Journal of Health Economics 5 (1986) 129–151. North-Holland

PROVIDER BEHAVIOR UNDER PROSPECTIVE REIMBURSEMENT

Cost Sharing and Supply

Randall P. ELLIS and Thomas G. McGUIRE*

Boston University, Boston, MA 02215, USA

Received April 1985, final version received March 1986

This paper develops a model in which physicians choose the level of services to be provided to their patients. We show that if physicians undervalue benefits to patients relative to hospital profits, prospective payment, a system in which hospitals receive a payment dependent on the diagnosis-related group within which a patient falls, can lead to too few services being provided. In contrast, a 'cost-based' reimbursement system is shown to result in too many services being provided. Competition between hospitals for physicians will tend to augment both of these problems. We also examine a mixed reimbursement system, in which hospital reimbursements are paid partly prospectively and partly cost-based. This system is shown under a variety of circumstances to be superior to the other two reimbursement systems by improving the incentives for the efficient level of services, reducing incentives to unnecessarily admit or reclassify patients, and reducing risk to providers.

1. Introduction

The Health Care Financing Administration's (HCFA) Medicare program, the largest single buyer of health care in the United States, began paying hospitals prospectively for inpatient care starting in 1983. After the system is fully in place by 1987, hospitals will receive, with certain exceptions, a payment dependent on the diagnosis-related group (DRG) within which a patient falls.[1] Other payers will be following Medicare's lead, in some cases, involuntarily. A few states now require all payers within the state to pay by DRG for hospital care.

Prospective payment shifts cost-control policy away from demand-side interventions, such as deductibles and coinsurance, to cost sharing on the

*Research for this paper was supported by a grant from the Pew Charitable Trust. Initial work on some of the ideas in this paper was performed under contract from the National Institute of Mental Health to Professional Management Associates, Inc. We are grateful to Jerry Cromwell, Richard Frank, Kyle Grazier, Howard Goldman, Will Manning, Susan Payne, Chris Ruhm, Pankaj Tandon, and two anonymous referees for helpful comments on earlier versions of the paper.

[1]For a description of the details of Medicare's prospective payment system, see Vladeck (1984).

supply side. By paying only a prospective amount, HCFA requires the hospital or other provider to bear the marginal costs of treatment. Cost sharing on the supply side may be superior to cost sharing on the demand side for three reasons. First, cost sharing on the demand side is tightly bound by Zeckhauser's (1970) dilemma: the tradeoff between risk spreading and appropriate incentives, as he called it. The more insurance, the less financial risk the patient bears, but the greater is the efficiency loss to moral hazard. The Zeckhauser tradeoff will be less acute on the supply side. Most hospitals discharge many hundreds of Medicare patients a year. For the institution, risk due to random variation in treatment costs is small.[2] Since a hospital can bear risk more readily than a single patient, cost sharing on the supply side can more easily be geared to providing appropriate incentives.

Second, providers may be a better target for cost containment measures than consumers simply because they are better able to control costs. For certain types of care, such as inpatient surgery, the physician or hospital is often the dominant decision-maker, with only a modest role for the patient. Demand-side incentives may have less effect than supply-side incentives.

Third, as a practical matter, it is difficult to take away health insurance coverage from a population in the interest of controlling cost. Medicare beneficiaries, unions, and other groups have all stood firm in resistance to increased demand-side cost sharing. In contrast, imposing cost sharing on providers has been surprisingly easy.

How well prospective payment works to promote cost-effective health care depends on how providers respond to the change in the method of reimbursement. One concern is that providers will 'game' a prospective system by reclassifying patients into higher paying categories. Gaming will affect social costs if providers increase the number of admissions or undertake procedures, such as operating room procedures, that move a patient into a higher paying classification. To the extent that gaming takes the form of a simple reclassification of patients, with no additional resources consumed, it will increase the payer's cost, but this will be offset by an increase in the net revenue of the provider.[3] The main issue of social cost and benefit with prospective payment is the degree to which the quantity of treatment supplied is affected by supply-side cost sharing, and whether this reponse is more or less than optimal.

This paper develops a simple model of provider response to the cost

[2]This statement refers to unsystematic risk under prospective payment. Systematic risk, risk associated with receipt of an unrepresentative (e.g., more severe) case load within the classification system, is not ameliorated by the law of large numbers. We consider systematic risk briefly below. For discussion of the types of risks faced by hospitals under prospective payment, see McGuire (1985).

[3]Simborg (1981) popularized the term 'DRG-creep' for this aspect of gaming. Carter and Ginsburg (1985) examined the empirical significance of DRG creep and found that it explains about three quarters of the increase in the case mix index observed between 1981 and 1984.

sharing imposed by prospective payment. Prospective payment is intended to reduce the quantity of hospital services used. How much response is enough and how much is too much has not been specified. Our model suggests a standard by which to judge the desirable degree of supply response. The unit of analysis is the hospital episode. We will be concerned primarily with the quantity of services supplied during an episode, deferring until after the main development of the model any discussion of the effect of alternative reimbursement systems on the number of admissions.

We model the physician as the key decision-maker who selects the level of services to be provided to a patient and trades off the benefits to the patient against the benefits to the hospital. In the terminology of the principal–agent literature, the physician is the 'agent' for two 'principals': the patient and the hospital. While this is obviously a simplification, it allows us to incorporate in one decision-maker the conflicting interests created by prospective payment systems. Alternative reimbursement systems change the incentives placed on providers of care. Under the old, cost-based reimbursement system, the patient and the hospital shared a general interest in more treatment – the patient for more benefits and the hospital for more revenue. This alliance is struck down by prospective payment. The more care a hospital provides under prospective payment, the lower the net revenues it receives. How the interests of patients and providers are balanced out is the key to the success of the new payment systems. As we will show, prospective payment provides the right incentives to maximize social welfare only in the extreme case in which the physician is a 'perfect agent'. The perfect agent, as we define it, values one dollar of benefit to the hospital (i.e., profit) equally with one dollar of benefit to the patient.

In formulating this model, we are particularly concerned with the possibility that physicians are not perfect agents but, because of the economic power of hospitals, are induced to be more attentive to the financial interests of providers. Our model allows this important possibility to be easily represented. To deal with the potential problem, we propose a mixed reimbursement system, which is part prospective and part cost-based, and can be regarded as a generalization of the form of payment systems now in place. This system is shown to mitigate several likely problems with a prospective payment system. Specifically, the mixed system reduces the risk to providers, reduces the incentives to admit and reclassify patients and reduces the tendency for undesirable competitive behavior by hospitals. These desirable features are summarized at the close of the paper.

2. A model of physician response to cost sharing

The model developed in this section focuses on the response of physicians to the cost sharing imposed by prospective payment. Three actors are

involved: patients, who are assumed to be fully insured and accept the prescribed treatment; physicians, who make the decisions about levels of treatment; and hospitals. Treatment to patients is assumed to be provided by both hospitals and physicians. Reflecting current reimbursement practice, reimbursements for physician services are assumed to be cost based, and separate from hospital services. Initially, we make physician compensation invariant to the costs incurred and services provided by hospitals, although later, when considering the impact of market competition, we allow for the possibility that hospitals also compensate physicians directly.

Before continuing it is worth noting that our approach is similar to that taken by Pauly (1980) and Pauly and Redisch (1973), who regard the physician as the key decision-maker, managing all inputs in the production of health services. Our model differs from that approach, however, in that rather than incorporating financial variables as constraints on physician behavior, we incorporate benefits to consumers and to hospitals as arguments directly in the physician's utility function.[4]

2.1. Patients

Let the patient's total benefits from health treatment during a single episode be $B(q,s)$, where q is the quantity of hospital services provided (e.g., room and board, and X-rays), and s represents physician inputs. To streamline the analysis, we ignore here any possible substitutability or complementarity between physician inputs and hospital services, and assume that the physician's input for a given episode is fixed. Hence we will write simply $B(q)$. This appears reasonable if, for instance, the physician decision-maker is a surgeon, all of whose inputs occur on the first day.[5] Marginal benefit is $b(q)$. As depicted in fig. 1, the total benefit function is assumed to reach a maximum at some quantity q', after which point total benefits fall. Total benefits will fall after q' both because of the time-price of receiving treatment and because of the risk of infection and other iatrogenic illness associated with a continued hospital stay. Benefit to the patient, measured in dollar terms, is assumed to be equal to the full social benefit from treatment. Implicit in this is the notion that the benefits are those that would be perceived by a well-informed consumer. Patients are assumed to passively accept treatment prescribed by their physicians. Patients can be thought of as being fully insured for hospital care and willing to accept as much benefit

[4]Our approach also differs from that of Feldstein (1968), Newhouse (1970), or Feldstein (1971). The formulation is also similar to models of physician behavior where the physician is concerned about the 'ethics' of treatment [Evans (1974)]. Models of physician behavior are reviewed in Detsky (1978).

[5]The significance of substitutability between physician and other inputs is examined in the appendix. Reflecting levels of physician treatment substantially complicates the analysis without providing major new insights.

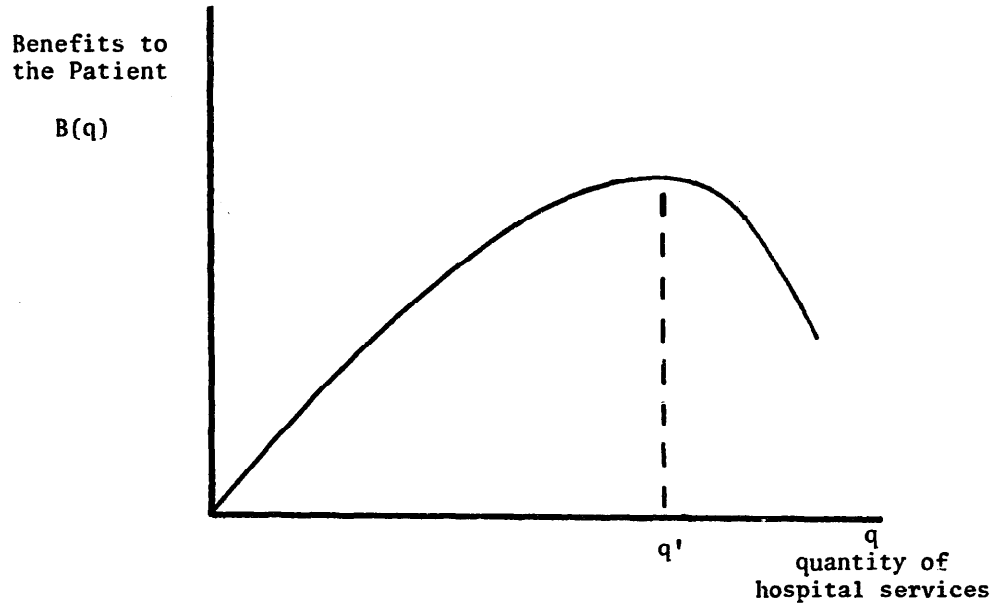

Fig. 1. A patient's benefit function.

from treatment as the hospital is willing to supply. We start our analysis by considering the case of a single patient before considering the case of numerous, diverse patients.

2.2. Hospitals

We will refer to net revenue, surplus, or profit to the hospital interchangeably,

$$\pi(q) = R(q) - C(q). \tag{1}$$

Eq. (1) states simply that profit on the treatment of a patient is equal to revenue less cost. Revenue, $R(q)$, depends on the payment system, while total cost, $C(q)$, depends on the cost of treatment. The marginal cost of treatment will be denoted $c(q)$. In what follows we will frequently focus attention on the simple case where the cost of hospital treatment is linear, so that total costs can be written as a constant, c, times the quantity provided.

2.3. Physicians

The physician is assumed to be interested in both the hospital's profit and the benefits to the patient. The utility function of the physician is thus $U(\pi(q), B(q))$. Given that profits are a good, and $B(q)$ reaches a maximum at q', typical indifference curves of the physician will look like the curves in fig. 2. The shape of the indifference curves shown depends upon the relative

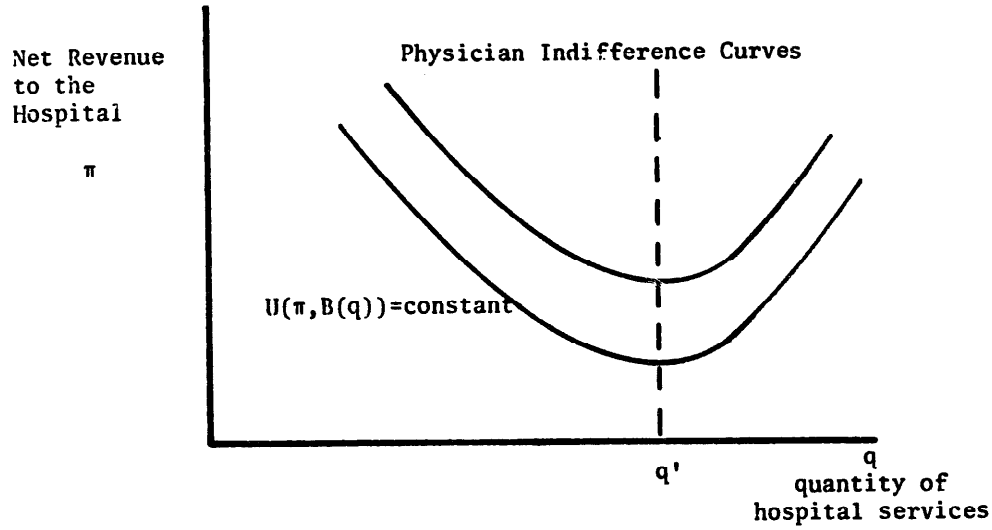

Fig. 2

weights placed by the physician on hospital profits and the benefits of treatment to the patient.

For a utility maximizing physician, the first-order condition for an optimum can be written as[6]

$$\frac{\partial U}{\partial B}\frac{\mathrm{d}B}{\mathrm{d}q}+\frac{\partial U}{\partial \pi}\frac{\mathrm{d}\pi}{\mathrm{d}q}=0. \tag{2}$$

For subsequent use, it will prove useful to write the physician's marginal rate of substitution between profits and quantity of treatment as follows:

$$MRS_{\pi,q}=\frac{(\partial U/\partial B)b(q)}{(\partial U/\partial \pi)}=MRS_{\pi,B}b(q)=\alpha b(q), \tag{3}$$

where $\alpha=MRS_{\pi,B}$. The variable α, which may itself depend on π and q, is equal to the rate at which the physician is willing to trade off one dollar of hospital profit for one dollar of patient benefit. As long as the net revenue and benefits for each patient are small relative to the total amounts for the hospital, then α will be nearly constant, and the physician's indifference curves will be vertically parallel. The parameter α can usefully be viewed as

[6]Throughout this paper we will generally assume that the second order conditions that are required hold. Sufficient but not necessary conditions for a global maximum are that $B(q)$ and $R(q)-C(q)$ both be strictly concave and that the utility function $U(\pi,B)$ be strictly quasi-convex. We also ignore in general the possibility of corner solutions. The possibility of some minimum level of services being needed to qualify for the prospective payment is considered below, however.

an index of the degree to which the physician, in its role as agent for the patient, takes into account the patient's interest.

When α is greater than one, the physician attaches too much importance to benefits to the patient relative to hospital profits. When α lies between zero and one, the physician considers patient benefit from treatment, but the patient's welfare receives less weight than the hospital's financial interest. In the extreme case of $\alpha=0$, the physician has no regard for the welfare of the patient and indifference curves are horizontal lines. When $\alpha=1$, we say the physician is a 'perfect agent': a dollar of benefit to the patient is given the same value as a dollar of net revenue to the hospital. Note that when $\alpha=1$ the slope of the hospital indifference curves are equal to the negative of the patient's marginal benefit function.

2.4. Physician behavior under cost-based payment

In an idealized cost-based reimbursement system, the revenue function for a hospital is identical to the cost function. Hence,

$$\pi(q)=R(q)-C(q)=0. \tag{4}$$

Since $d\pi/dq=0$, the physician who tries to maximize utility in such a cost-based system will choose the q such that $dB/dq=0$, i.e., the quantity of services that maximizes benefits to the patient. In a cost-based payment system, the physician will try to fully accommodate the demand of the fully-insured patient. This solution will describe the physician's decision as long as α is positive, that is, as long as the physician has any regard for the welfare of the patient.

The cost-based reimbursement system described here is one in which profits to the hospital are always zero. If the revenue generated from any hospital service actually exceeds marginal costs, then hospitals may be able to increase profits by providing services beyond q'. In such a setting, physicians may recommend (or tolerate) services even beyond the point at which benefits to the patient are maximized under cost-based reimbursement. This form of 'X-inefficiency' is a problem common to many regulated settings, and is one of the arguments in favor of prospective payment systems.

3. Physician behavior under prospective payment

In a purely prospective payment system, the revenue a hospital receives to care for a patient for a given episode of treatment is a constant, which we call a, and is independent of the quantity of services provided.

$$R(q)=a \tag{5}$$

and

$$\pi = a - C(q), \tag{6}$$

Using the first-order condition for utility maximization (2) and the definition of α as in (3), we can write

$$\alpha b(q) = c(q). \tag{7}$$

For the simple case where the marginal cost of treatment is constant, we can write

$$\alpha b(q) = c.$$

The physician's choice of the quantity of services to be supplied by the hospitals, as described by (8), is illustrated in fig. 3. Physician utility is maximized subject to a net revenue function which falls at a rate of c for each unit of service provided. The higher net revenue line shown in fig. 3 goes through the point of original supply (O, q'). In this case, the prospective amount a is equal to cq'. The quantity chosen, q^*, is where the slope of the physician's indifference curve between profits and the level of hospital services is equal to the marginal cost.

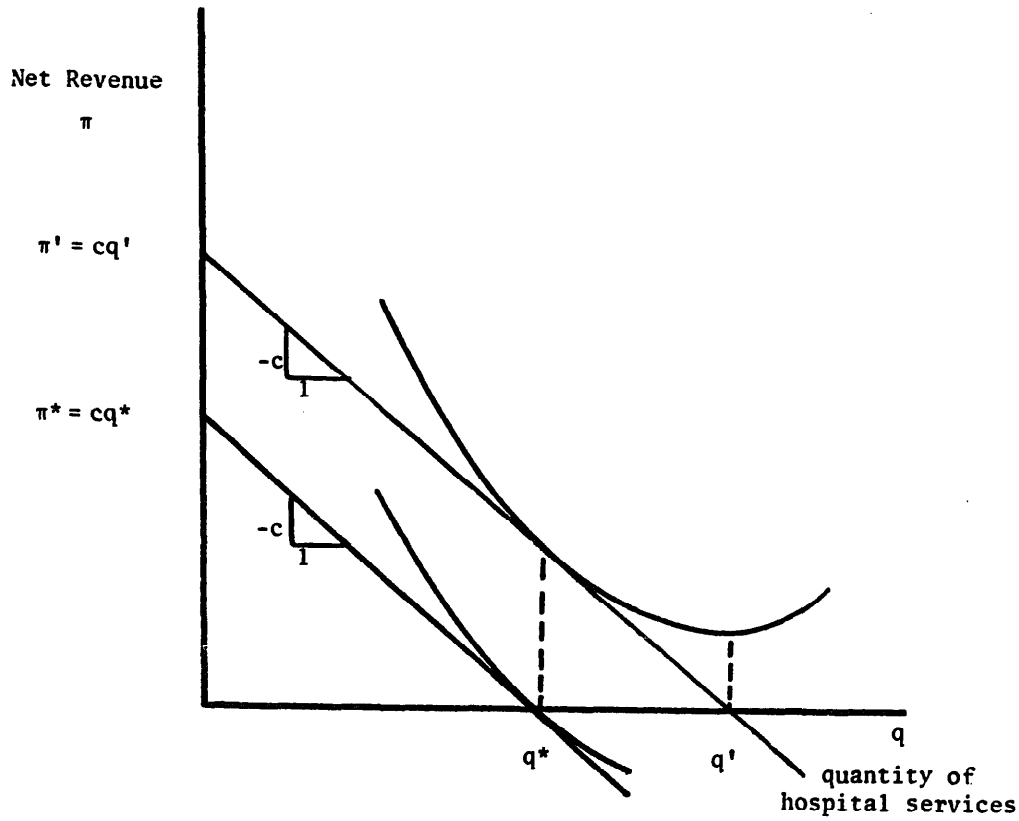

Fig. 3. Physician's choice of hospital services under prospective reimbursement.

3.1. Perfect agency: The promise of prospective payment

In a prospective payment system, the hospital bears all of the marginal costs of treatment while the patient derives all of the marginal benefits. In order to make the efficient quantity choice, the physician must fully take into account both the marginal cost and the marginal benefits of treatment. This takes place when $\alpha = MRS_{\pi,B} = 1$, i.e., the physician weighs a dollar of benefit to a patient equally with a dollar of profit to the hospital. It can be seen by (8) that when $\alpha = 1$, and the physician is a perfect agent, the physician sets marginal benefit equal to marginal cost.

If the payer sets the prospective amount so as to cover the costs of services provided under a cost-based system, cq', then hospitals will make profits. These profits are likely to be temporary, however. After a few years of experience with prospective payment, DRG's are scheduled to be 're-calibrated'. The new payment levels for each DRG category will be reduced approximately in proportion to the average supply response. Hence if the level of services is reduced to q^*, the new prospective payment amount will be cq^*. This net revenue line is shown in fig. 3. Reducing payment to this level maintains the marginal incentives to physicians to reduce the supply of hospital services, but transfers the net gains from prospective payment away from the hospital to the payor and the patients.[7] If there are 'income effects' in the physician's preferences, indifference curves will not be vertical displacements of one another. The prospective payment amount will (by changing α) affect the quantity supplied. Nonetheless, after recalibration, if the physician is a perfect agent, the quantity supplied will be that which equalizes marginal benefit and marginal cost. Moreover, all of the benefits of the move towards the efficient supply will be transfered to the payer and patient. We regard this to be the promise of prospective payment.

3.2. Imperfect agency: The hazards of prospective payment

There is ample reason to be concerned that patient's interests will be lost in the restructuring of incentives in the health care system to reduce costs. Hospitals are powerful actors in the health care system, providing access to critical complementary resources to physician services. New contracting mechanisms and payment systems exert pressures to change patterns of care. In the context of a general oversupply of physicians, hospitals may be in a strong position to induce affiliated physicians to adhere to hospital-set standards of utilization. Limited evidence available to date suggests that new

[7]In a pure insurance context, all benefits would go to the enrollee/patient in the form of reduced insurance premiums. If premiums are partly subsidized, as in the case of Medicare, the group subsidizing the premiums shares in the benefits.

payment systems are having an impact, but that not all the reductions could be labeled 'wasteful' care.[8]

In our model, the financial interests of the hospital influence treatment decisions through the decision making of the shared agent, the physician. When α is less than one, the weight given to hospital net revenues is too great in relation to patient benefits. In the extreme case, the physician would be regarded as a tool of the hospital's interest, prescribing the minimum acceptable level of treatment to maximize hospital net revenue. This circumstance would occur when the hospital is in a very strong bargaining position relative to the physician. While this extreme circumstance is unlikely to occur, the market for health care may not be sufficiently competitive to justify belief that a hospital has no bargaining power relative to physicians. If patients are relatively ineffective in making economic choices to further their own self-interest, the incentives in prospective payment will lead to the organized, powerful hospital to have more influence on physician behavior. We believe the most likely case is for α to lie between 0 and 1, meaning the physician is an imperfect agent for the patient, giving too much weight to the hospital's financial interest. In this circumstance, it is easy to see, by our eq. (8), that too little treatment would be provided. This is a theoretical framework for understanding the appearance of 'undersupply' in prospective payment. The remainder of the paper is concerned with developing a payment system that leads to efficient supply of care in the imperfect agent case.

4. An alternative: A mixed reimbursement system

Consider now a revenue function which combines a prospective amount, a, with a partial cost-based payment. Let the fraction of costs paid be r with $0<r<1$. Here we can write

$$R(q)=a+rC(q). \tag{9}$$

In this mixed system, the prospective amount can be lowered as r is increased to maintain the same total payment per case. As we will now show, a positive payment per unit of q provided is necessary to induce the hospital to supply the efficient quantity when the physician is an imperfect agent.

For the case of a linear cost function we can write

$$\pi=a+(r-1)cq, \tag{10}$$

[8]See Davis et al. (1985) for a recent review of the evidence on the effects of prospective payment. The *New York Times* (Jan. 9, 1986, p. A11) recently reported that the Senate Special Committee on Aging, working with the General Accounting Office and the Inspector General of the Department of Health and Human Services, found that patients were being discharged 'quicker and sicker' under Medicare's new prospective payment system.

and the physician's problem is to

$$\max_q U(a+(r-1)cq, B(q)), \tag{11}$$

the solution to which is characterized by the first-order condition

$$\frac{\partial U}{\partial \pi}(r-1)c+\frac{\partial U}{\partial B}b(q)=0, \tag{12}$$

or

$$\alpha b(q)=(1-r)c. \tag{13}$$

For efficiency, we require $b(q)=c$, so the optimal r is described by

$$r=1-\alpha. \tag{14}$$

The optimal cost sharing on the supply side $(1-r)$ is equal to α, the index of agency. When $\alpha=1$, a perfect agent (i.e., the physician) will fully take into account benefits to the patient, and no additional hospital reimbursement is needed. It is optimal that the supplier bears the full marginal cost. As α falls, some decrease in cost sharing brought about by an increase in marginal revenue is necessary to counter the tendency of the imperfect agent to call for an undersupply of services.[9]

Fig. 4 illustrates graphically the difference between a prospective payment system with imperfect agents and the mixed reimbursement system proposed here. For the case where $\alpha<1$ the physician's indifference curves look flatter than those of the perfect agent, and treatment would be undersupplied (at q'') if a full prospective payment system were in place. Under a mixed system, the physician can be induced to call for treatment q^*. With the appropriate choice of r, the optimal quantity of treatment can be obtained. In the next section, we demonstrate that α has an empirical counterpart that can be used to guide this choice.

In a small way, the Medicare prospective payment system is already mixed because of the so-called 'outlier' policy. For a case in which the actual length of stay exceeds the geometric mean in a DRG by 1.94 standard deviations, or 20 days, whichever is less, or in which costs exceed the amount ordinarily payable by 50 percent or $12,000, the hospital receives some per diem reimbursement. This reimbursement applies, however, only to days exceeding the cutoff and affects only a small portion of all episodes.[10] The intention of

[9]This result confirms a theorem in the principal–agent literature that the optimal sharing ratio will be less than unity. See, e.g., Shavell (1979).

[10]The per diem payment is equal to approximately 60 percent of the average cost per day within a DRG. See Vladeck (1984) for more discussion.

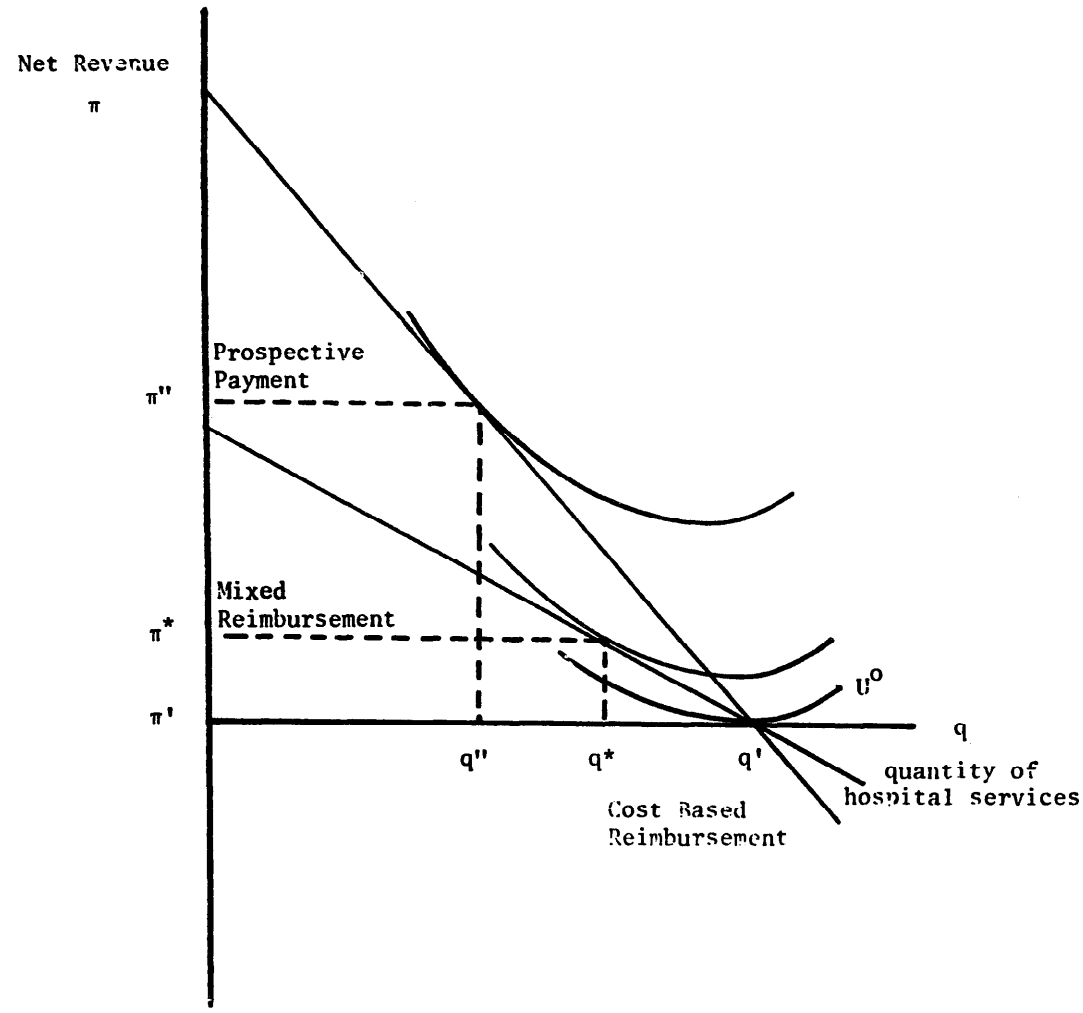

Fig. 4. A comparison of cost-based reimbursement, prospective payment, and mixed reimbursement.

the outlier policy is to reduce risk to facilities of high cost cases. Since the incentive effects within the outlier policy apply only to a small fraction of cases, their impact on provider behavior is less than in the mixed system proposed here.

5. Extensions of the basic model

In this section we examine several extensions of the basic model. The mixed system proposed here has a number of other advantages over a pure prospective system. These have to do with incentives regarding admissions and risk to providers in a prospective payment system. We first use the model to consider the case of diverse patient types falling in a single DRG. Here the risk reduction for providers resulting from the mixed reimbursement system is clearly seen. Next, the impact of prospective payment versus mixed reimbursement on the number of hospital admissions is addressed.

The likely impact of market competition is then considered. Finally, we use the model to consider the normative question of how one should decide whether the supply response from prospective payment is too much or too little.

5.1. Diversity of patient types

So far we have only considered the behaviour of a physician towards a single patient. Here we examine the impact of diverse patients in a prospective payment system. The diversity of interest in our model is diversity not accounted for in the patient classification system, such as DRGs. Two types of diversity are possible: diversity in the benefits from treatment, and diversity in the cost of providing treatment. We consider here the problem of diverse benefits, which allows us to still consider a single cost function.[11] In terms of our model, diversity in patient needs results in different total benefits functions, $B(q)$. Fig. 5 illustrates the case where there are two patients, one of whom benefits from a higher level of treatment than the other. The provider indifference curves shown reflect those of an imperfect agent. Three different levels of hospital treatment for each of the two patients are identified.

Under full cost reimbursement, quantities q_1' and q_2' are provided, where the indifference curves of the physician appropriate for each of the two individuals are just tangent to the horizontal axis. Under prospective payment, quantities q_1'' and q_2'' are provided: because of the imperfect agent problem, too little care is provided. Finally, under mixed reimbursement, quantities q_1^* and q_2^* are provided. The same fractional reimbursement rate r would be optimal if the provider weighs benefits to both patients equally.

An interesting and important outcome of the mixed reimbursement system proposed here is that it also reduces the risk to providers in net revenue per patient relative to the prospective payment system. As can be seen in fig. 5, profits vary more widely under a prospective payment system. This occurs because providers are partially reimbursed for incurred costs and hence are in effect partially insured. This partial insurance effect carries over regardless of whether the risk faced by the provider is systematic or unsystematic.

5.2. Incentives to admit and reclassify

A prospective payment system employs incentives to reduce treatment, except at one critical point: the decision to admit a patient. Questionable or

[11]The corresponding case of diverse costs with the same benefit function leads to findings which are the same as the case of diverse benefits considered here: namely that the same fractional reimbursement rate r will be optimal if the provider weighs benefits to all patients equally; and mixed reimbursement reduces provider risk.

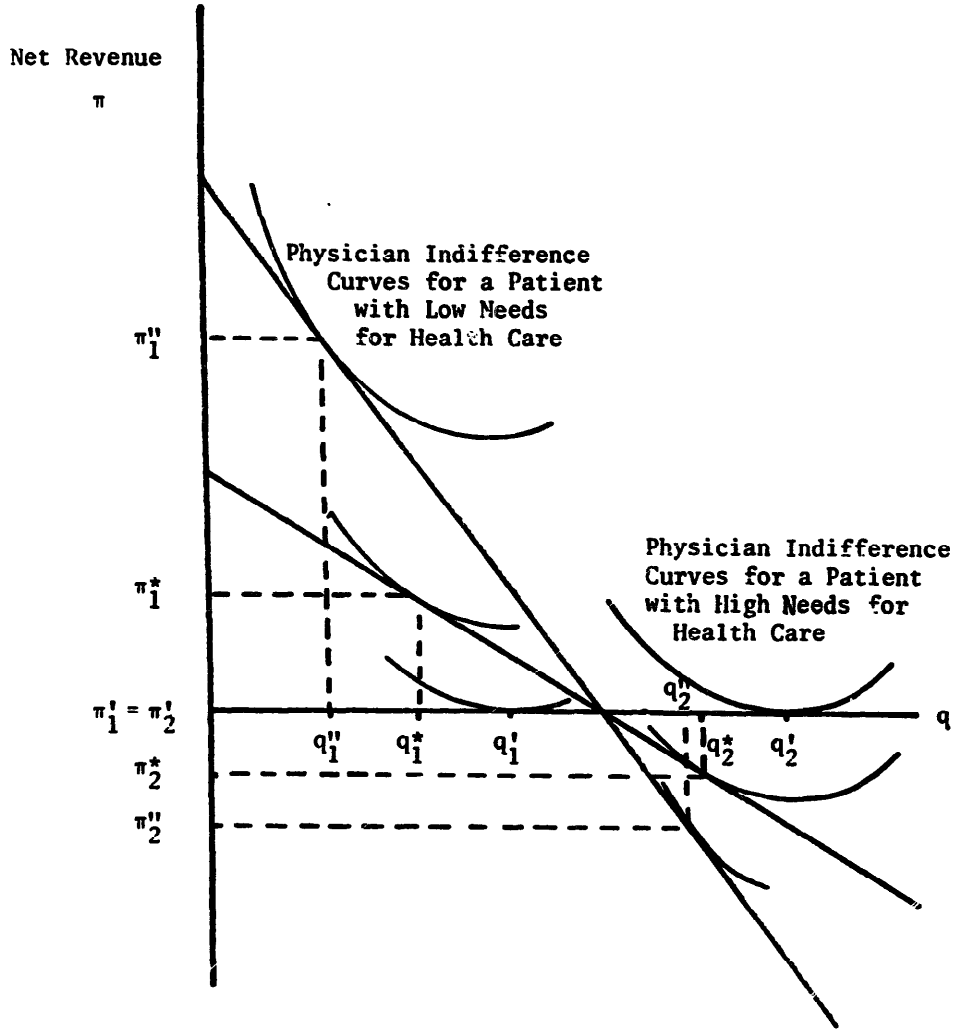

Fig. 5. Alternative reimbursement systems with two diverse patients.

low-value admissions are, in fact, especially profitable for hospitals. The full prospective amount is collected for an admission with a lower than average cost. This is of great concern to HCFA; it is primarily through increased admissions that the cost control goals of prospective payment may be thwarted.

The mixed reimbursement system also holds out the prospect of reducing the incentive for excessive admissions. Since hospitals are given a smaller lump sum for admitting patients under a mixed system, there is less incentive to admit low-cost patients.

A prospective payment system is feasible only if there is some minimum level of treatment (perhaps only one day) which must be provided in order to qualify for a prospective payment. Call this amount q^{min}. We define an admission to be inappropriate if $B(q) < C(q)$ for all $q > q^{min}$; i.e., the benefits of the admission are less than the costs at all levels of treatment greater than q^{min}. If the marginal benefit is assumed to be declining faster than marginal

cost at q^{min}, then the above condition can be shown to be equivalent to the condition that $B(q^{min}) < C(q^{min})$.

The problem of inappropriate admissions can then be conceptualized as follows. The provider will choose to admit a patient as long as

$$u(0, B(0)) < U(R(q^{min}) - C(q^{min}), B(q^{min})).$$

From this it can be readily seen that as long as q^{min} is below the average level of treatment for the DRG, then the revenue from the mixed system will be less than that of a prospective payment. Since the costs and benefits to the consumer remain the same at q^{min} in either payment system, the lower revenue under mixed reimbursement implies a lower incentive for excessive low-value admissions. Even the perfect agent will tend to admit low-value cases too often.

This same result is shown diagrammatically in fig. 6. The case shown is one where a physician would recommend admitting a patient under prospective payment, but would be discouraged from doing so under a mixed

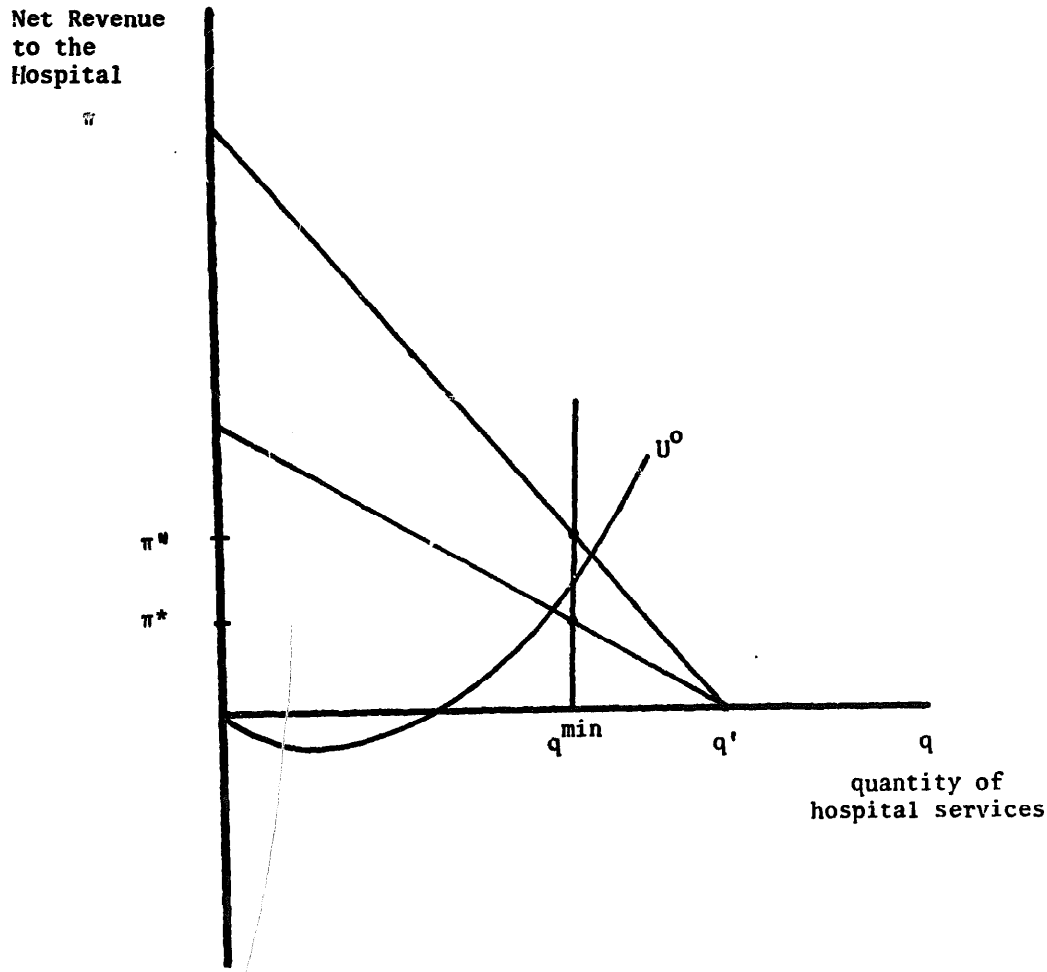

Fig. 6. A comparison of the incentives for inappropriate treatment.

reimbursement system. The relatively flat provider indifference curves reflect the fact that the patient benefits little from treatment. Relative to no treatment, the physician shown is made better off at q^{min} under a prospective payment system, but worse off under a mixed system. Partial marginal cost reimbursement will not always achieve this desired result, but it is notable that it at least moves in the right direction.

Prospective payment introduces the opposite problem with respect to high-cost admissions. If a physician anticipates that the hospital will incur a loss on a patient, that patient may be denied admission, even if the benefit of treatment exceeds the costs. This is the major 'access' issue with prospective payment. Once admitted, the perfect agent would provide the optimal care, but a hospital will have incentives to avoid accepting responsibility for patients likely to impose losses. Just as the mixed system, by reducing expected profits, decreases incentives to admit low-value cases, the mixed system decreases expected losses on high-cost cases, reducing the incentive to deny admission. This behavior is already illustrated in fig. 5 for the 'high need' patient.

The incentives introduced by a prospective payment system for 'DRG-creep' have already been mentionned above. As long as providers can increase their revenues without raising costs by choosing higher revenue-generating DRGs for their patients, then provider will have incentives to do so. The mixed reimbursement system, by providing a smaller prospective amount for each case, also reduces the incentives for this type of behavior.

5.3. Market competition

An important issue to be considered when examining the impact of alternative reimbursement mechanisms is their impact on market competition. Recent research by Farley (1985) and others indicate that the degree of competition among hospitals does have an important impact on hospital costs. The specific nature of competition in the health services sector is very poorly understood, however, with informational imperfections being an important general feature.

Available evidence suggests that health care consumers do a very limited amount of shopping around among physicians, and that, having chosen a physician, consumers accept most physician's recommendations quite passively. Patients do even less shopping around among hospitals for an inpatient episode. Indeed, the sporadic and often unforseeable nature of many admissions makes this quite difficult. In most cases the physician remains the key decision maker as to where, when, and for how long a patient will be hospitalized.

In contrast, hospitals appear to compete more or less actively for physicians, thereby attracting patients. Making office space and parking

available, offering the latest in medical equipment, and in some cases, making direct compensation to physicians, often for minimal tasks, are all mechanisms through which hospitals compete to attract physicians. The reimbursement system can have an impact on the *type* of physician and the *type* of physician patients that hospitals will try to attract.

Under cost-based reimbursement, as long as the reimbursement system more than covers costs, the physician goals of maximizing both benefits to the patient and profits to the hospital are perfectly compatible. Both the patient and the hospital would like to continue treatment as long as the benefits are positive. Hospitals have incentives to attract physicians who attach a great deal of importance to benefits to patients, and who believe in providing extensive hospital services. In the notation of our model, hospitals would like to compete most actively for 'high α' and 'high q' physicians. Both factors tend to increase the levels of hospital services provided, and increase hospital profits.

Under prospective payment, the system is exactly reversed. Since hospital profits decrease as the level of services increase, hospitals would like to compete for 'low α' and 'low q' physicians, i.e., physicians who place less weight on patient benefits relative to hospital profits, and physicians whose patients require fewer hospital services per episode.[12] In response to prospective payment we should expect to see competitive pressure on physicians to attach less importance to patient benefits and reduce the level of services provided by hospitals. 'High α' and 'high q' physicians may see their affiliations dropped from the most cost-conscious hospitals.

It is frequently asserted that 'competitive pressures' under prospective payment will keep hospitals from undersupplying services. We do not see the mechanism through which this will take place under prospective payment. If we assume, for the minute, that hospitals do compete directly for patients rather than physicians, then one has to recognize that not all patients are attractive to hospitals under prospective payment – only patients whose costs are less than the prospective payment amount. How are hospitals to attract low-cost patients and discourage high-cost ones?

Under cost-based reimbursement hospitals are interested in attracting all kinds of patients, since utilizing hospital capacity was an important means of lowering average fixed costs and increasing profits. Promoting a reputation for high quality services was a worthwhile strategy. Under prospective payment, such a strategy is not very attractive, since high quality services and a good reputation both reduce short run profits and are just as likely to attract high-cost as low-cost patients within a given DRG. In fact a good reputation may be more important to high-cost patients, who should be

[12]For instance, if there are two alternative modes of treatment for a patient that falls in a given DRG, hospitals will prefer the physician using the mode of treatment which requires less hospital resources.

more willing to search among hospitals for the best alternative. Promotional efforts targeted at attractive populations are one possible strategy, but this could be done by both low and high quality hospitals. Threats of malpractice suits will undoubtably work as a constraint on the most extreme forms of negligence by hospitals, however careful but profit conscious hospitals may still be able to reduce levels of services below what is considered acceptable.

The mixed reimbursement system discussed here moderates the competitive pressures on physicians that appear likely under prospective payment. By reducing the gains and losses to be made on each patient, hospital efforts to attract the right kind of physician, or the right kind of patient will be reduced. Together with the previously considered features, a mixed reimbursement system holds out the prospect of appropriately balancing the incentives for too much or too little treatment under cost-based or prospective payment systems.

5.4. *What is the desirable supply response?*

The purpose of prospective payment is to reduce the cost of hospital care. Are the financial incentives to reduce costs too strong or too weak? Is an observed reduction in quantity supplied enough or too much? We answer those questions here by using the supply response of the perfect agent as a standard against which to judge observed behavior.

Eq. (13) implicitly defines a supply curve to a single patient for a single episode of treatment. We need to emphasize that the supply curve derived here is not the usual supply curve based on the marginal cost of treatment. Instead, our supply curve reflects the physician's weighting of a perfectly informed consumer's own benefit from treatment.

For a given supplier, we can regard α as constant and differentiate (13) with respect to the payment per unit, rc, to find how quantity supplied will change with a change in the marginal payment. Letting rc be called p,

$$\frac{dq}{dp} = -\frac{dq/db}{\alpha}. \tag{15}$$

Eq. (15) states that the slope of the supply curve, dq/dp, is equal to the negative of the slope of the demand curve (dq/db) divided by α. When $\alpha = 1$, the marginal cost of treatment is constant, and the hospital is a perfect agent, then the supply response to cost sharing will be at the same rate as the demand response to cost sharing would be. Multiplying both sides of (15) by p/q transforms this equality into a statement about the relative elasticities. The normative standard for supply response suggested by our model is as follows. The supply response to prospective payment should be equal to the

demand response of a perfectly informed consumer to the elimination of insurance coverage.

5.5. A test for agency

The results presented above suggest how α can in principle be measured. If the elasticity of demand and the elastiticy of supply to alternative reimbursement rates can each be estimated consistently for a given market, then the ratio of the two can be used to infer the magnitude of α. Two conditions are needed for this measure to be precise: the consumer demand curve must be the same as the marginal benefit curve, and the marginal cost of treatment must be constant. Although these two conditions are unlikely to be met exactly, the ratio of these two numbers may nonetheless be a useful approximation.

Evidence about the elasticity of supply to the reimbursement rate is scant. Frank and Lave (1984) provide one such study for the supply of inpatient mental health services in response to coverage ceilings in Medicaid and find the response to be highly significant. Comparisons between the levels of treatment when care is free in a fee-for-service system and when care is provided by health maintenance organizations (HMOs) also provides an indication of the degree of supplier response. Luft (1980) surveys the literature and presents evidence that reductions in hospital admission rates of up to 40 percent may be achieved by HMOs. Manning et al. (1984) as part of the Rand *Health Insurance Study* (HIS) analyzed a single HMO in Seattle and found hospital admission rates to be about 40 percent lower in the HMO and overall expenditures to be about 25 percent lower compared to a fully insured fee-for-service system. This suggests that there is a significant degree of supplier response to the reimbursement rate.

Interpretation of relative demand and supply elasticities in terms of the revealed degree of agency of the provider is not straightforward. For one thing, the supply response in an HMO will be modified by the need to attract enrollees from fee-for-service plans which may provide a generous quantity and quality of care. This quality competition may predominate when premium enrollment costs are generally subsidized by employers. It can be expected that if not just one but all suppliers in an area were paid prospectively, the supply response by each provider would be much greater.[13] A provider might be unconcerned about quality competition for a captive or undesirable population, such as the Medicaid population studied by Frank and Lave (1984).

[13]The Minneapolis area, with its predominance of HMOs, may be a good example of this phenomenon. See Iglehart (1984) for a description of competition and utilization rates in Minneapolis.

6. Conclusion

Prospective payment imposes cost sharing on the supply side. It is an extreme form of cost sharing, analogous to no insurance on the demand side. Prospective payment will lead to efficient supply only in the unlikely case that the physician acts as the perfect agent of the payer, weighing the patient's benefit equally with the hospital's financial interests. If, as appears likely, physicians are induced to place more importance on hospital profits than on patient benefits, incentives in a prospective payment system will lead to an undersupply of services. Even in the case of perfect agency, prospective payment introduces undesirable incentives, such as the incentive to unnecessarily admit low-cost cases.

The paper develops a simple model of physician behavior to clarify the incentives of prospective payment. In addition to being a positive tool, the model suggests a normative standard for supply response. This paper proposes that the welfare standard for supply response be what demand response would have been to the same degree of cost sharing by a fully informed consumer. Evidence on supply response in health care is very limited; that which does exist suggests that the response is at least as great as response on the demand side. Our model provides a way to interpret this result. Providers are less than perfect agents.

Imperfect agency is our main argument for a mixed system of reimbursement, combining a prospective component and a component proportional to resources used. This more general payment method moderates the strong incentives to reduce care under prospective payment and will lead to a more efficient quantity of services under the conditions of imperfect agency. A mixed system was also demonstrated to discourage low-value admissions (relative to a pure prospective system) and discourage avoidance of high-cost cases.

In practical terms, inertia in the medical system should ameliorate any immediate concern about undersupply of services in response to prospective payment. Threat of malpractice, physican practice patterns, and patient expectations will constrain the elasticity of supply in the short run. Over a longer time, however, we expect that hospitals will compete for physicians that attach great importance to hospital profits, and constrain the cost and quantity of hospital services. Inadequate levels of hospital services loom as a definite possibility. It may be hard to imagine a U.S. health care system that does not respond with its nearly full technological capability to a patient's problem. It also seemed unlikely, at one point, that the long run response to be higher energy prices would be so great. Policy makers should be wary of overshooting with reimbursement policy, based on a short-sighted belief that the medical system is unresponsive to financial incentives. Twenty years of governmental financing of health care tell us otherwise.

Cost sharing on the supply side holds promise for policy. A complete

switch from no to full cost sharing on the supply side is, however, too radical a shift. The mixed system proposed here, combining cost-based reimbursement and prospective payment, is more likely to lead to an efficient level of supply, fewer inappropriate admissions, more appropriate competition among hospitals, less 'DRG-creep', and, at the same time, lower risk to providers.

Appendix

This appendix explicitly examines the case where the physician selects his/her own time input, s, as well as the hospital level of services, q. Both are assumed to affect benefits to the patient.

In this case we can write the physician's utility function to be maximized as

$$U(\pi(q), B(q,s), s). \tag{A.1}$$

The first-order conditions for a maximum are then

$$\frac{\partial U}{\partial \pi}\frac{\mathrm{d}\pi}{\mathrm{d}q}+\frac{\partial U}{\partial B}\frac{\partial B}{\partial q}=0, \tag{A.2}$$

$$\frac{\partial U}{\partial B}\frac{\partial B}{\partial s}+\frac{\partial U}{\partial s}=0. \tag{A.3}$$

Eq. (A.3) can be seen as implicitly defining a function $s^*(q)$, the desired provision of physician services conditional on the level of hospital services. Using $s^*(q)$ and $MRS_{\pi,B}=(\partial U/\partial B)/(\partial U/\partial \pi)$, eq. (A.2) can be written with explicit arguments as

$$MRS_{\pi,B}(q,s^*(q))\frac{\partial B(q,s^*(q))}{\partial q}+\frac{\mathrm{d}\pi(q)}{\mathrm{d}q}=0,$$

or, equivalently, using the notation in the main text,

$$\alpha(q)b(q)+\frac{\mathrm{d}\pi(q)}{\mathrm{d}q}=0. \tag{A.4}$$

This expression could be used in place of eq. (2) and (3) in the paper, and all of the same basic arguments could be made about the incentive effects of different reimbursement systems using this specification. The main text explicitly considers the case where $\partial s^*/\partial q=0$, i.e., the physician's inputs are independent of hospital service inputs.

Eqs. (A.2) and (A.3) can be further interpreted once something is said about the reimbursement mechanism for physicians. One plausible assumption is that the physician is reimbursed by a flat wage, w, per unit of time input. If this wage fully and exactly compensates the physician for his time input, then $\partial U/\partial s=0$ always. The physician will be indifferent between providing more or less services to the patient, other than for the benefits to the patient, and will provide s up to the point where $\partial B/\partial s=0$. This result is symmetric to the result for hospital services under cost-based reimbursement.

Under the assumption that q and s each have constant opportunity costs of c and w, respectively, the socially optimal levels of inputs of q and s are found by maximizing over q and s,

$$B(q,s)-cq-ws. \tag{A.5}$$

Note that any profits to hospitals and physicians are neglected since they are merely transfers to providers. From (A.5) it follows that services should be provided up to the point where

$$\frac{\partial B}{\partial q}=c, \tag{A.6}$$

$$\frac{\partial B}{\partial s}=w. \tag{A.7}$$

The 'first best' optimum characterized by (A.6) and (A.7) is unlikely to be met by any of the reimbursement mechanisms considered in this paper, since none of them will correct the likely overprovision of services by physicians due to cost-based physician reimbursement. In general, (A.6) will not be optimal given that (A.7) is not satisfied, and 'second best' optima need to be considered.

The optimal reimbursement system for hospitals in the 'second best' world in which physician reimbursement is cost based depends critically upon whether hospital and physician services are substitutes or complements. This relation is summarized by $s^*(q)$. If s and q are substitutes, then reducing compensation to q increases s, mitigating the cost savings, and prospective payment becomes less attractive. If on the other hand s and q are complementary, then prospective payment may also reduce physician services. In terms of our mixed system, substitutability increases the optimal reimbursement rate, r, while complementarity lowers the optimal r. Note that for the case considered in the main text q and s are assumed to be neither substitutes or complements ($s^*(q)=\bar{s}$) and eq. (A.6) remains the correct 'second best' optimum.

References

Carter, Grace, and Paul B. Ginsburg, 1985, The Medicare case mix index increase: Medical practice changes, aging and DRG creep, R-3292-HCFA (The Rand Corporation, Santa Monica, CA) June.

Davis, Karen, G. Anderson, S. Renn, D. Rowland, C. Schramm and E. Steinberg, 1985, Is cost containment working?, Health Affairs 4, no. 3, 81–94.

Detsky, A.S., 1978, The economic foundations of national health policy (Ballinger, Cambridge, MA).

Evans, R., 1974, Supplier-induced demand: Some empirical evidence and implications, in: M. Perlman, ed., The economics of health care and medical care (Wiley, New York).

Farley, Dean, 1985, Competition among hospitals: Market structure and its relation to utilization, costs and financial position, DHMS Publication no. (PHS) 85-3353 (National Center for Health Services Research, U.S. Department of Health and Human Services, Washington, DC).

Feldstein, Martin, 1971, Hospital cost inflation: A study of non-profit price dynamics, American Economic Review 61, no. 5, 853–872.

Feldstein, Paul, 1968, Applying economic concepts to hospital care, Hospital Administration 13, Winter, 68–89.

Frank, Richard G. and Judith Lave, 1984, The effect of benefit design on the length of stay of Medicaid psychiatric patients, Unpublished manuscript (University of Pittsburgh, PA).

Iglehart, John K., 1984, The Twin Cities' medical marketplace, New England Journal of Medicine 311, no. 3, 343–348.

Luft, Harold S., 1980, Assessing the evidence on HMO performance, Milbank Memorial Fund Quarterly 58, 501–536.

McGuire, Thomas G., 1985, Criteria for evaluating a case-mix classification system as a basis for prospective payment, Unpublished manuscript (Health Policy Center, Brandeis University, Waltham, MA).

Manning, Willard G., Arleen Leibowitz, George A. Goldberg, William H. Rogers and Joseph P. Newhouse, 1984, A controlled trial of the effect of a prepaid group practice on use of services, New England Journal of Medicine 310, no. 23, 1505–1530.

Newhouse, Joseph P., 1970, Towards a theory of non-profit institutions: An economic model of a hospital, American Economic Review 60, March, 64–74.

Pauly, Mark, 1980, Doctors and their workshops: Economic models of physician behavior (University of Chicago Press, Chicago, IL).

Pauly, Mark and Michael Redisch, 1973, The not-for-profit hospital as a physician's cooperative, American Economic Review 63, March, 87–99.

Shavell, Steve, 1979, Risk sharing and incentives in the principal-agent relationship, Bell Journal of Economics, Spring, 55–73.

Simborg, D.W., 1981, DRG creep: A new hospital-acquired disease, New England Journal of Medicine 304, no. 26, 1602–1604.

Vladeck, B.C., 1984, Medicare hospital payment by diagnosis-related groups, Annals of Internal Medicine 100, no. 4, 576–590.

Zeckhauser, Richard, 1970, Medical insurance: A case study of the trade-off between risk spreading and appropriate incentives, Journal of Economic Theory 2, 10–26.