Optimal payment systems for health services

R. Ellis

Journal of Health Economics

Journal of Health Economics 9 (1990) 375–396. North-Holland

OPTIMAL PAYMENT SYSTEMS FOR HEALTH SERVICES*

Randall P. ELLIS and Thomas G. McGUIRE

Department of Economics, Boston University, Boston, MA 02215, USA

Received January 1990, final version received July 1990

Demand-side cost sharing and the supply-side reimbursement system provide two separate instruments that can be used to influence the quantity of health services consumed. For risk-averse consumers, optimal payment systems – pairs of insurance and reimbursement plans – are characterized by conflict rather than consensus between patient and provider about the quantity of treatment. A model of conflict resolution based on bargaining theory is used to represent the outcome when the payment system creates divergences between desired demand and desired supply. Using that model, we describe the optimal combination of insurance and reimbursement systems that maximize consumer welfare.

1. Introduction

The central problem in U.S. health policy is designing a payment system that protects patients against financial risk without inducing inefficiently high levels of health service use. Any health payment system consists of two parts: the insurance system that structures patients' incentives to demand care, and the reimbursement system that structures providers' incentives to supply care. A well-established literature examines optimal insurance for consumers facing uncertain health costs in the presence of risk aversion and moral hazard [Arrow (1963), Zeckhauser (1970), Keeler et al. (1977), Besley, (1988)]. A much smaller but rapidly growing literature examines the characteristics of optimal reimbursement for providers in the presence of response to supply incentives [Dranove (1987), Ellis and McGuire (1986, 1988), Frank and Lave (1986, 1989)].

The purpose of this paper is to consider both insurance and reimbursement as elements of the optimal health payment system, with optimality being judged in relation to the social goals of minimizing consumer financial risk and of providing the efficient level of health care. An important point of

*Research for this paper was supported by grants from the Pew Charitable Trust and Grants Nos. R23-MH42335 and KO5MH00832-01 from the National Institute of Mental Health. We are grateful to Emmett Keeler, Michael Manove, Robert Rosenthal, the editor, several anonymous referees, and participants at seminars at Berkeley, Columbia, Harvard, MIT, and the ASSA meetings in Chicago for helpful comments; and to Mohammad Tareque for his careful work as a research assistant.

departure for the paper is the observation that not all possible payment systems (pairs of insurance and reimbursement plans) clear markets. More concretely, in an encounter between a patient and a health care provider, the level of services desired by the patient need not be the same as the level desired by the provider.

For example, quantity demanded regularly exceeds quantity supplied in health maintenance organizations (HMOs), where on the margin consumers often face no cost sharing,[1] and providers receive no additional payments for supplying care. Most members of HMOs have alternatives to the HMO, so that competition is some protection against undesirable provider restrictions on supply. Yet in settings such as Medicare and Medicaid where reimbursement incentives to restrict supply are being most aggressively pursued, patients may have few or no alternatives. For example, the lump-sum methods used by Medicare for reimbursing hospitals for the elderly, based upon the Diagnosis Related Group (DRG) in which the patient is classified upon discharge, provides strong incentives for providers to supply fewer services than well-insured Medicare beneficiaries may desire.

In section 2, we develop a model in which patients and providers may disagree over the level of health services to provide. We use a model of conflict resolution to reconcile differences between the quantity demanded and quantity supplied, where the solution concept is based on an adaptation of the Nash bargaining model due to Roth (1977) in which gains from bargaining to each party are expressed in relation to the 'point of minimal expectations' for each agent rather than a threat point. When the marginal benefit functions of each agent are linear, we show that Roth's model has a convenient analytical solution.

Section 3 uses this model to reconcile patient demand and provider supply when the patient is ill. Patient demand is modeled in a conventional way, with insurance affecting the quantity demanded. Although informational problems are often regarded as central to health care markets [Arrow (1963), Pauly (1988)] we follow the optimal insurance literature and disregard patient's difficulties in evaluating medical treatment. We also neglect the problem of adverse selection and consider a single health plan offered to a representative consumer. The supply of health care is more complicated because it involves the provider in the roles of self-interested supplier and agent for the patient. In our model, the degree of agency affects the supply curve for care, and the reimbursement system affects the desired quantity supplied. Once demand and supply are specified, we can map insurance and reimbursement systems into actual quantities provided.

The full model is presented in section 4, where the patient faces a risk of

[1]HMOs usually do charge a small, fixed visit fee for services, but within a single visit there are generally no additional charges. Hence, on the margin there is no cost sharing.

being either healthy or sick, and insurance reduces financial risk as well as influences quantity demanded. Conditions for social optimality are described. Section 5 solves the model for the optimal health payment system. The form of the optimal system, and whether it can achieve the first best, depends on the distribution of bargaining power between the provider and patient, the moral hazard in demand, risk aversion, and the strength of provider agency on behalf of the patient.

The central finding of the paper is that among all health payment systems, including those that do and those that do not lead to market clearing, consumer welfare is maximized by payment systems in which providers are not willing to supply all the care that the consumer demands when sick. Risk-averse consumers want insurance to reduce financial risk, but they do not want to pay for the extra quantity demanded because of moral hazard. Giving providers incentives to restrict supply via the reimbursement system can both protect consumers against financial risk and achieve (in some cases) efficient levels of health care consumption.

2. A model of conflict resolution

In health care, the quantity of service desired by consumers and the quantity desired by providers frequently depend upon complex payment and reimbursement schemes. Furthermore, prices do not automatically equate demand and supply to clear the market. Instead, an insurer can break the linkage between the demand price and the supply price, and use both prices separately to influence the quantity of services actually provided. This section develops a model in which demand and supply prices are set independently, and in which the buyer and seller can differ about the desired quantity traded.

The usual assumption made in situations with a divergence between demand and supply is that the short side of the market dominates. That is, if X_D^* and X_S^* are desired demand and desired supply, respectively, then the usual assumption is that $X=\min(X_D^*, X_S^*)$. For health services this assumption appears too simplistic. The importance of the 'physician–patient relationship' puts pressure on both sides to come to an agreement about a course of action. Because patients generally acknowledge that physicians have special expertise about health care, and because they may wish to avoid alienating their physician, patients may defer somewhat towards the desired supply of providers, X_S^*. Conversely, physicians and other providers have legal, ethical, and commercial reasons to care about patient satisfaction, and are therefore also under pressure to compromise the level of treatment provided towards X_D^*.

An alternative approach for representing the resolution of conflict between desired demand and desired supply is to use a simple function of X_D^* and X_S^*

such as $X=\delta X_D^*+(1-\delta)X_S^*$, with $0\leqq\delta\leqq 1$. The disadvantage of this 'reduced form' model is that it lacks a theoretical basis, and does not reflect marginal valuation of benefits and costs in its solution.

Instead of using either of these two approaches, we model the resolution of conflict using a bargaining model. Since prices are predetermined, consumers and suppliers must come to an agreement over quantities. We use a solution concept for the Nash (1950) bargaining problem developed by Roth (1977, 1979) to represent the outcome. Roth defines the 'point of minimal expectations' for an agent as the most desired outcome of the other agent. He shows that the Nash bargaining solutions using the point of minimal expectations as the disagreement point is the unique solution that satisfies four desirable properties: independence of positive affine transformations, symmetry, Pareto optimality, and independence of alternatives other than the point of minimal expectations.

One attractive feature of the Roth–Nash bargaining solution is that the outcome can be represented as equivalent to the solution of a maximization problem with a Cobb–Douglas welfare function. Relaxing the symmetry assumption, the exponent of the Cobb–Douglas welfare function reflects the bargaining power of the two agents [Roth (1979), Binmore (1980)]. In terms of our consumer–supplier bargaining model, each agent is assumed to bargain over the quantity to be provided (X) while taking as his point of minimal expectations the level of treatment desired by the other agent (X_S^* for the consumer, and X_D^* for the supplier). Using this approach, the outcome of bargaining can be represented as the solution to the problem:

$$\max_X \,[U(X)-U(X_S^*)]^{1-\gamma}[V(X)-V(X_D^*)]^{\gamma}, \tag{1}$$

where $U(X)$ and $V(X)$ are the utility functions of the consumer and supplier, respectively, $1-\gamma$ is the bargaining weight of the consumer, γ is the bargaining weight of the supplier, and $0\leqq\gamma\leqq 1$.[2]

In this paper, we focus on the analytically tractable case where both the consumer's and the supplier's utility functions are concave, quadratic functions that can be written as

$$U(X)=FX-\tfrac{1}{2}GX^2+K_1 \tag{2}$$

and

[2]Binmore, Rubinstein and Wolinsky (1986) demonstrate that the asymmetric Roth–Nash solution can be derived as the limiting case of Rubinstein's (1982) bargaining model that represents the bargaining process in extensive form. The asymmetry reflected in the bargaining weights are consistent with models in which either one agent is more impatient to reach agreement, or the agents have different probabilities of a breakdown in negotiations. This asymmetric model has proven useful in other recent applications [e.g., Svejnar (1986)].

$$V(X)=HX-\tfrac{1}{2}IX^2+K_2, \tag{3}$$

respectively, with F and H nonnegative, G and I strictly positive, and K_1 and K_2 arbitrary constants. It is straightforward to show that desired quantity demanded and the quantity supplied are

$$X_D^*=F/G \tag{4}$$

and

$$X_S^*=H/I, \tag{5}$$

respectively.

Since the Cobb–Douglas function (1) is strictly quasiconcave, and the two functions $V(X)$ and $U(X)$ are strictly concave, the bargaining problem represented by eq. (1) is guaranteed to have a unique solution. Moreover, it is readily verified that the derivatives of the objective function become infinite at the boundaries where $X=X_D^*$ and $X=X_S^*$, hence the solution is guaranteed to be an interior one. For this reason, solutions satisfying the first-order conditions will be unique interior maxima.

Substituting eqs. (2) and (3) into (1), using (4) and (5), taking first derivatives, setting equal to zero, and rearranging yields the following expression that characterizes the outcome of the bargaining process:

$$\frac{(\gamma-1)[F-GX]}{\gamma[H-IX]}=\frac{[F-\frac{1}{2}G(X+X_S^*)](X-X_S^*)}{[H-\frac{1}{2}I(X+X_D^*)](X-X_D^*)}. \tag{6}$$

This expression can be simplified to a cubic function in X that has one feasible real root for the general case with $0<\gamma<1$. The general cubic solution to this equation (shown in a mathematical appendix available from the authors upon request) is rather cumbersome. Fortunately, for the case where the consumer and supplier have equal bargaining weights ($\gamma=\frac{1}{2}$), it is readily verified by substitution that the following simpler and much more intuitive expression is a solution satisfying the above condition:

$$X=\tfrac{1}{2}[F/G+H/I]=\tfrac{1}{2}(X_D^*+X_S^*). \tag{7}$$

This gives us the interesting result that if both agents have equal bargaining power (i.e., $\gamma=\frac{1}{2}$), the Roth–Nash solution concept implies that the actual quantity agreed upon will be the simple average of the quantities desired by both the consumer and the supplier. We have shown the following result.

Proposition 1. If the utility functions of both the consumer and the supplier are concave and quadratic in quantities, then the Roth–Nash solution to the bargaining problem is a cubic with one feasible root, characterized by the solution to (6). For the special case where the consumer and supplier have equal bargaining power, the solution is $X=\frac{1}{2}(X_D^*+X_S^*)$, *i.e., the solution will lie halfway between desired demand and desired supply.* □

For much of the remainder of this paper we focus on the solution for symmetric bargaining power. We do this for reasons of tractability, not because we actually believe that consumers and providers have equal bargaining power. (Stronger bargaining power by providers seems more likely.) Furthermore, in a more general model with competition and asymmetric information, bargaining power could be expected to depend on these factors. Here we take bargaining power as exogenous. After presenting analytic results for the case of symmetric bargaining power we use simulation techniques to derive solutions for asymmetric bargaining power.

3. Moral hazard and agency

In this section, we apply the solution concept above to reconcile differences between the quantity of health care demanded and supplied conditional on the consumer being sick. For the reasons stated above, the patient and the provider are bound to come to an agreement. Moral hazard determines the shape of the patient's demand curve, and the insurance coverage determines the desired point on the demand curve. Agency (along with moral hazard) shapes the provider's supply curve, and the reimbursement system determines the desired point of supply. Although the patient has insurance, in this section, the patient is assumed to be risk neutral.

3.1. The patient's utility function

For simplicity, suppose that when healthy the consumer's utility depends upon only a single non-health good, N. When ill, the patient's utility depends upon two goods: N, and the consumption of health services, X. Suppose also that N yields a constant marginal utility, normalized to one. Let Y be income and P be the premium that must be paid by the patient even when ill. For convenience, we measure X in dollars, normalized so that the cost per unit is one. Define c to be the level of patient cost sharing, which is subject to the important boundary constraints

$$0 \leqq c \leqq 1. \tag{8}$$

Using this notation, the patient's utility when healthy, U_{H}, is simply N, which will be equal to $Y-P$, and the patient's utility when ill, U_{I}, is

$$U_{\mathrm{I}}=N+B(X)-K-Y-P-cX+B(X)-K, \tag{9}$$

where $B(X)$ is the total benefits from treatment, and K is an arbitrary constant. The constant K can be thought of as the fixed cost of being sick, the reference point from which benefits to treatment are measured. In order to apply our results from the last section, we assume that $B(X)$ is concave and quadratic:

$$B(X)=aX-\tfrac{1}{2}bX^2 \tag{10}$$

with $a>0$, $b>0$. The premium is fixed, and so the patient chooses X when sick while ignoring the effect on the premium. By substituting eq. (10) into (9) it is straightforward to verify that for a given c the quantity of health services which maximizes consumer utility is

$$X_{\mathrm{D}}^{*}=\begin{cases}(a-c)/b, & c<a, \qquad (11)\\ 0, & \text{otherwise.} \qquad (12)\end{cases}$$

Since cases where $c>a$ correspond to corner solutions where the consumer does not demand any services when sick, we shall rule out these cases by assuming $a>1$. This assumption will need to be further strengthened when risk aversion is introduced.

3.2. The provider's objective function

The provider is assumed to be concerned about profits, and about health benefits to the patient.[3] Concern about the latter occurs because the provider serves as an agent for the patient. Both profits and patient benefits depend on the quantity of services provided, X. We represent the provider's objective function in simple fashion as

$$V(X)=\alpha B(X)+\Pi(X). \tag{13}$$

The provider's preferences are characterized by a constant marginal rate of

[3] An alternative formulation of agency would have the provider care about patient utility rather than health benefits. We believe our formulation better captures the agency role of the health provider. Providers only observe and feel responsible for patient's health, not their overall welfare. Also, provider reputations and malpractice concerns are affected by treatment benefits. Our formulation does create an asymmetry between benefits and costs of health care to the patient, however. The implications of this asymmetry are reexamined in a subsequent footnote.

substitution, α, between profits, $\Pi(X)$, and benefits from treatment, $B(X)$, for an individual patient.[4] Agency is captured by the parameter α. We call the provider a 'perfect' agent when $\alpha = 1$ and the provider values patient health benefits equivalently with provider profits. Imperfect agency is introduced by allowing the provider to undervalue benefits to the patient relative to profits, which corresponds to $\alpha < 1$. Since one can construct arguments for why one might find $\alpha > 1$, a form of 'super agency', we also examine cases with $\alpha \geqq 1$.

The provider's net revenue or profit depends upon the quantity of services the patient receives as well as the reimbursement system. For this paper we characterize reimbursement as consisting of two parts: a lump-sum payment, R, that is independent of the level of services provided, and a cost-based component that we write as $(1-s)X$.

The reimbursement system for the provider, summarized by R and s, is analogous to the insurance system for the patient, defined by P and c. The parameter R corresponds to an insurance premium paid to the provider to accept part of the financial risk of supplying health care to the patient. The parameter s is the degree of supply-side cost sharing. The provider regards R and s as given when deciding desired supply to a patient, just as the patient regards the insurance premium and the coinsurance rate as given when choosing his demand for care. We assume that the lump-sum payment R is set so as to equate total service costs across all patients with total revenues, that is, $R \equiv E[X] - E[(1-s)X] = sE[X]$.

Since the notion of supply-side cost sharing is less commonly used than the demand-side analog (the coinsurance rate) it is worth identifying s under two common reimbursement systems. A fully prospective payment system is characterized by $s = 1$, with full payment to providers in the form of the prospective payment R. A cost-based reimbursement system is characterized by $R = 0$, $s = 0$. Intermediate between these two extremes are reimbursement systems that have some degree of prospective payment $(R > 0)$ and some provider responsibility for costs at the margin $(0 < s < 1)$. As in Ellis and McGuire (1986) we call these 'mixed systems' of reimbursement.

Using this notation, provider profits from a single patient can be written as

$$\Pi(X) = R - sX. \tag{14}$$

Substituting in expressions for patient benefit and net revenue, the provider's objective function becomes

[4]Since total provider profits are determined over many patients, the marginal utility of profits to the provider is likely to be relatively constant when treating a single patient and hence the assumption of constant marginal rate of substitution between profits and patient benefits is not as unreasonable as it might at first seem.

$$V(X)=\alpha[aX-\tfrac{1}{2}bX^2]+R-sX. \tag{15}$$

The desired supply of services by providers is

$$X_S^*=\begin{cases}(a-s/\alpha)/b, & s<\alpha a, \quad (16)\\ 0, & \text{otherwise.} \quad (17)\end{cases}$$

As can be seen from eq. (16), the lower is α, that is, the less providers value patient benefits in relation to their profits, the more responsive are providers to incentives in the reimbursement system represented by the supply-side cost sharing s. The analytical solutions discussed in this paper require that eq. (16) applies so that desired supply is positive. We also impose the important boundary constraints on s:

$$0\leqq s\leqq 1. \tag{18}$$

3.3. The bargaining solution and efficient health care use

Since both the patient's utility and the provider's objective functions are quadratic in X and satisfy the conditions necessary for Proposition 1, the Roth–Nash solution to the bargaining problem for the symmetric bargaining power case ($\gamma=\frac{1}{2}$) will be:[5]

$$X(c,s)=a/b-(c+s/\alpha)/(2b). \tag{19}$$

The solution described by (19) can be interpreted in light of the efficient level of X, where marginal social benefits equal marginal social costs. In our model this corresponds to $B'(X)=1$. By stating the condition for efficiency in this way, we regard only the patient's benefits from health treatment. Although the provider is an agent for the patient and the patient's health benefits enter into the provider's objective function, provider welfare is assumed to be indicated solely by net revenue. Since, as stated above, the prospective component of reimbursement, R, is set so as just to cover costs for the average patient, expected net revenue of the provider is zero, and social welfare is identical to patient utility. Because patients must bear the full cost of health care through P or c, their utility will be maximized where their marginal benefit equals marginal cost.

It can easily be seen from eq. (10) that $B'(X)=1$ corresponds to $X=(a-1)/b$. From (19), the social optimum is achieved by any payment system with $(c+s/\alpha)/2=1$. Given the constraints that $c\leqq 1$, $s\leqq 1$, achieving the first

[5]Here, $F=a-c$, $G=b$, $H=\alpha a-s$, and $I=\alpha b$.

best is infeasible whenever $\alpha > 1$, and the second best solution is $c = s = 1$. For $\alpha < 1$, multiple payment systems will achieve the social optimum. All that is required is that there be enough demand and supply-side cost sharing in combination so that $c + s/\alpha = 2$. Since desired supply falls as α decreases, supply-side cost sharing, s, must be decreased in order to offset imperfect agency.

A number of important observations can be made about the results so far that carry through to the more complete model that includes patient risk aversion. First, the bargaining model points out that there can be a built-in problem of excess health costs *even without insurance* because of the agency role of the provider. Even if patients have no insurance and pay full social cost for health care ($c = 1$), unless there is sufficient supply-side cost sharing, desired supply may exceed desired demand and result in overconsumption. Second, the model suggests an alternative interpretation for the 'supplier-induced demand' in health care markets noted by Evans (1974), Fuchs (1978), and others. In our model, providers influence utilization directly without having to alter patient's demand: the quantity of services provided may lie off of the patient's demand curve.

A third observation is that even in this simple model without risk aversion where there are two policy instruments (c, s) and one social goal (efficient health care use), the first best may not be attainable because of boundary constraints on the range of demand and supply-side cost sharing (i.e., $0 \leqq c \leqq 1$ and $0 \leqq s \leqq 1$). These boundary constraints will remain prominent in the next section.

Finally, these results preview the constructive role to be played by strong provider bargaining power and imperfect agency. It can be foreseen that in our model the first-best allocation when the patient is risk averse will be characterized by $c = 0$. Is supply-side cost sharing strong enough to fix the level of X at the efficient level in this model given $c = 0$? For the symmetric bargaining power case, only if, from the condition for efficiency, s can be set at 2α. This is feasible only when providers are relatively weak agents for patients. Alternatively, the potency of supply-side cost sharing can be enhanced by stronger provider bargaining power.

4. Patient risk aversion

4.1. Expected utility framework

We assume there are two states of the world that differ according to the state of health of the consumer and the marginal utility of income. The consumer is healthy with probability $(1 - \delta)$ and ill with probability δ, with $0 < \delta < 1$. When healthy, the marginal utility of income is assumed to be η, versus λ when ill. When $\eta < \lambda$ the individual is risk averse and hence willing to pay for insurance at actuarially fair prices. When $\lambda<\eta$, the consumer can be considered a risk lover, and will prefer not to have any insurance.[6] Since we are allowed one normalization, we choose the convenient normalization that the expected marginal utility of income is one, which means

$$(1-\delta)\eta+\delta\lambda\equiv 1. \tag{20}$$

Clearly, with this notation, $\lambda>\eta$ is equivalent to $\lambda>1$, and we will use this representation of risk aversion throughout the remainder of the paper.

The individual benefits from health services only when ill, but must pay a premium P in both states of the world. In terms of the notation introduced above, net income available to spend on non-health goods when healthy is $Y-P$ and when ill, $Y-P-cX$. Recalling that the quadratic function $B(X)$ represents benefits from treatment, the patient's expected utility function can be expressed as

$$EU=(1-\delta)\eta(Y-P)+\delta[\lambda(Y-P-cX)+B(X)-K] \tag{21}$$

$$=Y-P+\delta[B(X)-\lambda cX-K], \tag{22}$$

where for the second line we have used normalization equation (20). The actuarially fair insurance premium is

$$P=\delta(1-c)X. \tag{23}$$

Using this framework, the consumer's desired demand when ill is found by maximizing the term in square brackets in eq. (21), the utility when ill, while holding P constant. This yields

$$X_{\mathrm{D}}^{*}=\begin{cases}(a-c\lambda)/b, & c<a/\lambda, \qquad (24)\\ 0, & \text{otherwise}. \qquad (25)\end{cases}$$

[6]Since our representation of risk aversion is somewhat unconventional, it deserves a brief comment. The most common way of incorporating risk aversion into models of consumer behavior is to define a utility of income function that is concave in income, such that the marginal utility of income declines as income available for purchasing non-health services increases. With this conventional representation, the marginal utility of income changes only because of changes in out-of-pocket costs resulting from health care expenditures, and a fully insured consumer has a constant marginal utility of income.

We use an alternative mechanism, which is to assume that the marginal utility of income (λ) depends upon the state of the world. A convenient implication of this formulation is that the patient's demand curve for X [which depends on $B(X)$ and λ] can be specified independent of the degree of insurance coverage. Although we believe that the most plausible assumption is that consumers value money more when they are ill ($\lambda>1$), even if they are fully insured, we also briefly examine cases where $\lambda<1$ and $\lambda=1$ since this provides further insights.

This expression is similar to eq. (11), except for the appearance of λ: for a consumer that is not fully insured ($c>0$), as the marginal utility of income when ill increases, fewer health care services are demanded.

4.2. The bargaining solution and efficiency conditions

The consumer utility function and desired demand are as described above, while provider utility and desired supply remain as defined above in eqs. (15) and (16). When desired demand and supply are positive ($c<a/\lambda$ and $s<\alpha a$), the conditions of Proposition 1 are satisfied[7] and the Roth–Nash bargaining solution yields:

$$X(c,s)=a/b-(\lambda c+s/\alpha)/(2b). \tag{26}$$

The efficiency conditions change only slightly now that we have introduced risk aversion. Analogous to the line of argument above, maximizing the expected utility of consumers is equivalent to maximizing social welfare since the reimbursement system sets provider profits to zero. Consumer utility is affected by insurance coverage (c) and consumption of health care (X). The social optimum can be found by maximizing eq. (22) over c and X, subject to (23) and (8). It is readily shown that for the case of risk aversion, $\lambda>1$, the social optimum satisfies the two conditions:

$$c^*_{\mathrm{SO}}=0 \tag{27}$$

and

$$B'(X)=a-bX=1. \tag{28}$$

Hence for a risk-averse consumer, the socially optimal level of X, X^*_{SO}, is

$$X^*_{\mathrm{SO}}=(a-1)/b. \tag{29}$$

For the case where $\lambda<1$ the social optimum is characterized by

$$c^*_{\mathrm{SO}}=1, \tag{30}$$

$$X^*_{\mathrm{SO}}=(a-\lambda)/b. \tag{31}$$

As should be expected, for a risk-loving consumer, the social optimum is achieved with no insurance. Since health care is paid for from dollars when ill, by consuming health care up until the marginal benefit is equal to the

[7] Here, $F=a-\lambda c$, $G=b$, $H=\alpha a-s$, and $I=\alpha b$.

marginal utility of income when ill. For the case where $\lambda = 1$, so that the consumer is risk neutral, the consumer is indifferent about the level of insurance.

5. Optimal payment systems

This section considers the problem of attaining the social optimum just described with two payment system parameters, c and s, when health care use is determined by the bargaining solution. Mathematically, this requires maximizing the expected utility of the consumer with the additional constraint that quantity of health care use is determined by the bargaining solution. The policy maker is assumed to understand how patient and provider interests affect health use, and selects demand and supply-side payment policies accordingly. Initially, we continue with the case of symmetric bargaining power, and examine the characteristics of payment systems which achieve the first best social optimum when that is possible. While maintaining the symmetry assumption, we next examine the role of moral hazard and the degree of agency. Finally, we turn to cases of unequal bargaining power, where first patients and then providers solely determine the levels of treatment. We also discuss what is probably the most realistic case, in which providers have stronger bargaining power than patients when determining levels of treatment.

5.1. Symmetric bargaining power

When patients and providers have equal bargaining power ($\gamma = \frac{1}{2}$), the level of services will be determined by eq. (26). The problem of maximizing social welfare can therefore be written as:

$$\max_{c,s} E[U] = Y - P + \delta[B(X) - \lambda cX - K], \tag{32}$$

subject to eqs. (23), (26), (8), and (18). The complete solution to this optimization problem is derived in an appendix, available from the authors upon request, where it is shown that nine solution regimes are possible for various parameter values, corresponding to different combinations of α, λ, δ, a, and b. These regimes correspond to whether c and s are on or off their upper and lower boundaries, or in between. Fortunately, the intuition behind the main findings can be seen without showing the complete derivation of results. We will focus here on the intuition of the results rather than a formal derivation.

Under what conditions can the two payment system parameters (c and s) be chosen so that the outcome of bargaining achieves the social optimum, and what will this payment system look like? It was shown that when the consumer is risk averse ($\lambda > 1$), the social optimal is characterized by eqs. (27) and (29), while the solution to the bargaining problem satisfies eq. (26). Hence, the bargaining solution will achieve the social optimum only if $c^* = 0$ and $\frac{1}{2}(\lambda c^* + s^*/\alpha) = 1$. This in turn implies $c^* = 0$ and $s^* = 2\alpha$ are necessary conditions for the social optimum to be achieved. Since s is subject to the boundary constraint $s \leqq 1$, this payment system is feasible only if $2\alpha \leqq 1$. Thus we have now shown:

Proposition 2. If the consumer and provider have equal bargaining power ($\gamma = \frac{1}{2}$), *the consumer is risk averse* ($\lambda > 1$), *and the provider is an agent for the patient when ill* ($\alpha > 0$), *then the optimal payment system achieves the first-best social optimum characterized by eqs.* (27) *and* (29) *only if*

$$2\alpha \leqq 1. \tag{33}$$

The optimal payment system is characterized by full insurance for the consumer, $c^* = 0$, *and a 'mixed' reimbursement system for providers, with* $s^* = 2\alpha$. □

Condition (33) can be interpreted as requiring that the provider is a sufficiently imperfect agent that the level of under-provision desired by the provider is sufficient to counteract the overconsumption of the consumer when fully insured. Note that when the first-best social optimum can be achieved, the optimal payment system is independent of the degree of moral hazard, which in our model is fully captured by the constant term (a) in the demand curve.[8]

What are the characteristics of optimal payment systems when condition (33) is not satisfied? If the consumer is risk averse ($\lambda > 1$) and agency is strong, so that (33) is not satisfied, then the first-best cannot be achieved, and too many services are used. It turns out that whenever the first best is unattainable, optimal payment systems are always characterized by fully prospective payment, i.e., $s^* = 1$. This occurs because the supply-side incentives to reduce care by using prospective payment are not strong enough to offset the overconsumption caused by full insurance.

The optimal degree of cost sharing when the first best is not attainable

[8]This result is not sensitive to the assumption that physicians care only about patient benefits from treatment rather than total patient utility: when patient cost sharing is zero, as it is at the first best, a model in which the provider cares about patient utility instead of patient benefits from treatment yields the identical result.

depends upon α, λ, δ, and a. Payment systems with fully prospective payment and no insurance, partial insurance, or full insurance are all possible. Determination of the optimal insurance given the reimbursement system is fully prospective reflects the trade-off between risk protection and incentives to use the efficient level of health care.

An important point to make before proceeding is that the most prevalent payment system used in the United States, cost-based reimbursement with partial insurance ($s=0$, $0<c<1$), is *never* an optimal payment systems. Starting from such a payment system, consumer welfare can always be improved by imposing some supply-side cost sharing. It is interesting to note that recent trends in the U.S. toward prospective payment have been exactly in this direction.

5.2. Moral hazard and agency

Moral hazard has played a central role in the literature on the economics of insurance. Second-best pricing implies that the insurance coverage should decrease as moral hazard (or the demand elasticity) increases [Zeckhauser (1970)]. An important implication of Proposition 2 is that under certain conditions supply incentives are sufficiently strong that moral hazard is irrelevant to the optimal health payment system. When supply incentives are weak, and cannot fully achieve the efficient level of health services, X^*, in the presence of full insurance, however, moral hazard returns to a prominent position in this second-best world.

In our model moral hazard is captured entirely by the constant term of the demand curve, a.[9] As $a \rightarrow 1$, the moral hazard problem becomes more severe, since the demand curve becomes more elastic at a given price. Here we examine the relationship between different combinations of a and α while maintaining the assumption of symmetric bargaining power. Although the results are shown analytically in an appendix (available upon request), since the key relationships are easiest to see graphically, we illustrate the results here using specific numerical values for several key parameters.

Fig. 1 illustrates the payment systems appropriate for various combinations of a and α while assuming symmetric bargaining power ($\gamma=\frac{1}{2}$), and mild risk aversion ($\lambda=1.2$). To fully specify the model, we also assume $\delta=0.1$. Once these three parameters are specified, four different payment

[9]As mentioned in section 4, we assume for the remainder of this paper that $a>\lambda$ and $a>2\alpha$ so that desired demand and desired supply are both positive and the bargaining solution (26) is appropriate. This imposes constraints on the combinations of parameters for which our bargaining model is appropriate, but the restriction does not rule out parameter combinations that are economically interesting.

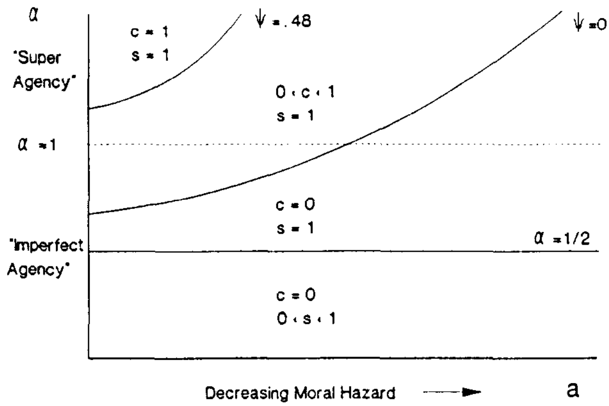

Fig. 1. Characteristics of optimal payment systems when there is symmetric bargaining power ($\gamma=0.5$) and mild risk aversion ($\lambda=1.2$). Shown here are the coinsurance rate (c) and provider cost sharing (s), for different values of moral hazard (a) and agency (α).

regimes are possible. The three boundaries between these regions are shown as solid lines.[10]

Note first that when α is less than $\frac{1}{2}$, the optimal payment system is one with full insurance ($c=0$) and a mixed system ($0<s<1$), as previously discussed. As α increases, the optimal provider reimbursement system switches over to one of fully prospective payment at the threshold where $2\alpha=1$. Although initially full insurance can be maintained, as α increases, eventually supply-side incentives are not strong enough to prevent overconsumption and the optimal reimbursement system becomes one with partial insurance. If the provider is a 'super agent', ($\alpha>1$) trying to overprovide services to the consumer, and the moral hazard problem is severe (a close to one) then a system with fully prospective payment and no insurance can be optimal.

Holding α, the degree of agency, constant at 1, while increasing a, increases the level of insurance coverage that is optimal. Hence our model, as illustrated in fig. 1, reinforces the conventional finding that as the moral hazard problem declines, it is optimal to provide more and more complete insurance coverage to consumers. Our model suggests two reinforcements to the finding. The first and most important refinement is that if the provider is a highly imperfect agent, then the optimal payment system may be independent of the degree of moral hazard.

[10]The boundary condition separating the regime with partial insurance from the regime with full insurance is upward sloping, and is defined by $\psi=0$, where $\psi=2.4-0.8/\alpha-0.8a$. The boundary condition separating the regime with partial insurance from the regime with no insurance is $\psi=\lambda(4-3\lambda)=0.48$. The general expression for ψ is derived in an appendix available from the authors.

The second refinement is that if the provider is a 'super agent' who tends to over-provide services, then a payment system with no insurance may be optimal even for moderate levels of risk aversion and moral hazard. In the conventional optimal insurance model, it is optimal to at least partially insure a risk-averse consumer. In our model that need not be so if providers systematically overprovide services.

5.3. Unequal bargaining power

Our analysis of optimal payment systems has thus far only considered cases of symmetric bargaining power ($\gamma=0.5$). This has been done for convenience since the symmetric bargaining case has a single analytical solution. To explore asymmetric bargaining power, this section first considers the two extreme cases where the consumer alone determines the level of treatment ($\gamma=0$) and then the opposite case where the provider alone determines the level of treatment ($\gamma=1$). We then use simulation results to examine what we believe to be the most realistic case, one in which the provider has stronger bargaining power than the patient ($\frac{1}{2}<\gamma<1$).

5.3.1. Consumer sovereignty ($\gamma=0$)

Optimal payment systems when the consumer alone determines the level of treatment will be invariant to the level of provider cost sharing. Hence the reimbursement parameter s may be ignored. The social problem is to choose c so as to maximize eq. (22) subject to (23), (8), and relation (24) instead of (26). It can readily be shown that the optimal choice of c depends on

$$c^* = \frac{\lambda - a(\lambda - 1)}{[2-\lambda]\lambda}. \tag{34}$$

The optimum is $c=c^*$, unless $c^*<0$, in which $c=0$ is optimal. It is readily verified that for any $a>1$, $\lambda>1$, then as $\lambda\to 1$, $c^*\to 1$, i.e., that no insurance is optimal if there is no risk aversion. Also, it can be seen that as λ increases, c^* decreases until at $\lambda=a/(a-1)$, $c^*=0$, and full insurance is optimal. Only in cases where $\lambda\leqq 1$, corresponding to risk neutrality or risk-loving behavior, does the payment system achieve a first-best solution satisfying (30) and (31), with $c=1$. This confirms the usual optimal insurance result that with moral hazard and risk aversion, partial insurance achieves only a second best.

5.3.2. Provider sovereignty ($\gamma=1$)

Optimal payment systems when the provider alone determines the level of treatment is easy to derive. Since cost sharing has no effect on the level of treatment, c can be set so as to achieve the optimal level of insurance (zero

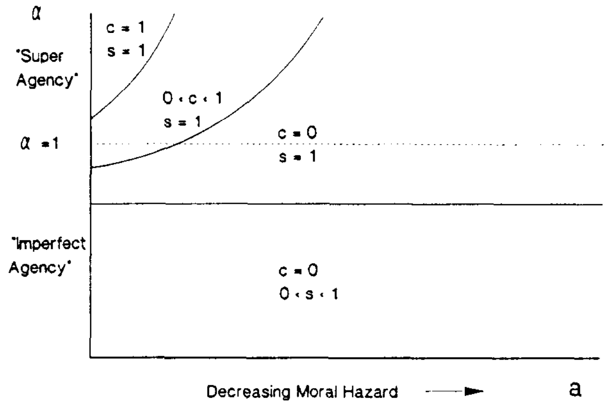

Fig. 2. Characteristics of optimal payment systems when there is strong provider bargaining power ($\gamma=0.8$) and mild risk aversion ($\lambda=1.2$). Shown here are the coinsurance rate (c) and provider cost sharing (s), for different values of moral hazard (a) and agency (α).

or one, depending upon whether the consumer is risk averse or risk neutral). The reimbursement parameter s should be set so as to achieve the optimal level of health services. Comparing the providers desired supply shown in eq. (16) with the socially optimal level of consumption shown in eq. (29), it can be seen that when consumers are risk neutral, the payment system will achieve the social optimum only if $s=\alpha$, and $\alpha\leqq 1$. This is the same result as in Ellis and McGuire (1986).

What is the optimal payment system when $\alpha\geqq 1$? For risk-averse consumers ($\lambda>1$), the optimal system is $s^*=1$, $c^*=0$. Here one would like to raise s further to reduce demand, but the constraint $s\leqq 1$ is binding. Raising c has no effect, since the quantity demanded by the consumer is immaterial when $\gamma=1$.[11]

5.3.3. *Strong provider bargaining power* ($\gamma>\frac{1}{2}$)

Optimal payment systems when providers are stronger bargainers than consumers but do not solely determine the level of treatment is perhaps the most realistic case. Unfortunately, analytic solutions to the unequal bargaining weight case appear intractable. Therefore, we examined the nature of different payment regimes using a simulation model that solved for the bargaining outcomes and searched for the optimal payment system numerically. The outcome of one set of simulations when $\gamma=0.8$ are shown in fig. 2.

[11]For risk lovers, optimal payment systems of course involve no insurance. It is easily shown that for $\lambda<1$, $\alpha\lambda<1$ the optimal payment system is $0<s^*=\alpha\lambda<1$, $c^*=1$. Finally, the optimal payment system for cases where $\lambda<1$, $\alpha\lambda\geqq 1$, is $s^*=1$, $c^*=1$.

The boundaries between the different payment regimes, shown as solid lines in fig. 2, correspond to those in fig. 1. In fig. 2, these boundaries were derived numerically rather then analytically, since they cannot be expressed as simple analytical expressions.

In contrast with fig. 1, a larger set of combinations of a and α result in the first-best social optimum being achieved with $c=0$, and $0<s<1$. As with symmetric bargaining, the level of moral hazard is irrelevant if α is sufficiently low, since supply-side incentives are sufficiently strong as to offset the consumer's overconsumption when fully insured. Regions in which only second best is achieved, and partial or full insurance are optimal have also diminished in size.

The results of the simulations with asymmetric bargaining power confirm what would be expected intuitively. In effect, increased provider bargaining power increases the range over which the mixed reimbursement system is optimal, and reduces the set of parameter combinations in which partial insurance is optimal. Strong provider bargaining also makes achieving the first best easier.

6. Conclusion

The problem of designing an optimal health care payment system must be recast when it is recognized that payment instruments on both the demand and the supply side can be used to attain the social goals of efficient utilization and minimization of patient financial risk. Neither the literature on optimal insurance nor the literature on optimal provider reimbursement takes account of the other side of the policy coin. Empirical evidence from the health sector leaves no doubt that both demand and supply-side payment practices influence utilization.

In this paper we have proposed a model of treatment determination which is consistent with the evidence that both demand- and supply-side incentives matter and which is capable of addressing the questions of the optimal combination of payment strategies. Utilization is modeled as the outcome of bargaining between patients and providers, with the solution being influenced by the bargaining strength of the two parties, the agency of the provider, and the demand- and supply-side payment incentives. In the explicit model we develop, it was necessary to make a number of simplifying assumptions which are useful to keep in mind when reviewing our conclusions: patients and physicians have full information about treatment benefits, patients cannot choose physicians, bargaining power is exogenous, physicians are risk neutral, payment systems are costlessly administered, and financial instruments are the only controls on behavior. The specific features of our results are not robust to changes in these assumptions. Two central findings do stand up well, however, and these are what we emphasize here.

First, conflict rather than consensus is part of socially optimal payment systems. Payment systems that achieve the desired balance between protecting consumers from financial risk and controlling costs are characterized by generous insurance coverage and financial incentives on providers to control costs. This results in conflict, because consumers would like to receive more services than providers want to give them. For our specific assumptions, the first best (if feasible) is achieved by full insurance for consumers and what we refer to as a mixed reimbursement system for providers, with some part of payment prospective and some part of payment cost based.[12] This finding helps to explain why many consumers choose to enroll in health payment plans, such as HMOs, where they do not always have their way about treatment.

Second, supply-side policies are the preferred instruments for cost control. If cost is a problem, the first question that should be asked is, 'what can be done with the provider reimbursement system?' rather than the traditional, 'how can insurance benefits be reduced?' A specific application of this idea is that, as we showed, cost-based reimbursement is *never* part of an optimal health care payment system. Our paper provides theoretical support for the general movement away from cost-based reimbursement that has recently occurred in private and public health plans. As some prospectiveness in the supply-side payment system is introduced, the use of insurance design for purposes of cost control can be relaxed, and coverage for patients can be improved. Currently observed coinsurance rates in the range 20–30% may be too high once adequate incentives to control costs are introduced on the supply side.

Our analysis makes some other points, but these conclusions are more tied to our specific assumptions. Although achieving the socially optimal level of treatment and degree of consumer insurance is easier in our model than in a world where only insurance or only supply-side payment strategies are available, it is striking that in many plausible scenarios achieving the first best is still not possible. Additional policy instruments, including non-financial controls on utilization, and competition may brighten the prospects for achieving efficient outcomes.

Achieving the first best is easiest if providers are strong bargainers but relatively imperfect agents. A counter-intuitive finding in the paper is therefore that imperfect agency and strong provider bargaining power can be good things. In the hands of a beneficent policy maker, powerful supply-side

[12]Under different assumptions, full insurance may not be part of an optimal payment system. For example, the transaction costs of providing insurance may imply that full insurance is optimal only after some deductible is exceeded. Or, if the decision to seek treatment is modeled separately from the decision about the intensity of services during treatment, demand side cost sharing may reemerge as part of a desired payment system. The fact that most HMOs charge a small fixed fee for each visit may be an appropriate measure for them to take in order to reduce the moral hazard problem of unnecessary visits.

policies are what make full insurance possible for consumers. With weak providers or those who value patient health benefits highly, even a fully prospective payment may not provide adequate incentives to moderate the level of treatment in our model, and partial or no insurance may be necessary to prevent overconsumption. Stronger bargaining weight for providers or less agency is desirable in that it makes it easier to set low consumer cost sharing and still achieve the first best. A more general model would recognize that powerful and dominant suppliers represent a double-edged sword. With asymmetry of information, such as when the provider knows patient health benefits with error, the exclusive reliance on the supply-side to set quantity poses greater risk. It would be relatively easy to construct a model in which a balance of supply and demand side incentives would come out of a need to minimize the cost of mistakes, or perhaps even to moderate the degree of conflict, which itself may entail costs.

This paper stresses the importance of provider and patient bargaining power, provider agency, and the response of desired supply to reimbursement incentives. Before now, these elements have been treated separately in the distinct demand and supply-side payment literatures. Our intregration of these literatures is done with relatively simple approaches to bargaining power, agency, and supply response. Further research should consider regarding these factors as themselves subject to the play of more fundamental forces associated with competition, information, and preferences.

References

Arrow, K., 1963, Uncertainty and the welfare economics of medical care, American Economic Review 53, 941–973.

Besley, T.J., 1988, Optimal reimbursement health insurance and the theory of Ramsey taxation, Journal of Health Economics 7, 321–366.

Binmore, K.G., 1980, Nash bargaining theory, Part II, ICERD, London School of Economics, D.P. 80/14.

Binmore, K.G., A. Rubinstein and A. Wolinsky, 1986, The Nash bargaining solution in economic modeling, Rand Journal of Economics 17, 176–188.

Dranove, D., 1987, Rate setting by diagnosis related groups and hospital specialization, Rand Journal of Economics 18, 417–427.

Ellis, R.P. and T.G. McGuire, 1986, Provider behavior under prospective reimbursement: Cost sharing and supply, Journal of Health Economics 5, 129–152.

Ellis, R.P. and T.G. McGuire, 1988, Insurance principles and the design of prospective payment systems, Journal of Health Economics 7, 215–237.

Evans, R., 1974, Supplier-induced demand: Some empirical evidence and implications, in: Mark Perlman, ed., The economics of health and medical care (Macmillan, London).

Frank, R.G. and J.R. Lave, 1986, Length of stay of medicaid psychiatric patient, Journal of Human Resources 21, 321–338.

Frank, R.G. and J.R. Lave, 1989, A comparison of hospital responses to reimbursement policies for medicaid psychiatric patients, Rand Journal of Economics 20, 588–600.

Fuchs, V., 1978, The supply of surgeons and the demand for operations, Journal of Human Resources 13 (Suppl.), 35–56.

Keeler, E., D. Morrow and J.P. Newhouse, 1977, The demand for supplementary health insurance, or do deductibles matter?, Journal of Political Economy 85, 789–802.

Nash, J.F. Jr., 1950, The bargaining problem, Econometrica 18, 155–162.
Pauly, M., 1988, Is medical care different? Old questions, new answers, Journal of Health Politics, Policy and Law 13, 227–238.
Roth, A.E., 1977, Independence of irrelevant alternatives, and solutions to Nash's bargaining problem, Journal of Economic Theory 16, 247–251.
Roth, A.E., 1979, Axiomatic models of bargaining (Springer, Berlin).
Rubinstein, A., 1982, Perfect equilibrium in a bargaining model, Econometrica 50, 97–110.
Svejnar, J., 1986, Bargaining power, fear of disagreement, and wage settlements: Theory and evidence from U.S. practice, Econometrica 54, 1055–1078.
Zeckhauser, R., 1970, Medical insurance: A case study of the trade-off between risk spreading and appropriate incentives, Journal of Economic Theory 2, 10–26.