HARVARD UNIVERSITY
JOHN F. KENNEDY SCHOOL OF GOVERNMENT

Faculty Research Working Papers Series

Market Structure, Commitment, and Treatment Incentives in Health Care

Nolan H. Miller

February 2004
RWP04-007

Market Structure, Commitment, and Treatment Incentives in Health Care.

Nolan H. Miller[*]

February 13, 2004

Abstract

People are more distrustful of managed care organizations (MCOs) than traditional health plans, a phenomenon that has become known as "managed-care backlash." In a model of the relationship between a patient, insurer, and physician, this paper shows that when the roles of insurer and provider are combined into a single player (as in a staff-model HMO), the equilibrium insurance plan departs from the social optimum, due to the fact that the HMO cannot credibly commit to providing non-least-cost care. In contrast, when the insurer and provider roles are separate, as in fee-for-service insurance, the equilibrium reimbursements for the physician implement the first-best treatment regime at first-best cost. Thus, the relative inability of MCOs to commit to non-least-cost care may account for at least part of managed-care backlash.

[*] John F. Kennedy School of Government, Harvard University. I thank Nabil Al-Najjar, Sandeep Baliga, Karen Eggleston, Elizabeth Lacey, Joe Newhouse, Jim Malcomson, Dan Spulber, Sri Sridharan and Richard Zeckhauser for helpful comments. The financial support of the State Farm Companies Foundation is gratefully acknowledged. All errors are my own. nolan_miller@harvard.edu

1 Introduction

There is a growing body of evidence that people are distrustful of managed care organizations (MCOs) and believe that MCOs are unlikely to provide necessary care should they become severely ill. For example, one study found that only 30% of MCO members trust their health plan to provide the right level of care, as opposed to 55% of people in traditional plans, and that 61% of MCO members believed their health plan was more concerned with saving money than with giving patients the best treatment, compared with 34% of people in traditional plans.[1] This paper presents a partial explanation for this backlash based on differences in the ability of the insurers to make credible commitments to consumers in the managed-care and traditional health plan markets.

Determining the proper treatment for a patient exhibiting nonspecific symptoms is complex. Different diseases can generate similar symptoms, and it is often impossible to reduce the probability of one kind of mistaken diagnosis without increasing the probability of another. For example, it may be impossible to reduce the likelihood of mistakenly performing surgery on a patient who could be treated with drugs without increasing the risk of treating a patient who truly requires surgery with drugs. This is just a manifestation of the familiar statistical trade-off between type I and type II errors.

Because mistakes are inevitable, the best that can be done for patients is to strike a balance between the probabilities of the various types of errors in light of the costs and benefits associated with each. This paper is concerned with studying to what extent health care providers succeed in maximizing patient welfare, and whether performance depends on the structure of the health care sector, i.e., whether the roles of payer and provider are separate or integrated.

In order to study the role of payer-provider integration, this paper focuses on two polar examples. The case where the payer and provider functions are separate is captured by the traditional fee-for-service (FFS) insurance environment, where there is an arm's-length relationship between the insurer and physician. Under this arrangement, physicians are compensated by insurers based on the services they perform. The other extreme case, where the payer and provider functions are completely integrated, is best captured by staff-model HMOs. In this case, physicians are salaried employees of the HMO.[2]

[1] The statistics come from Blendon et al. (1998), which reports the results of a 1997 Kaiser/Harvard/Princeton Survey Research Associates survey. Dranove's (2000) discussion of the evolution of the health care industry discusses the "general distrust" of the managed care industry and reports the results of a number of corroborating studies.

[2] While a prime example, staff-model HMOs have become much less prevalent in recent years. Consequently,

This paper considers the interaction between insurers, physicians, and patients in both the HMO and FFS environments. In each case, the insurer receives a premium payment from patients and in exchange pays for the patients' health care, either through directly employing physicians (HMO) or paying independent physicians based on the type of treatment provided (FFS). Physicians receive a noisy, non-verifiable signal of each patient's health condition and choose which of two possible treatments the patient should receive.

Two incentive problems are of particular importance in understanding the relationship between market structure and performance. The first, which we denote the "insurer-physician" problem, is concerned with how closely the insurer can control the physician's behavior. For example, the insurer may desire to rein in the physician's tendency to prescribe high-cost treatments according to his own judgement rather than the insurer's more restrictive guidelines. When the insurer and physician are combined, as in an HMO, the insurer can directly control the physician's behavior. In effect, it is as if the payer also chooses how the patients should be treated. In contrast, when the payer and provider functions are separate, as in the FFS environment, the insurer must rely on the incentives provided by its contract with the physician to induce the behavior it desires. Thus the set of treatment protocols that the FFS insurer can promise to patients will be limited to those that it can induce the physician to follow. This problem is complicated by the fact that the result of the physician's examination is private and non-verifiable, so that contracts between the insurer and physician cannot be based on this information.[3]

We call the second incentive problem the "insurer-patient" problem. This incentive problem concerns the insurer and patient and revolves around whether the patient will expect the insurer to implement its promised treatment rule. Because the physician's appraisal of the patient's condition is not verifiable and the patient's premium is sunk at the time care is provided, profit-maximizing insurers will be tempted to provide only low-cost care. Thus, patients will expect to receive least-cost care unless something can be done to convince them otherwise.

With respect to this problem, the FFS insurer's disadvantage in not being able to directly control the physician's behavior becomes an advantage. Because the insurer and physician relate

after laying out the two extreme cases, the paper also discusses how the results apply to other arrangements, such as group-model HMOs, which have become the predominant form of managed care organization in recent years (Harvard Managed Care Industry Center Group, 2002).

[3] An additional aspect of the incentive problem between patients and their insurance companies is the issue of whether the insurer will cover all illnesses and perform all treatments. However, since a proper analysis of these issues would require an explicit model of the competition between insurance companies, consideration of these questions is beyond the scope of this paper.

through an arm's-length contract, this contract can be used by the insurer to convince patients that it will, in fact, provide a higher standard of care. On the other hand, the HMO's advantage in directly controlling its physician's incentives becomes a disadvantage. Since the HMO has no means of credibly committing to provide a particular standard of care, patients expect the HMO to provide only lowest-cost care. Given that, the patient's concerns become self-fulfilling. Their concern over cost cutting makes patients unwilling to pay premiums large enough for the insurer to finance anything but the lowest-quality care.

As the preceding discussion suggests, FFS insurers are relatively more vulnerable to the insurer-physician problem, while the HMO structure is relatively more vulnerable to the insurer-patient problem. The main results of this paper derive the equilibrium insurance plan under each market structure and show that while the FFS equilibrium insurance plan coincides with the patient-welfare maximizing plan, the HMO equilibrium is generally not welfare-maximizing. Intuitively, the difference in the equilibrium outcomes arises because in the FFS environment the reimbursement made to the physician following each type of treatment can be chosen so that the physician implements the socially optimal treatment rule, overcoming the insurer-physician incentive problem. Since the contract between the physician and insurer depends only on the treatment performed (which is verifiable) such contracts can be used by the insurer to credibly commit to the socially optimal treatment rule, overcoming the insurer-patient incentive problem. The insurer's commitment to the physician, in effect, is also a commitment to the patient.

Since the HMO employs the physician, it can monitor the physician's behavior and make sure that the physician behaves as desired. Thus the HMO overcomes the insurer-physician incentive problem. However, because there is no arm's-length relationship between payer and provider, the HMO is unable to credibly commit not to implement a cost-minimizing treatment regime, and so the insurer-patient incentive problem persists. As a result, the equilibrium insurance plan in the HMO market will generally differ from the socially optimal insurance plan. This inability to convince consumers that it will provide non-least-cost care may be one justification for managed-care backlash.

There is a large literature on issues of incentives in health care.[4] A number of papers have focussed the interaction between the quality of care provided and the method of provider payment. See for example Ma (1994), Ma (1997), Ma and McGuire (1997), and Ellis (1998). Other papers

[4] See Gaynor (1994) for a useful survey of the market for physicians' services.

focus on the interaction between provision of incentives through supply-side cost sharing, under which providers are less than fully reimbursed for the costs of treatment, and demand-side cost sharing, under which patients are less than fully insured for the cost of treatment. See for example Ellis and McGuire (1986), Seldon (1990), and Ellis and McGuire (1990). Ellis and McGuire (1993) provides an interesting survey of this literature. A robust result in this area is that optimal incentive schemes tend to involve full insurance with providers being paid according to a scheme that is a "mixture" of capitation and partial cost reimbursement. This bears some resemblance to the optimal wages for fee-for-service providers derived in this paper, which involves a cost-based component and a capitation component that sets the physician's profit equal to zero.

The paper most closely related to this one is Chetty (1998). Chetty (1998) considers a problem very similar to the one presented here, although he adopts a stochastic general equilibrium approach in which the premium paid by the patient and the size of the malpractice award in the event of mistaken treatment adjust to bring about equilibrium. Chetty argues that in such a model the outcomes under HMO insurance and fee-for-service insurance are identical. However, he suggests that a game-theoretic investigation of the strategic interaction between the insurer, physician, and patient might yield additional insight. The analysis presented in this paper is such an analysis, and the results differ substantially from those of Chetty. The main difference, the fact that fee-for-service schemes outperform HMO plans, arises from the fact that the present model recognizes the way in which arm's-length contracts between the insurer and physician can allow the insurer to commit to behaviors that the HMO cannot, which is not possible in Chetty's model.

Malcomson (2003) uses a model similar to the one considered here to study the quite different question of whether, in a setting such as the FFS setting in this paper, it is better for the payer to assign the two treatments to the same diagnosis-related group (DRG), making the same payment for each, or assign the two treatments to different DRGs and allowing the reimbursements for the two treatments to differ. The paper characterizes optimal reimbursements in both cases and shows how multiple DRGs can be used to improve patient outcomes by providing incentives to providers and reducing their informational rents.

The remainder of this paper is as follows. Section 2 describes the general model. Section 3 derives the socially optimal treatment rule. Section 4 derives the equilibrium in the FFS and HMO environments and discusses the main results. Section 5 considers some extensions of the basic analysis. Section 6 concludes. An appendix contains a particularly long proof.

2 The Model

Consider the interaction between a patient, insurer, and physician. The patient purchases health coverage from the insurer, who contracts with the physician to provide the patient's care. The physician receives a private, nonverifiable signal of the patient's condition, chooses one of two treatments, and incurs the cost of treatment.

The patient has utility function for money $u(\cdot)$, which is strictly increasing, strictly concave, and twice differentiable. The consumer's initial wealth is w. The patient's total utility is:

$$U(H, w) = H + u(w),$$

where H is the level of the patient's health.[5]

There are two types of patients, denoted a and b. The patient's type can be thought of as which of two conditions he has.[6] The probability that the patient is type a is given by $\pi \in (0, 1)$. The patient does not know his own type and cannot treat himself. Because of this, he must receive health care services from a physician. The physician is assumed to be risk neutral and profit maximizing.

The physician's information about the patient's type consists of a privately observed, nonverifiable signal $x \in [0, 1]$. The distributions of signals for patients of type a and b are given by continuous probability density functions $a(x)$ and $b(x)$, respectively. We assume that $a(x)$ is strictly increasing on $[0, 1]$ and $b(x)$ is strictly decreasing on $[0, 1]$. Thus, as x increases the relative likelihood of that signal being generated by a type a patient increases as well.

Let $f_a(x) = \pi a(x)$ and $f_b(x) = (1 - \pi) b(x)$ give the joint density of a patient generating signal x and being of type a or b, respectively. Let $f(x) = f_a(x) + f_b(x)$. Following the same logic, define functions $F_a(x) \equiv \int_0^x f_a(s)\, ds$, $F_b(x) \equiv \int_0^x f_b(s)\, ds = F(x) - F_a(x)$, and $F(x) \equiv \int_0^x f(s)\, ds$, which give the cumulative probability that a type a patient generates a signal smaller than x, a type b generates a signal smaller than x, and that any patient generates a signal smaller than x, respectively.

There are two treatments, A and B. A patient of type a who is treated with A realizes final

[5] Since the health benefits of treatment are not insurable, we model the health shock as non-monetary.

[6] It is assumed that the patient is ill with probability one. A more general model would assume that the patient becomes ill with probability $\lambda \in (0, 1)$ and, conditional on becoming ill contracts disease a with probability π and disease b with probability $(1 - \pi)$. However, adding such complexity does not affect the results, and so it is omitted for the sake of notational simplicity.

health h_{Aa}. Health outcomes h_{Ba}, h_{Ab}, and h_{Bb} are similarly defined. It is assumed that

$$h_{Aa} > h_{Ba} \text{ and } h_{Bb} > h_{Ab}. \tag{1}$$

The expected cost of treating a patient of type $t \in \{a, b\}$ with treatment $T \in \{A, B\}$ is given by c_{Tt}. For most of the analysis, we make the following assumption, which, combined with the structure imposed on the distribution of signals above, ensures that the expected cost of treating a patient with treatment A relative to that of treating with B is decreasing in x:[7]

$$c_{Ba} + c_{Ab} - c_{Aa} - c_{Bb} > 0. \tag{2}$$

In place of (2), Chetty (1998) makes the stronger assumption that:

$$c_{Aa} < c_{Ba} \text{ and } c_{Bb} < c_{Ab}. \tag{3}$$

Condition (3) implies that treatment A is the least-cost treatment for type-a patients and treatment B is the least-cost treatment for type-b patients.[8] Since (3) clearly implies (2), all of the results in this paper continue to hold if (3) is assumed instead of (2).[9] However, Chetty's assumption (3) rules out important cases that are permissible under (2). For example, (3) rules out the case where that treatment A is always cheaper than treatment B, i.e., $c_{Aa} < c_{Ba}$ and $c_{Ab} < c_{Bb}$, as might be the case where A is drug treatment and B is surgery. It is possible for such a cost structure to satisfy (2).[10] We briefly consider situations where (2) does not hold in Section 5.1.

Throughout the paper, we assume that the costs are borne by the provider of care, i.e., the physician. When the insurer is an HMO, the integration of the payer and provider functions implies that the cost is effectively borne by the insurer. For the purposes of the model, the critical component of the cost of care is the part that falls on the party who decides which treatment the patient should receive, i.e., the physician. Consequently, additional costs, such as the cost incurred by a hospital or laboratory, do not affect the qualitative results.

[7] To see this, differentiate $\left(\frac{f_a(x)}{f(x)}c_{Ba} + \frac{f_b(x)}{f(x)}c_{Bb}\right) - \left(\frac{f_a(x)}{f(x)}c_{Aa} + \frac{f_b(x)}{f(x)}c_{Ab}\right)$ with respect to x and use the assumptions on $f_a(x)$ and $f_b(x)$ above.

[8] Taken together with (1), (3) makes it natural to think of treatment A as the proper treatment for condition a and similarly for treatment B and condition b.

[9] Condition (2) is also implied if mistakes are costly in the sense that $c_{Bb} < c_{Ba}$ and $c_{Aa} < c_{Aa}$.

[10] However, not all such cost structures do, i.e., $c_{Aa} < c_{Ba}$ and $c_{Ab} < c_{Bb}$ does not imply (2).

The insurer is risk neutral, profit maximizing, and one firm in a perfectly competitive industry. The insurer enters into two types of contracts: contracts with the physician and contracts with the patient. While the signal received by the physician is non-contractible, the treatment provided to the patient is. Thus a contract between the insurer and physician consists of reimbursements w_A and w_B paid to the physician following treatments A and B respectively.

The contract between the patient and the insurer has two parts, the *premium* and the *treatment rule.* The premium is the price, p, that the patient pays to the insurer for health coverage. The *treatment rule* defines the terms of the insurance policy. Formally, let a *treatment rule* be a function $T(x) : [0,1] \rightarrow \{A, B\}$ that specifies for any signal x which treatment a patient generating that signal will receive. An *insurance plan* consists of a premium and a treatment rule.

3 Welfare-Maximizing Insurance Plans

We begin by characterizing the insurance plan that maximizes patient welfare, which we denote $(p^*, T^*(x))$. Since patients are risk averse, insurers are risk neutral, and the insurance market is competitive, patients are fully insured and insurers earn zero profit under the welfare-maximizing plan. Taking this into account, the welfare-maximizing treatment rule, $T^*(x)$, solves:

$$\begin{aligned}&\max_{T(x)} \int_{T_A} \left(h_{Aa} + u\left(w-p\right)\right) f_a\left(s\right) + \left(h_{Ab} + u\left(w-p\right)\right) f_b\left(s\right) ds \\ &+ \int_{T_B} \left(h_{Ba} + u\left(w-p\right)\right) f_a\left(s\right) + \left(h_{Bb} + u\left(w-p\right)\right) f_b\left(s\right) ds\end{aligned} \tag{WMP}$$

where $T_A = \{x|T(x) = A\}$, $T_B = \{x|T(x) = B\}$, and

$$p = c_{Aa} \int_{T_A} f_a(s)\, ds + c_{Ab} \int_{T_A} f_b(s)\, ds + c_{Bb} \int_{T_B} f_a(s)\, ds + c_{Ba} \int_{T_B} f_b(s)\, ds.$$

Because a treatment rule can be changed on a set of signals of Lebesgue measure zero without altering the expected utility it offers, there will be a multitude of transfer rules that solve (WMP). However, due to the structure we have placed on the model, the relative likelihood of signal x arising from a type a patient rather than a type b patient increases. Consequently, there is always a welfare-maximizing treatment rule that takes a particularly simple (and intuitively plausible) form.

Lemma 1 *There exists an x^* such that the treatment rule that specifies $T(x) = B$ for $x < x^*$ and $T(x) = A$ for $x \geq x^*$ solves (WMP).*[11]

Proof. *See the Appendix.* ■

Lemma 1 establishes that A is the appropriate treatment for high-x patients and B is the appropriate treatment for low-x patients. We call treatment rules with this structure "cut-off rules" and the $\hat{x}$ that divides patients treated with B (low-x patients) from those treated with B (high-x patients) as the "cut-off value." We will often identify a cut-off rule by its cut-off value.

Using Lemma 1, (WMP) can be rewritten as:

$$\max_{x^* \in [0,1]} u(w - p^*) + \int_0^{x^*} (f_a(s) h_{Ba} + f_b(s) h_{Bb}) \, ds \tag{4}$$
$$+ \int_{x^*}^{1} (f_a(s) h_{Aa} + f_b(s) h_{Ab}) \, ds,$$

where $p^* = F_a(x^*) c_{Ba} + F_b(x^*) c_{Bb} + (\pi - F_a(x^*)) c_{Aa} + ((1-\pi) - F_b(x^*)) c_{Ab}$. Since the objective function is continuous and x^* lies in a closed interval, a solution to this problem exists.

Differentiating the objective function in (4) with respect to x^* to yields first-order necessary condition:

$$-u'(w - p) \frac{dp}{dx^*} + f_a(x^*)(h_{Ba} - h_{Aa}) + f_b(x^*)(h_{Bb} - h_{Ab}) = 0, \tag{5}$$

where $\frac{dp}{dx^*} = f_a(x^*)(c_{Ba} - c_{Aa}) + f_b(x^*)(c_{Bb} - c_{Ab})$. In order to focus on the most interesting case, we assume x^* is unique and interior, i.e., $0 < x^* < 1$.[12]

The terms of (5) are interpreted as follows. The first term, $-u'(w-p)\frac{dp}{dx^*}$, is the change in utility due to the fact that increasing x^* changes the premium, weighted by the marginal utility of income. The next term, $f_a(x^*)(h_{Ba} - h_{Aa})$, is the marginal decrease in health benefits given that when x^* is increased, the marginal type a people who were treated with A will now be treated incorrectly with B. The last term, $f_b(x^*)(h_{Bb} - h_{Ab})$, is the corresponding term due to the fact that the marginal type b people who were treated with A are now treated correctly with B. Hence the second and third terms capture the health benefit trade off between increasing the number of type-a people who are mistakenly treated with B and decreasing the number of type-b people who

[11] Other solutions to (WMP) are essentially the same, differing from this rule only on a set of measure zero.

[12] If we assume (3) instead of (2), it can be shown that the objective function in (4) is concave, and therefore that x^* is unique. However, even if there are multiple maximizers, the main results of the paper persist. That is, the FFS equilibrium will coincide with *one of* the maximizers, and the HMO equilibrium will generally not coincide with any of them.

are mistakenly treated with A, while the first term represents the marginal impact of this change in the treatment rule on the premium and consequently on the patient's final utility from wealth.

4 Equilibrium Analysis

The timing of the game between the insurer, physician, and patient is as follows.

1. The insurer offers the physician a contract $(w_A, w_B, T(x))$ specifying the payments w_A and w_B to be made following each treatment and the treatment rule $T(x)$ that the physician should follow. The physician may either accept or reject this contract. If accepted, the insurer-physician contract is observed by the patient.

2. The insurer offers the patient treatment plan $(p, T(x))$, which the patient may either accept or reject.

3. The patient becomes ill, and the physician receives signal $x \in [0,1]$ about the patient's condition, based upon which he chooses whether to treat the patient with A or B. Following treatment the physician is reimbursed according to the contract with the insurer.

Our notion of equilibrium requires that the patient, insurer, and physician all act optimally. The term "provider" is used in the definitions to refer to the party making the treatment decisions. In the FFS model, the physician is the provider, whereas in the HMO model, treatment decisions are made by the HMO, which combines the insurer and physician roles.

At stage 1, we require:

Incentive Compatibility: Given contract $(w_A, w_B, T(x))$, for each realized signal x, $T(x)$ specifies the treatment that maximizes the provider's expected profit.

Participation: Given contract $(w_A, w_B, T(x))$, the provider's ex ante expected profit is non-negative.[13]

Incentive Compatibility and Participation are important for two reasons. First, if $(w_A, w_B, T(x))$ satisfies these properties, then, given reimbursements w_A and w_B, the physician will follow treatment rule $T(x)$. Second, if the insurer's contract with the physician satisfies these properties,

[13] Alternative specifications of the Physician's participation condition are discussed in Section 5.2.

then the patient should expect the physician to follow treatment rule $T(x)$ because the patient knows that it is in the physician's best interest to do so. To emphasize the second point, we call an insurer-physician contract $(w_A, w_B, T(x))$ **credible** if it satisfies Incentive Compatibility and Participation. Patients have no reason to believe an insurer will deliver standard of care $T(x)$ unless it is backed by a credible insurer-physician contract. For convenience, we will sometimes refer to the treatment rule, rather than the contract, as being credible. In this case it should be understood that a credible treatment rule is accompanied by some reimbursements that make it credible.[14]

At stage 2, we require that the insurer-physician contract be credible, as well as:

Actuarial Fairness: Given the treatment rule, the premium is equal to the expected cost of a patient's care.

Constrained Welfare Maximization: The insurance contract $(p, T(x))$ maximizes the patient's welfare from among all insurance contracts consisting of a credible treatment rule and its actuarially fair premium.

Actuarial Fairness and Constrained Welfare Maximization embody the assumption that the insurance market is competitive, and thus that the market acts to provide patients with the best possible fairly-priced insurance contract. An equilibrium insurance plan consists of a credible insurer-physician contract and the associated actuarially fair premium that maximize the patient's expected welfare.

4.1 Separate Insurer and Provider: Fee-For-Service

In this section, we consider the case where the insurer and provider are separate – i.e. there is an insurance company that pays an independent physician according to a prespecified fee schedule, as in fee-for-service (FFS) insurance. The equilibrium must satisfy the four requirements described above.

Formally, Incentive Compatibility insists that:

$$T(x) = A \text{ if and only if } w_A - \frac{f_a(x)}{f(x)} c_{Aa} - \frac{f_b(x)}{f(x)} c_{Ab} \geq w_B - \frac{f_a(x)}{f(x)} c_{Ba} - \frac{f_b(x)}{f(x)} c_{Bb}, \tag{6}$$

[14] Generally, if there are reimbursements that make treatment rule $T(x)$ credible, then there is a continuum of reimbursements that make $T(x)$ credible.

and Participation requires that:

$$\int_{\{T(x)=A\}} w_A - f_a(x)\, c_{Aa} - f_b(x)\, c_{Ab} dx + \int_{\{T(x)=B\}} w_B - f_a(x)\, c_{Ba} - f_b(x)\, c_{Bb} dx \geq 0. \tag{7}$$

Actuarial fairness implies that the premium is given by:

$$p = \int_{\{T(x)=A\}} f_a(x)\, c_{Aa} + f_b(x)\, c_{Ab} dx + \int_{\{T(x)=B\}} f_a(x)\, c_{Ba} + f_b(x)\, c_{Bb} dx, \tag{8}$$

and Constrained Welfare Maximization implies that the equilibrium insurance plan solves:

$$\begin{aligned}
&\max_{T(x)} \int_{T_A} (h_{Aa} + u(w-p))\, f_a(s) + (h_{Ab} + u(w-p))\, f_b(s)\, ds \\
&+ \int_{T_B} (h_{Ba} + u(w-p))\, f_a(s) + (h_{Bb} + u(w-p))\, f_b(s)\, ds \\
&\text{s.t. } (6),\ (7),\ \text{and}(8).
\end{aligned}$$

To determine the set of credible treatment plans, rewrite (6) as:

$$T(x) = A \text{ if and only if } \left(\frac{f_a(x)}{f(x)} c_{Ba} + \frac{f_b(x)}{f(x)} c_{Bb}\right) - \left(\frac{f_a(x)}{f(x)} c_{Aa} + \frac{f_b(x)}{f(x)} c_{Ab}\right) \geq w_B - w_A. \tag{9}$$

From (9), it is straightforward to characterize the set of incentive-compatible treatment rules.

Lemma 2 *For any $\hat{x} \in [0,1]$, there exist payments w_A and w_B such that the physician maximizes his expected profit by treating patients of type $[0, \hat{x})$ with B and patients of type $[\hat{x}, 1]$ with A.*

Proof. Noting that $f(x) = f_a(x) + f_b(x)$, the left side of the inequality in (9) is decreasing in x, since

$$\begin{aligned}
\frac{d}{dx}\left(\left(\frac{f_a(x)}{f(x)} c_{Ba} + \frac{f_b(x)}{f(x)} c_{Ab}\right) - \left(\frac{f_a(x)}{f(x)} c_{Aa} + \frac{f_b(x)}{f(x)} c_{Ab}\right)\right) &= \\
\left(f_a'(x)\, f_b(x) - f_a(x)\, f_b'(x)\right) \frac{(c_{Ba} - c_{Bb}) + (c_{Ab} - c_{Aa})}{(f_a(x) + f_b(x))^2} &< 0.
\end{aligned} \tag{10}$$

Hence, setting

$$w_B - w_A = \frac{f_a(\hat{x})}{f(\hat{x})} (c_{Ba} - c_{Aa}) + \frac{f_b(\hat{x})}{f(\hat{x})} (c_{Bb} - c_{Ab})$$

implements the desired treatment rule. ■

Lemma 2 establishes that there are payments that make any cut-off treatment rule incentive compatible. Since the inequality in (9) depends on w_A and w_B only through their difference, a constant can be added to both without affecting the physician's incentives. If the physician is paid reimbursements

$$\begin{aligned} w_A(\hat{x}) &= \frac{f_a(\hat{x})}{f(\hat{x})}c_{Aa} + \frac{f_b(\hat{x})}{f(\hat{x})}c_{Ab} + k, \text{ and} & (11) \\ w_B(\hat{x}) &= \frac{f_a(\hat{x})}{f(\hat{x})}c_{Ba} + \frac{f_b(\hat{x})}{f(\hat{x})}c_{Bb} + k, & (12) \end{aligned}$$

then the physician's profit is given by:

$$\begin{aligned} & F(\hat{x})\, w_A(\hat{x}) + (1 - F(\hat{x}))\, w_B(\hat{x}) \\ & - (F_a(\hat{x})\, c_{Ba} + F_b(\hat{x})\, c_{Bb} + (\pi - F_a(\hat{x}))\, c_{Aa} + ((1-\pi) - F_b(\hat{x}))\, c_{Ab}) \\ = \;& k + F(\hat{x}) \left(\frac{f_a(\hat{x})}{f(\hat{x})}c_{Aa} + \frac{f_b(\hat{x})}{f(\hat{x})}c_{Ab} \right) + (1 - F(\hat{x})) \left(\frac{f_a(\hat{x})}{f(\hat{x})}c_{Ba} + \frac{f_b(\hat{x})}{f(\hat{x})}c_{Bb} \right) \\ & - (F_a(\hat{x})\, c_{Ba} + F_b(\hat{x})\, c_{Bb} + (\pi - F_a(\hat{x}))\, c_{Aa} + ((1-\pi) - F_b(\hat{x}))\, c_{Ab}). \end{aligned}$$

Thus, for each $\hat{x}$, choosing constant k equal to:

$$\begin{aligned} k(\hat{x}) &= (F_a(\hat{x})\, c_{Ba} + F_b(\hat{x})\, c_{Bb} + (\pi - F_a(\hat{x}))\, c_{Aa} + ((1-\pi) - F_b(\hat{x}))\, c_{Ab}) & (13) \\ & \quad -F(\hat{x}) \left(\frac{f_a(\hat{x})}{f(\hat{x})}c_{Aa} + \frac{f_b(\hat{x})}{f(\hat{x})}c_{Ab} \right) + (1 - F(\hat{x})) \left(\frac{f_a(\hat{x})}{f(\hat{x})}c_{Ba} + \frac{f_b(\hat{x})}{f(\hat{x})}c_{Bb} \right), \end{aligned}$$

ensures that the Participation condition, (7), is satisfied. This establishes Lemma 3.

Lemma 3 *Any cut-off treatment rule with $\hat{x} \in [0, 1]$ is credible in the fee-for-service environment.*

Given cut-off rule $\hat{x}$, the the actuarially fair premium is:

$$\begin{aligned} p(\hat{x}) &= F(\hat{x})\, w_B + (1 - F(\hat{x}))\, w_A & (14) \\ &= k(\hat{x}) + F(\hat{x}) \left(\frac{f_a(\hat{x})}{f(\hat{x})}c_{Ba} + \frac{f_b(\hat{x})}{f(\hat{x})}c_{Bb} \right) \\ & \quad + (1 - F(\hat{x})) \left(\frac{f_a(\hat{x})}{f(\hat{x})}c_{Aa} + \frac{f_b(\hat{x})}{f(\hat{x})}c_{Ab} \right). \end{aligned}$$

The final equilibrium requirement is Constrained Welfare Maximization. Recall that the insurance plan that maximizes patient welfare has cut-off value x^* and premium p^*. Since this plan is credible (Lemma 3), the equilibrium of the FFS environment involves the welfare-maximizing insurance plan.

Proposition 1 *In the fee-for-service environment, the equilibrium involves insurers paying reimbursements*

$$\begin{aligned} w_A(x^*) &= \frac{f_a(x^*)}{f(x^*)} c_{Aa} + \frac{f_b(x^*)}{f(x^*)} c_{Ab} + k(x^*), \text{ and} \\ w_B(x^*) &= \frac{f_a(x^*)}{f(x^*)} c_{Ba} + \frac{f_b(x^*)}{f(x^*)} c_{Bb} + k(x^*), \end{aligned}$$

and charging premium p^*. Physicians treat patients of type $[0, x^*)$ with B and patients of type $[x^*, 1]$ with A.

Proof. *It only remains to verify that the actuarially fair premium for the efficient rule, found by evaluating (14) at $\hat{x} = x^*$, is equal to the welfare-maximizing premium, p^*. This is easily done by substitution (and must be so, since the efficient decision rule is implemented and both the insurer and physician expect to earn zero profit).* ■

Proposition 1 establishes that the equilibrium insurance plan in the fee-for-service environment is the welfare-maximizing plan. Whenever the welfare-maximizing insurance plan is credible, it must be the equilibrium plan. Hence the key to the fee-for-service environment maximizing patient welfare is that, by judiciously choosing its reimbursements, the insurer makes it incentive compatible for the Physician to provide the welfare-maximizing treatment rule. This serves as a credible commitment to patients.

4.2 Integrated Insurer and Provider: HMO

The HMO environment differs from the FFS environment in that the HMO's integrated insurer-provider structure replaces the arm's-length contracting between the insurer and physician in the FFS environment. Thus the roles played by the physician and insurer in the FFS environment are now both played by the HMO. Most importantly, this implies that the physician and insurer now have the same preferences: both seek to maximize the HMO's expected profit.

Once again, we begin by determining which treatment rules are credible. The critical difference between the fee-for-service and HMO environments is that in the HMO environment, since reimbursements w_A and w_B are purely internal transfers, they do not affect the HMO's expected profit, and therefore cannot be used for incentive purposes, i.e., to make credible commitments to patients. To see this, note that, given signal x, if the HMO treats with A, it expects profit:

$$-w_A + w_A - \left(\frac{f_a(x)}{f(x)} c_{Aa} + \frac{f_b(x)}{f(x)} c_{Ab} \right). \tag{15}$$

If the HMO treats with B, its expected profit is:

$$-w_B + w_B - \left(\frac{f_a(x)}{f(x)} c_{Ba} + \frac{f_b(x)}{f(x)} c_{Bb} \right). \tag{16}$$

Thus, in the HMO environment, an Incentive Compatible treatment rule must satisfy:

$$T(x) = A \text{ if and only if } \left(f_a(x) c_{Ba} + f_b(x) c_{Bb}\right) - \left(f_a(x) c_{Aa} + f_b(x) c_{Ab}\right) \geq 0. \tag{17}$$

Since the left-hand side of the inequality in (17) is independent of w_A and w_B, the insurer cannot use its reimbursement schedule to change the Physician's incentives, and therefore (since the left-hand side of the inequality in (17) is monotone in x) there is only one incentive-compatible treatment rule. This treatment rule is the cut-off rule defined by x^H, where x^H solves:

$$f_a\left(x^H\right)\left(c_{Ba} - c_{Aa}\right) = f_b\left(x^H\right)\left(c_{Ab} - c_{Bb}\right), \tag{18}$$

if (18) has a solution. Otherwise, x^H is equal to 0 if $f_a(0)(c_{Ba} - c_{Aa}) > f_b(0)(c_{Ab} - c_{Bb})$, or equal to 1 if $f_a(1)(c_{Ba} - c_{Aa}) < f_b(1)(c_{Ab} - c_{Bb})$.

The actuarially fair premium associated with cut-off treatment rule x^H is:

$$\begin{aligned} p^H &= F_a\left(x^H\right) c_{Ba} + F_b\left(x^H\right) c_{Bb} \\ &\quad + \left(\pi - F_a\left(x^H\right)\right) c_{Aa} + \left((1-\pi) - F_b\left(x^H\right)\right) c_{Ab}. \end{aligned} \tag{19}$$

Since there is only one credible treatment rule, insurance plan $\left(p^H, x^H\right)$ (trivially) satisfies Constrained Welfare Maximization, and therefore it is the equilibrium insurance plan.

Comparing $\left(p^H, x^H\right)$ with (p^*, x^*) yields Proposition 2 as an immediate consequence.

Proposition 2 *The equilibrium insurance plan in the HMO environment is* $\left(p^H, x^H\right)$*, which coin-*

cides with welfare-maximizing insurance plan (p^*, x^*) only if x^* satisfies:

$$f_a(x^*)(h_{Ba} - h_{Aa}) + f_b(x^*)(h_{Bb} - h_{Ab}) = 0. \tag{20}$$

Proof. *The welfare-maximizing treatment plan is characterized by:*

$$-u'(w-p)\frac{dp}{dx^*} + f_a(x^*)(h_{Ba} - h_{Aa}) + f_b(x^*)(h_{Bb} - h_{Ab}) = 0, \tag{21}$$

where $\frac{dp}{dx^*} = f_a(x^*)(c_{Ba} - c_{Aa}) + f_b(x^*)(c_{Bb} - c_{Ab})$. By (18), $\frac{dp}{dx^*} = 0$ if $x^* = x^H$, from which the conclusion follows. ■

Although it is possible for (20) to hold and hence for the HMO equilibrium to maximize patient welfare, for almost all parameter values (i.e., generically), (20) will not hold. Hence, the HMO equilibrium will generally not maximize patient welfare. At its base, the HMO's preferred treatment rule differs from the welfare-maximizing treatment rule because the former considers only the cost of care, while the latter also takes into account patients' health benefits.

4.3 Market Structure and Performance

Propositions 1 and 2 establish that the FFS environment outperforms the HMO environment. This difference arises from the ability of the fee-for-service insurer to use its contract with the Physician to accomplish two tasks. First, it is able to induce the Physician to follow the welfare-maximizing treatment rule. Second, because this contract is enforceable and observed by the patient, it allows the insurer to commit to a treatment plan that provides more than least-cost care. While it may appear that the incentive problem is the more challenging, this is not the case. In this model, the HMO's failure to achieve efficiency arises from its inability to convince patients it will not provide lowest-cost care. Given this, the patients are unwilling to pay for any higher standard of care, and so the HMO equilibrium involves lowest-cost care. This is true despite the fact that the HMO management can induce its physicians to follow any treatment rule.

Seen in this light, we can recast the discussion in terms of commitment. The FFS environment is one in which the insurer has full commitment power (i.e., it can commit to any treatment rule), and the HMO environment is one in which the insurer has no commitment power.[15] Of course,

[15] Indeed, the fact that the FFS environment must outperform the HMO environment can be seen from arising from the fact that the set of behaviors the FFS insurer can commit to includes the behavior that the HMO can commit to.

in the real world, FFS insurer's lack full commitment power, and HMO insurer's do not totally lack commitment power. Nevertheless, it remains true that the arm's length relationship between physician and insurer in the FFS environment provides the insurer with a method of commitment that the HMO insurer lacks. Indeed, this commitment power is exactly the physician autonomy, albeit directed by the reimbursement schedule, that insurance customers cite as attracting them to more expensive fee-for-service insurance.

The role of commitment accounts for the difference between our results and those of Chetty (1998), who finds in a similar model that the equilibrium outcome in the FFS and HMO environments are identical. Chetty (1998) recognizes that in the fee-for-service environment the insurer faces the problem of inducing the physician to behave as it wants, while the integrated HMO can directly control the physician's behavior. However, he does not consider that the insurer faces the additional problem of credibly communicating its intended behavior to patients. Thus, while it is true that the FFS insurer cannot induce the Physician to behave in any way that the HMO cannot, it is also true that the FFS insurer can credibly commit to a larger set of behaviors than the HMO. Competition then drives the FFS insurer to adopt the plan that maximizes patient welfare and thereby outperform the HMO.

In our model, the HMO most closely resembles a staff-model HMO, where the physicians are salaried employees. Although once the predominant form of managed care organization, staff-model HMOs have become much less prevalent in recent years. According to an analysis by the Harvard Managed Care Industry Study Group (2002), in 1998 only 7% of managed care enrollees belonged to HMOs that employed salaried physicians. The vast majority of the remainder belonged to HMOs that compensated physicians primarily through capitation (58%) or primarily through fee-for-service payments (35%).

In the context of the present analysis, the behavior of HMOs compensating physicians using capitation payments is likely to be very similar to that of the staff-model HMO. Since the capitation payment received by the physician group does not vary with the treatment provided to patients, the contract between the HMO and the physician group cannot be used to make commitments to patients. And, given that it is capitated, the physician group will have the same incentive to provide only least-cost care as the HMO. Consequently, the (theoretical) outcome for the capitated group-model HMO will be the same as for the staff-model HMO.

To the extent that the contract between an HMO compensating physicians on a fee-for-service basis can be used to make commitments to patients, the (theoretical) performance of these organizations should resemble the FFS environment studied in this paper. However, by virtue of the limits it places on patients' choices, it may be difficult for the HMO to convince its patients that the physicians are, in fact, operating independently. If the HMO is unable to accomplish this, patients may believe that the HMO will attempt to induce physicians to cut costs, leading its performance to resemble that of the staff-model HMO.

In the present analysis, we allowed only a single commitment device, the insurer-physician contract. However, other commitment devices can help the HMO to perform better. For example, reputational concerns may induce the HMO to provide more costly care as a way of attracting new customers and retaining older ones. To the extent that patients understand these mechanisms, they may be effective in credibly communicating the HMO's intentions.

Recently, a number of managed care organizations in California,[16] Massachusetts,[17] and other states have begun adopting compensation policies that reward physicians for high-quality care. Although there are certainly other contributing factors, the analysis in this paper suggests that such policies may be an attempt by HMOs to make the same sorts of commitments to providing higher-quality care than FFS insurers do. By making public commitments to their doctors that they will be rewarded for providing high-quality care, the HMO may simultaneously convince its potential customers that they will not attempt to skimp. In the language of the present analysis, these policies may make high-quality treatment plans credible, leading to improved performance by the HMO sector relative to FFS.

5 Alternative Specifications

5.1 Cost Structure

Until now, we have assumed that treatment costs satisfy (2), i.e., declining relative expected cost of treatment A. This assumption ensures that, given fixed payments w_A and w_B, the provider follows a cut-off rule where patients generating high signals are treated with A and patients generating low

[16] Fong, Tony. "Bonus plan will reward doctors for quality of care," *San Diego Union-Tribune*, January 16, 2002, pg. C-1. Benko, Laura B. "Bonus Time; California associate plans to reward physicians for good patient care," *Modern Healthcare*, January 14, 2002, pg 18.

[17] Kowalcayk, Liz. "For doctors, bonuses for quality care," *The Boston Globe,* November 7, 2002, pg. A1.

signals are treated with B. If (2) does not hold, i.e.,

$$c_{Ba} + c_{Ab} - c_{Aa} - c_{Bb} < 0, \tag{22}$$

then the provider's incentives are reversed. Since (22) implies that the expected cost of treatment A relative to that of treatment B *increases* in x, the provider will choose to treat high-signal patients with B and low-signal patients with A. The interesting case is where the efficient treatment rule continues to treat high-x patients with A and low-x patients with B, and so we maintain this assumption.[18] Thus, under (22), provider behavior is "backward" relative to the welfare-maximizing rule. Nevertheless, the main result of this paper, that fee-for-service arrangements outperform HMOs, persists.

The argument is straightforward and applies the same ideas developed above. Begin with the HMO. Since the HMO is not able to make commitments to provide anything other than least-cost care, the only credible treatment rule is the least-cost one. Note that under (22), the least-cost treatment rule is "backward," treating high-x patients with B and low-x patients with A. Let $\tilde{x}^H$ (which may be an endpoint) denote the cut-off value for this treatment rule.

Next, consider the FFS environment. As before, the insurer can use the reimbursements it pays to the physician to commit to various treatment rules. In particular, under (22), reimbursements w_A and w_B make the treatment rule that treats $[0, \tilde{x})$ with A and $[\tilde{x}, 1]$ with B is incentive compatible if:

$$w_B - w_A = \frac{f_a(\tilde{x})}{f(\tilde{x})}(c_{Ba} - c_{Aa}) + \frac{f_b(\tilde{x})}{f(\tilde{x})}(c_{Bb} - c_{Ab}). \tag{23}$$

It is straightforward to verify that the right-hand side of (23) is monotone in $\tilde{x}$ and therefore that any "backward" cut-off rule is incentive-compatible in the FFS environment.

Although it is somewhat complicated to derive the equilibrium in the FFS environment (since it involves finding the welfare-maximizing "backward" treatment rule, which generally depends on the shape of $u(\cdot)$), doing so is not necessary in order to show the FFS environment once again outperforms the HMO. To see why, note that backward cut-off rule $\tilde{x}^H$ is credible in the FFS environment, and so the FFS environment must perform at least as well as the HMO. In fact, the FFS strictly outperforms the HMO unless the cost-minimizing backward cut-off rule $\tilde{x}^H$ is also the

[18] This can be true despite (22) if the difference in benefits under the two treatments is large relative to the difference in cost. If it becomes efficient to treat high-x patients with B and low-x patients with A, then this case is exactly like our earlier case with the roles of high and low signals reversed.

welfare-maximizing backward cut-off rule. Thus, superior commitment power once again leads the FFS structure to outperform the HMO.[19]

5.2 Participation Constraints

In the preceding analysis, the provider's participation decision is modeled as taking place before the patient's signal is observed. That is, the provider must expect to break even on each patient. However, due to the fact that patients differ in the expected cost of caring for them, under any scheme that breaks even ex ante there will be signals following which the provider expects to earn a positive profit and signals following which the provider expects to earn a negative profit. Further, it will most likely be that the provider expects to earn money overall on all patients receiving one treatment and lost money on all those treated with the other. If it is possible for the provider to refuse to treat a patient after having examined him or else refuse to provide one of the treatments altogether, then the relevant requirement is not an ex ante participation constraint, but rather an interim one, i.e., upon observing any signal x, the provider must expect to break even.

In this section we briefly explore the impact the requirement that the physician must break even on each patient,[20] returning to our initial assumption, (2), that as x increases the expected cost of treatment A decreases relative to treatment B. In either market structure, the essence of the interim participation constraint is that the payment the provider receives for treating a patient must cover the expected cost of the patient's care, conditioned on the patient's signal. Since the question is more straightforward in the FFS case, we begin there.

The Incentive Compatibility, Actuarial Fairness, and Constrained Welfare Maximization requirements remain the same.[21] Thus it remains that any cut-off rule is incentive-compatible. In particular, reimbursements satisfying

$$w_B - w_A = \frac{f_a(\hat{x})}{f(\hat{x})}(c_{Ba} - c_{Aa}) + \frac{f_b(\hat{x})}{f(\hat{x})}(c_{Bb} - c_{Ab}), \tag{24}$$

[19] While the FFS equilibrium is still superior to the HMO equilibrium, it need not be welfare-maximizing.

[20] The analysis in the case where the physician must break even on each treatment is similar.

[21] By leaving the actuarial fairness requirement unchanged, we are assuming that the *insurer* still faces an ex ante participation constraint.

induce cut-off rule $\hat{x}$. The interim participation constraints require that for all x:

$$\begin{aligned} w_A &\geq \frac{f_a(x)}{f(x)} c_{Aa} + \frac{f_b(x)}{f(x)} c_{Ab}, \text{ and} \quad (25) \\ w_B &\geq \frac{f_a(x)}{f(x)} c_{Ba} + \frac{f_b(x)}{f(x)} c_{Bb}. \quad (26) \end{aligned}$$

Note that, for any $\hat{x}$, there are wages w_A and w_B satisfying (24), and (25), (26). Thus any cut-off rule $\hat{x}$ is credible even with interim participation constraints. Actuarial fairness holds that the premium must equal the insurer's expected cost of care:

$$p = (F(x) w_B + (1 - F(x) w_A)). \quad (27)$$

The constrained welfare-maximizing insurance plan therefore solves:

$$\max_{x \in [0,1]} u(w - p) + \int_0^x (f_a(s) h_{Ba} + f_b(s) h_{Bb}) \, ds + \int_x^1 (f_a(s) h_{Aa} + f_b(s) h_{Ab}) \, ds$$

subject to (24), (25), and (26). Denote the solution x^{**}.

Although it is not particularly instructive to further characterize the solution to this problem, a number of observations are immediate. First, since the premium differs under ex ante and interim participation constraints, generally x^{**} will differ from x^{*}. Thus, the imposition of interim participation constraints distorts the treatment rule. Second, since the interim participation constraints imply that the physician earns a non-negative profit following every signal and interim expected treatment costs differ, the physician earns a strictly positive profit following some signals and hence a strictly positive profit overall. Because this profit implies a higher premium and different standard of care, the imposition of interim participation constraints reduces patient welfare relative to the case of only ex ante participation constraints.

The question of how to integrate interim participation constraints into the HMO structure is somewhat less straightforward than the FFS case, due to the lack of separation between the insurer and provider roles. When the HMO accepts the patient's premium, it agrees to cover the patient's treatment cost in the event that he needs care. Because of this, the HMO is obligated to provide the patient with *some* treatment. Since we have assumed that the HMO can directly control the physician's incentives, i.e., that it can force the physician to follow any treatment rule the HMO desires, it also reasonable to assume that the HMO can force its physicians to treat all patients, regardless of their signals. Further, since the HMO is responsible for the patient's care, if one HMO physician is able to refuse a particular patient, that patient will only turn to another HMO physician. Thus, one reasonable conclusion to draw is that the imposition of the interim participation constraint has no effect on the equilibrium. The HMO follows cut-off rule x^H and charges p^H. In any case, because the patient's condition is "covered," and the HMO will have to pay for some treatment, however the interim participation constraint is modeled it is likely to have less of an impact on the HMO market than the FFS market

To the extent that interim participation constraints are important, their imposition may lead the HMO to outperform FFS. However, it need not necessarily do so. In the presence of interim participation constraints, the FFS physician earns an information rent. However, given that the physician will earn a rent, the FFS equilibrium maximizes patient welfare from among all possible cut-off rules. On the other hand, because the HMO exercises tighter control over its physicians, it is unaffected by the imposition of the interim participation constraint. However, it is still unable to commit to providing anything but lowest-cost care. Thus, whether FFS or HMO performs better in the presence of interim participation constraints depends on whether the distortion due to the physician's information rent or the HMO's inability to commit to non-cost-minimizing behavior is more harmful.

It is unclear just how important are interim participation issues in FFS markets. On the one hand, codes of medical ethics prohibit physicians from refusing to treat a patient merely because the patient is expected to be high cost. This idea is supported in the legal system by "patient abandonment" law, which prohibits a physician from abandoning a patient without giving the patient ample time to find an equally qualified replacement. On the other hand, even if a physician does not "dump" a patient outright, there are any number of ways the physician can encourage the patient to seek another doctor, including delaying appointments and having an unpleasant bedside manner.

6 Conclusion

This paper has considered the interaction between a patient, physician, and insurer as it relates to decisions about which treatment the patient should receive when diagnosis is inherently uncertain. The main result is that the FFS insurer is able to use the arm's length relationship between itself and physicians in order to make commitments to provide a higher standard of care that HMOs, which lack such commitment devices. Consequently, the FFS market equilibrium maximizes patient welfare, while the HMO equilibrium does not.

The commitment gap studied in this paper provides a partial explanation for the phenomenon of managed-care backlash described in the introduction.[22] In the context of this paper, it is argued that this distrust stems from (1) potential customers' belief that HMOs have a strong incentive to choose the least-cost treatment available, and (2) the fact that the HMO's inability to make credible commitments to the contrary leaves it powerless to alter this impression. In light of this, HMOs attempting to earn consumers' trust may face an uphill battle. If consumers believe the HMO will provide only least-cost care, they will only be willing to join if the HMO's premium is low. However, if the HMO receives only small payments from its customers, it might be unable to provide higher-cost care even if it wanted to. Thus, HMOs face a challenge in breaking out of this circle of mistrust. One promising method may be adopting systems that reward physicians for providing quality care (such as those mentioned earlier) or instituting a broad system of HMO "report cards." However, the efficacy of such systems in improving consumer trust has yet to be established.

Several qualifications to the main result are in order. First, as discussed above, if there is a real possibility that the FFS physician will refuse to treat patients or offer treatments that are expected to be high cost, then this may improve HMO performance relative to FFS, and, if such issues are very important, may even lead the HMO to outperform the FFS market.

Second, throughout this analysis, we have assumed that the providers in the FFS and HMO frameworks have the same cost structure. However, this may not be the case. The purpose of HMOs is to control costs, and to the extent that HMOs are able to enact real cost reductions (as opposed to providing less generous care), the efficiency gain due to these cost reductions must be weighed in favor of the HMO when comparing the performance of the two market structures.

Third, although the paper shows that properly chosen reimbursements induce the efficient treatment rule in the FFS model and that competition in the FFS market will drive the equilibrium toward efficiency, it assumes that the insurers properly understand the physician's production function. However, understanding physician costs is not easy. Physicians perform many different

[22] Rosenthal and Newhouse (2002) provide an alternative, though related, explanation for backlash. They argue that backlash against managed care arises from the fact that people believe that MCO ration care without regard to consumers' preferences about which services should be rationed.

treatments for many different conditions, and some costs (such as office staff and rent) are not directly attributable to patient care. Further, the insurer may not be able to directly observe the physician's costs.

Finally, in this paper we have assumed perfect competition in the insurance sector. It remains an open question how the conclusions presented here would be affected by the presence of market power. Nevertheless, while these other factors may be important, the issues of trust and credible commitment raised in this paper will continue to play a role in determining the relative performance of FFS- and HMO-type markets, even under more general circumstances.

References

[1] Blendon, R., M. Brodie, J. Benson, D. Altman, L. Levitt, T. Hoff, and L. Hugick. 1998. Understanding the Managed Care Backlash. *Health Affairs.* 17 : 80 - 94.

[2] Chetty, V. K. 1998. Stochastic technology, production organization and costs. *Journal of Health Economics* 17 : 187-210.

[3] Dranove, D. 2000. *The Economic Evolution of American Health Care.* Princeton University Press: Princeton.

[4] Ellis, R. P. 1998. Creaming, skimping and dumping: provider competition on the intensive and extensive margins. *Journal of Health Economics* 17 : 537-555.

[5] Ellis, R.P. and T.G. McGuire. 1986. Provider behavior under prospective reimbursement: cost sharing and supply. *Journal of Health Economics* 5 : 129-151.

[6] Ellis, R. P. and T.G. McGuire. 1993. Supply-side and demand-side cost sharing in health care. *Journal of Economic Perspectives* 7 : 135-151.

[7] Ellis, R. P. and T.G. McGuire. 1990. Optimal payment systems for health services, *Journal of Health Economics* 9 : 375-396.

[8] Gaynor, M. 1994. Issues in the industrial organization of the market for physician services. *Journal of Economics and Management Strategy* 3 : 211-255.

[9] Harvard Managed Care Industry Center Group. 2002. Managed Care: An Industry Snapshot. *Inquiry* 39: 207-220.

[10] Ma, C. A. 1994. Health care payment systems: cost and quality incentives. *Journal of Economics and Management Strategy* 3 : 93-112.

[11] Ma C. A. and T.G. McGuire. 1997. Optimal health insurance and provider payment. *American Economics Review* 87 : 685-704.

[12] Malcomson, J. 2003. The Specification of Daignosis-Related Groups. University of Oxford Department of Economics Discussion Paper #162.

[13] Rosenthal, M. and J. Newhouse. 2002. Managed Care and Efficient Rationing. *Journal of Health Care Finance.* 28 : 1-10.

[14] Selden, T. M. 1990. A model of capitation. *Journal of Health Economics* 9 : 397-409.

A Proof of Lemma 1

Proof of Lemma 1: Let

$$x^* = \sup_{x\in[0,1]} \{x|\Pr(t < x|T(x) = A) = 0\}.$$

That is, let x^* be the largest signal such that the set of signals smaller than x^* that are treated with A has measure zero. If $x^* = 0$ or $x^* = 1$, this completes the proof. If $x^* \in (0,1)$, by the definition of x^*, it must be that there exists an $x^+ > x^*$ such that

$$\Pr\left(x^* < t < x^+|T(x) = A\right) = x^+ - x^*.$$

Let $T_B^+ = \{x > x^*|T(x) = B\}$. If $\Pr\left(T_B^+\right) = 0$ then this completes the proof. If $\Pr\left(T_B^+\right) > 0$, then choose $T_B^{++} \subset T_B^+ \cap (x^+, 1]$ such that $0 < \Pr\left(T_B^{++}\right) < x^+ - x^*$ and let $T_A^+ \subset [x^*, x^+]$ such that $\Pr\left(T_A^+\right) = \Pr\left(T_B^{++}\right)$. Thus T_A^+ and T_B^{++} are two sets of signals that occur with equal probability and such that every signal in T_B^{++} is larger than every signal in T_A^+. Next, define a new treatment rule $T'(x)$ such that

$$T'(x) = \begin{cases} B & x \in T_A^+ \\ A & x \in T_B^{++} \\ T(x) & \text{otherwise} \end{cases}.$$

Hence $T'(x)$ differs from $T(x)$ only in that set T_A^+ is treated with B instead of A and set T_B^{++} is treated with A instead of B. Let

$$\begin{aligned} q_{T_A^+} &= \int_{T_A^+} f_a \\ q_{T_B^{++}} &= \int_{T_B^{++}} f_a \end{aligned}$$

be the probability of a type a player conditional on drawing a signal in T_A^+ or T_B^{++} respectively. Since $f_a^{'} > 0$, this implies that $q_{T_A^+} < q_{T_B^{++}}$. Also, note that

$$\int_S f_a = \int_S f - f_b$$

and that $\int_{T_A^+} f = \int_{T_B^{++}} f \equiv q$.

The change in expected health is given by

$$\begin{aligned} &\int_{T_A^+} f_a(s)(h_{Ba} - h_{Aa}) + f_b(s)(h_{Bb} - h_{Ab})\,ds \\ &- \int_{T_B^{++}} f_a(s)(h_{Ba} - h_{Aa}) + f_b(s)(h_{Bb} - h_{Ab})\,ds. \end{aligned}$$

Rearranging terms, this is equal to

$$\begin{aligned} &(h_{Ba} - h_{Aa})\left(\int_{T_A^+} f_a(s)\,ds - \int_{T_B^{++}} f_a(s)\,ds\right) \\ &+ (h_{Bb} - h_{Ab})\left(\int_{T_A^+} f_b(s)\,ds - \int_{T_B^{++}} f_b(s)\,ds\right) \end{aligned}$$

or

$$(h_{Ba} - h_{Aa})\left(q_{T_A^+} - q_{T_B^{++}}\right) + (h_{Bb} - h_{Ab})\left(q_{T_B^{++}} - q_{T_A^+}\right).$$

Since $h_{Ba} - h_{Aa} < 0$, $h_{Bb} - h_{Ab} > 0$, and $q_{T_A^+} < q_{T_B^{++}}$ this difference is positive. Thus the expected health benefit of treatment rule $T'(x)$ is greater than the expected health benefit of treatment rule $T(x)$.

The change in expected cost due to implementing treatment rule $T'(x)$ instead of treatment rule $T(x)$ is given by

$$\begin{aligned}&\int_{T_A^+} f_a(s)(c_{Ba}-c_{Aa})+f_b(s)(c_{Bb}-c_{Ab})\,ds\\&-\int_{T_B^{++}} f_a(s)(c_{Ba}-c_{Aa})+f_b(s)(c_{Bb}-c_{Ab})\,ds.\end{aligned}$$

Rearranging terms:

$$\begin{aligned}(c_{Ba}-c_{Aa})\quad&\left(\int_{T_A^+} f_a(s)\,ds-\int_{T_B^{++}} f_a(s)\,ds\right)\\&+(c_{Bb}-c_{Ab})\left(\int_{T_A^+} f_b(s)\,ds-\int_{T_B^{++}} f_b(s)\,ds\right)\end{aligned}$$

which is equal to

$$\begin{aligned}&(c_{Ba}-c_{Aa})\left(q_{T_A^+}-q_{T_B^{++}}\right)+(c_{Bb}-c_{Ab})\left(q_{T_B^{++}}-q_{T_A^+}\right)\\&((c_{Ba}-c_{Aa})-(c_{Bb}-c_{Ab}))\left(q_{T_A^+}-q_{T_B^{++}}\right)\end{aligned}$$

By definition,

$$c_{Ba}-c_{Bb}>0>c_{Aa}-c_{Ab}$$

which implies that

$$(c_{Ba}-c_{Aa})-(c_{Bb}-c_{Ab})>0$$

Thus $((c_{Ba}-c_{Aa})-(c_{Bb}-c_{Ab}))\left(q_{T_A^+}-q_{T_B^{++}}\right)<0$, and the change in expected cost is negative. Hence the proposed change increases expected health and decreases expected cost and thus expected premium, improving the patient's expected utility and contradicting the possibility that $\Pr\left(T_B^{++}\right)>0$.■