corresponding cost functions in eqs. (4) and (5) by $\bar{c}(t_2)$.[11] If the hospital neither observes a patient's cost nor practices dumping, then the efficient quality enhancement and cost reduction efforts can be implemented, analogous to the result of the model with deterministic costs in Section 3. If the hospital observes a patient's severity and, hence, his treatment cost, it may decide to refuse admission to a patient if the price is smaller than the cost. In this case, if p is the prospective price and $p < \hat{c}$, the hospital's profit becomes

$$\mu(t_1) \int_0^{p} (p - c) dF(c; t_2) - \gamma(t_1 + t_2);$$

those patients whose costs are between p and $\hat{c}$ will not be accepted by the hospital.

For a prospective payment rate of $p < \hat{c}$, the hospital's profit-maximizing choices of t_1 and t_2 will be given by the following first order conditions:

$$\mu'(t_1) \int_0^{p} (p - c) dF(c; t_2) = \gamma'(t_1 + t_2) \tag{11}$$

$$\mu(t_1) \int_0^{p} (p - c) dF_2(c; t_2) = \gamma'(t_1 + t_2). \tag{12}$$

We will assume that patients who are refused treatment at the

[10] That is, for $t_2 < t_2'$, $F(c; t_2) \leq F(c; t_2')$, with a strict inequality for some c.

[11] Since c is now independent of t_1, the partial derivative of c with respect to t_1 in eq. (4) will be set at zero.

hospital will seek treatment at a public hospital and will receive care at a fixed quality. The gross benefit will be given by

$$\int_0^p W(t_1)dF(c;\, t_2) + \int_p^{\hat{c}} L\, dF(c;\, t_2),$$

where L denotes the consumers' benefit of receiving care at the public hospital. Hence, the regulator's objective can be written as:

$$\int_0^p [W(t_1) - c\mu(t_1)]dF(c;\, t_2) + \int_p^{\hat{c}} [L - c\mu(t_1)]dF(c;\, t_2) - \gamma(t_1 + t_2),$$

or simply

$$W(t_1) - \bar{c}(t_2)\mu(t_1) - \gamma(t_1 + t_2) - [1 - F(p;\, t_2)][W(t_1) - L], \quad (13)$$

where the last term in expression (13) can be taken to be the costs of dumping that are not directly due to diminished cost reduction incentives.

For any prospective price set between 0 and $\hat{c}$, the implementation of the efficient mix of quality and cost efforts is infeasible. The hospital's choice of t_1 and t_2 is given by eqs. (11) and (12). Thus, for any $p < \hat{c}$, then the solution from eqs. (11) and (12) will differ from t_1^* and t_2^*. This is true even if the nonincentive costs of dumping (the last term in expression [13]) can be eliminated completely. Indeed, from eq. (10), the efficient allocation can be achieved only if $W'(t_1^*)/\mu'(t_1^*) > \hat{c}$, that is, when the price for motivating the efficient quality enhancement effort level is above the support of the treatment cost distribution, so that dumping will not occur.

Dumping may also be detrimental for quality incentives. In the absence of dumping, the hospital's marginal return from increasing its qualify effort is $\mu'(t_1)(p - \bar{c}(t_2))$. Comparing this with the expression on the left hand side of eq. (11), we see that for $p < \hat{c}$, this marginal return will be smaller than that under dumping. In other words, dumping ensures that the hospital earns strictly positive profit from each patient and may encourage an overinvestment in quality to increase total demands.

Because of these undesirable effects of dumping, there is no a priori reason that prospective payment must perform better than cost reimbursement. For example, when cost reduction effort is ineffective, then the efficient cost reduction effort will be near its minimum and the performance of cost reimbursement is not excessively poor; this will be true when the value of $c_2(t_1,\, t_2)$ is insignificant. On the other hand, when the cost of dumping is high, then the prospective payment rate may be set too high, resulting in excessive quality and cost efforts and poor overall welfare; this will be true when consumers' benefit from services at a public hospital, L, is very low. Indeed, cost reimbursements and prospective payments implement different kinds of incentive trade-offs; while cost reimbursement gives only incentives for quality enhancement efforts, prospective payments may result in dumping and in suboptimal quality and cost efforts. The relative performance of these two systems will depend on their particular applications.

For policy implications, our result suggests another advantage of refining the classifications of DRG on which current prospective payment systems are based. A refinement of an existing classification introduces new categories to narrow the range of cost variations within any particular discharge category. Hence, the likelihood of finding $W'(t_1^*)/\mu'(t_1^*) > \hat{c}$ will tend to increase.

Our result also suggests that the current practice of adjusting the prospective payment according to patient severity, or casemix, may not mitigate the perverse cost reduction incentives due to dumping. Whenever a hospital's prospective payments are adjusted downward, possibly due to its lower average costs in a previous year, its cost reduction incentives will be further reduced, because the likelihood of dumping is inversely related to the prospective rate. On the other hand, providing extra compensations to hospitals for treating extremely high-cost patients—outlier payments—will be at least neutral to cost incentives; under both dumping and the supplemental outlier payments, the hospital would not be responsible for those high costs. Moreover, as we shall demonstrate next, optimally designed outlier payments can actually restore the efficient mix of quality and cost efforts.

The implementation of the efficient allocation of efforts, t_1^* and t_2^*, can be achieved by a piecewise linear-reimbursement rule that can be interpreted as a mixture of fully cost-based and fully prospective payment systems. Recall that when the hospital does not practice dumping, the prospective reimbursement rate that implements the efficient cost and quality efforts is denoted by $p^* = W'(t_1^*)/\mu'(t_1^*)$ (see eq. [10]). If dumping is used, we must have $p^* < \hat{c}$.

Consider conditioning the payment on cost, and let $p(c)$ denote the regulator's payment to the hospital when the actual cost of treatment is c. To avoid dumping, the payment must be above cost, so that for $0 \leq c \leq \hat{c}$, $p(c) \geq c$. Now choose c^* in $[0, \hat{c}]$ to satisfy

$$\int_0^{\hat{c}} (p^* - c)dF(c; t_2^*) = \int_0^{c^*} (c^* - c)dF(c; t_2^*). \tag{14}$$

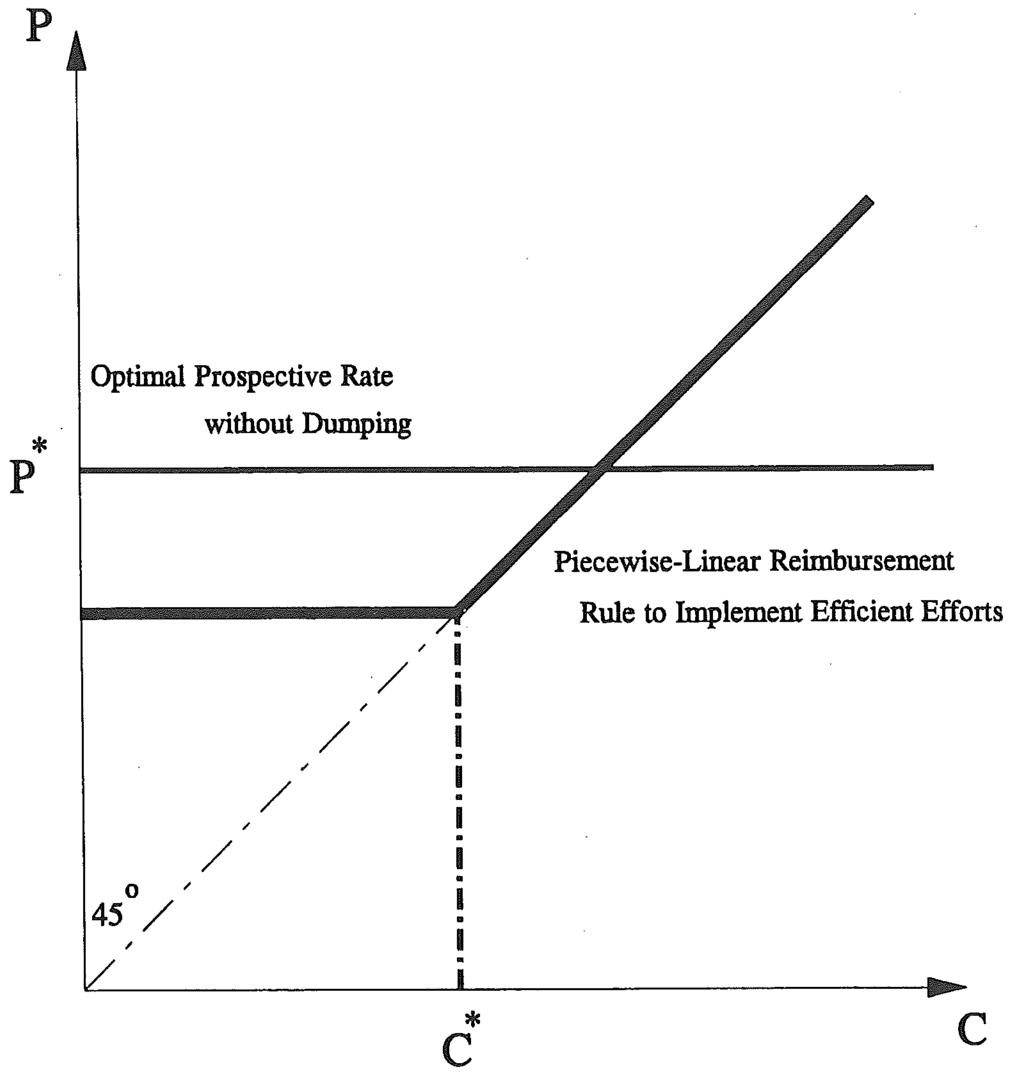

FIGURE 1. REIMBURSEMENT RULE TO IMPLEMENT EFFICIENT EFFORTS.

Such a c^* exists, is less than p^*, and is unique.[12] Then set the reimbursement rule $p(c)$ to be

$$p(c) = \begin{Bmatrix} c^* & \text{for } c \leq c^* \\ c & \text{for } c > c^* \end{Bmatrix}.$$

Figure 1 illustrates this reimbursement rule. For costs that are below c^*, the hospital is reimbursed by a prospective payment equal

[12] Clearly, the expression on the right-hand side of eq. (14) is continuous and strictly monotonic in c^*. For $c^* = 0$, it is strictly less than the left-hand side expression; for $c^* = p^*$, strictly greater. By continuity and monotonicity, there must exist a unique number, c^*, in $[0, p^*]$ such that eq. (14) holds.

to c^*; for costs above c^*, the hospital is reimbursed its costs. From eq. (14), the choice of c^* ensures that the hospital earns a revenue level equal to that from a pure prospective payment system when dumping is ruled out. Thus, the hospital is given the incentive to provide the efficient cost and quality efforts.

The reimbursement rule also can be interpreted as a form of "sliding-scale" regulation; Lyon (1993) provides an analysis of general properties of sliding-scale regulations. Under this form of regulation, profit sharing is incorporated into a simple price-cap mechanism. Whenever the rate of return is either below or above a predetermined range, the price will be adjusted. Under our reimbursement rule, the hospital's payment will be increased to cover costs whenever these costs are above a preset level; otherwise, the hospital is fully responsible for its costs.

5. Efficiency and Quality Discrimination

In this section we assume that a hospital can use quality discrimination or cream skimming to circumvent the intended purpose of the prospective payment system. Cream skimming refers to those activities a hospital may engage in to attract patients that are low cost; in our framework, the hospital uses different levels of quality to discriminate against patients with different severities. For example, high qualities may be offered to those patients with lower severity and treatment costs.

Cream skimming is possible because the payment classification system has failed to distinguish between patients whose treatment costs vary systematically within the same discharge category. More fundamentally, the hospital may obtain some information that allows it to use different procedures for illnesses under a single diagnosis. Therefore, one category of discharge and payment for the hospital may include two or more different types of patients who systematically differ in their severities and treatment requirements.

We assume that there are two types of patients, A and B. Type A patients only require a fraction, θ, of the (nonstochastic) treatment costs of type B patients. Let t_1^j and t_2^j be respectively the quality and cost reduction efforts the hospital chooses for type j patients, $j = A$, B. The unit cost of treatment for type A and B patients is $\theta c(t_2^A)$ and $c(t_2^B)$, respectively, with corresponding demands $\mu(t_1^A)$ and $\mu(t_1^B)$.

Under cost reimbursement, the systematic cost variations among subpopulations of patients will be insignificant to the hospital, since all the costs will be paid for. Because there is no incentive to discriminate according to quality, a uniform quality will be offered. Cost reduction effort, however, remains at the minimal level. This also implies that if the hospital does not practice cream skimming, then the efficiency result in Section 3 will apply, and prospective payment will strictly dominate cost reimbursement. We now compare the two payment systems when the hospital uses quality to discriminate patients.

Suppose the prospective payment per discharge is p, then for effort choices t_i^j, $i = 1, 2$, $j = A, B$, the hospital's net profit is

$$\mu(t_1^A)[p - \theta c(t_2^A)] + \mu(t_1^B)[p - c(t_2^B)] - \gamma(t_1^A + t_2^A + t_1^B + t_2^B).$$

We examine the profit-maximizing choices of t_i^j, $i = 1, 2$, $j = A, B$. The first order conditions with respect to t_1^A, t_2^A, t_1^B, and t_2^B are as follows:

$$\mu'(t_1^A)[p - \theta c(t_2^A)] = \gamma'(t_1^A + t_2^A + t_1^B + t_2^B) \qquad (15)$$

$$-\theta c'(t_2^A)\mu(t_1^A) = \gamma'(t_1^A + t_2^A + t_1^B + t_2^B) \qquad (16)$$

$$\mu'(t_1^B)[p - c(t_2^B)] = \gamma'(t_1^A + t_2^A + t_1^B + t_2^B) \qquad (17)$$

$$-c'(t_2^B)\mu(t_1^B) = \gamma'(t_1^A + t_2^A + t_1^B + t_2^B). \qquad (18)$$

Generally, the hospital chooses different quality and cost efforts for the two classes of patients. At the maximum $t_1^A \leq t_1^B$ only if $t_2^A \leq t_2^B$.[13] We now turn to the welfare implication of quality discrimination.

First consider a superior pricing classification system in which the two classes of patients with systematic cost differences can be distinguished. Given these distinct classifications, the two classes of patients can be assigned different prices, say, p^A and p^B. Cream skimming within a subclass will not occur any more, and the efficiency result in Section 3 will apply. That is, p^A and p^B will be set at $W'(t_1^{A*})/\mu'(t_1^{A*})$ and $W'(t_1^{B*})/\mu'(t_1^{B*})$, respectively. Here, t_1^{A*} and t_1^{B*} are the efficient efforts for quality for the two groups of consumers, and t_1^{A*}, t_2^{A*}, t_1^{B*}, t_2^{B*} maximize

$$W(t_1^A) + W(t_1^B) - \theta c(t_2^A)\mu(t_1^A) - c(t_2^B)\mu(t_1^B) - \gamma(t_1^A + t_2^A + t_1^B + t_2^B). \qquad (19)$$

We now return to the inferior classification system that can only set one price for both classes of patients. First, when the hospital uses quality discrimination, the regulator must face the constraints eqs. (15)–(18), when its objective (19) is maximized. Observe, however,

[13] To prove this, suppose that $t_1^A \leq t_1^B$. Then we have $\mu(t_1^A) \leq \mu(t_1^B)$ and $\mu'(t_1^A) \geq \mu'(t_1^B)$. From eqs. (15), (17), and $\mu'(t_1^A) \geq \mu'(t_1^B)$, we have $p - \theta c(t_2^A) \leq p - c(t_2^B)$ and $c(t_2^B) \leq \theta c(t_2^A)$. Hence, $t_2^A \leq t_2^B$.

that only two of these four constraints are effective. Because the hospital fully internalizes costs, the first order conditions with respect to cost reduction efforts, eqs. (16) and (18), will not be effective constraints for the regulator. Hence, under quality discrimination the maximum welfare is given by the maximization of expression (19) subject to eqs. (15) and (17). Second, when the hospital does not practice cream skimming, then it will supply the same quality for the two classes of patients; thus, the regulator achieves the (constrained) maximum of expression (19) subject to the constraint $t_1^A = t_1^B$.

In general, the regulator's preferences on prospective payments with and without quality discrimination is ambiguous. But notice that the result in Section 3 implies that the regulator prefers prospective payment to cost reimbursement whenever the hospital does not discriminate patients by quality. So whenever the maximum welfare from the prospective payment with cream skimming ranks higher than that from without, prospective payment must be superior than cost reimbursement. Somewhat surprisingly, sometimes the regulator can actually achieve the unconstrained maximum of expression (19), which is infeasible without quality discrimination!

To see this possibility, suppose that $W'(t)/\mu'(t)$ is a constant (independent of t). In this case, identical prices will be set for both types A and B patients even if the two types can be assigned different DRGs. Then let the regulator set the prospective price p equal to this constant $(W'(t)/\mu'(t))$. It is then easy to check that the regulator's constraints under cream skimming, eqs. (15)–(18), exactly become the first order conditions for the unconstrained maximization of the regulator's objective (19)! Then, despite the coarse classification in the given pricing system, prospective payment with quality discrimination will implement that efficient allocation were the payment classification system sufficiently rich to capture all systematic cost variations within any single price category.[14]

A sufficient condition for $W'(t)/\mu'(t)$ being independent of t is that the demand function $\mu(t)$ is separable in quality t and price d, and that W measures consumer surplus. That is, μ can be written in the form $f(t)g(d)$ or $f(t) + g(d)$, where f and g are two functions, and d is the patient's (exogenous) out-of-pocket expenditure when seeking treatment from the hospital. In these cases consumer surplus W is $\int_d^\infty f(t)g(y)dy$ or $\int_d^\infty [f(t) + g(y)]dy$.

[14] Again, using a similar argument in Section 3, one can verify that the hospital will earn nonnegative profits.

6. Conclusion

In this paper, we have evaluated cost reduction and quality enhancement incentives of cost reimbursement and prospective payment systems. If the hospital cannot dump high-cost patients, and if the pricing classification system truly reflects all systematic cost variations, then prospective payment can implement the efficient allocation of cost reduction and quality enhancement efforts. When the hospital avoids high-cost patients through dumping, it lacks the incentive to reduce these costs. Hence, the prospective payment system may fail to achieve a balance between quality and cost incentives. It is quite possible that simple cost reimbursement performs better. We find that in the presence of dumping, a combination of prospective payment and cost reimbursement can align the proper incentives and achieve the efficient allocation.

When there are systematic cost differences within a single pricing category under the prospective payment system, then the hospital can use quality to discriminate patients. The comparison of the efficiency properties between prospective payment and cost reimbursement systems is ambiguous. But we find that for a broad class of environments, not only does the prospective payment system together with cream skimming perform better than one in which cream skimming is absent, it can actually achieve an efficient allocation corresponding to a refined classification with systematic cost differences fully reflected by prospective prices. As a corollary, in this environment prospective payment is superior than cost reimbursement.

Our normative approach allows us to evaluate claims about the performance of payment systems. The emphasis in this paper is on the allocation of efforts by the hospital on cost reduction and quality enhancement activities. Our results indicate that payment systems determine incentive trade-offs. One must consider the hospital's available counterstrategies to react to the policies and whether the use of such strategies can neutralize or further enhance the intended effects.

References

Allen, R. and P. Gertler, 1991, "Regulation and the Provision of Quality to Heterogenous Consumers: The Case of Prospective Pricing of Medical Services," *Journal of Regulatory Economics,* 3, 361–375.

Armstrong, M., R. Rees, and J. Vickers, 1991, "Optimal Regulatory Lag Under Price Cap Regulation," manuscript, Oxford University.

Brennan, T.J., 1989, "Regulating by Capping Prices," *Journal of Regulatory Economics,* 1, 133–147.

Cabral, L.M.B. and M.H. Riordan, 1989, "Incentives for Cost Reduction Under Price Cap Regulation," *Journal of Regulatory Economics*, 1, 93–102.

Clemenz, G., 1991, "Optimal Price-Cap Regulation," *Journal of Industrial Economics*, 39, 391–408.

Ellis, R.P., 1993, "Creaming, Skimping and Dumping: Provider Competition for Patients," Boston University Discussion Paper.

——— and T.G. McGuire, 1986, "Provider Behavior under Prospective Reimbursement," *Journal of Health Economics*, 5, 129–151.

Dranove, D., 1984, "An Empirical Study of a Hospital-Based Home Care Program," *Inquiry*, 22, 59–66.

———, 1987, "Rate-setting by Diagnosis Related Groups and Hospital Specialization," *Rand Journal of Economics*, 18, 417–427.

Holmstrom, B. and P.R. Milgrom, 1991, "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," *Journal of Law, Economics, & Organization*, Special Issue, 7, 24–52.

Lyon, T.P., 1993, "A Model of Sliding-Scale Regulation," manuscript, Indiana University.

Newhouse, J.P., 1983, "Two Prospective Difficulties with Prospective Payment, or, Its Better to Be a Resident than a Patient with a Complex Problem," *Journal of Health Economics*, 2, 269–274.

———, 1984, "Cream Skimming, Asymmetric Information, and a Competitive Insurance Market," *Journal of Health Economics*, 3, 97–100.

Pauly, M., 1984, "Is Cream Skimming a Problem for the Competitive Medical Market?" *Journal of Health Economics*, 3, 87–95.

Pope, G.C., 1989, "Hospital Nonprice Competition and Medicare Reimbursement Policy," *Journal of Health Economics*, 8, 147–172.

Russell, L.B, 1989, *Medicare's New Hospital Payment System*, Washington, DC: The Brookings Institution.

Shleifer, A., 1985, "A Theory of Yardstick Competition," *Rand Journal of Economics*, 16, 319–327.

Spence, A.M., 1975, "Monopoly, Quality and Regulation," *Bell Journal of Economics*, 6, 417–429.