Managed Care, Physician Incentives, and Norms of Medical Practice: Racing to the Bottom or Pulling to the Top?

by

David J. Cooper
Department of Economics
Weatherhead School of Management
Case Western Reserve University
10900 Euclid Avenue
Cleveland, OH 44106
djc13@po.cwru.edu

and
James B. Rebitzer
Department of Economics
Weatherhead School of Management
Case Western Reserve University
10900 Euclid Avenue
Cleveland, OH 44106
jbr@po.cwru.edu
National Bureau of Economics Research
The Levy Economics Institute of Bard College

9/24/02

We would like to thank Esther Gal-Or, Mari Rege, and Larry Samuelson for their useful comments, as well as seminar participants at Case Western Reserve and the Veteran Administration's 13th Annual Health Economics Conference. We would like to thank Bethia Cullis for her work as a research assistant on this paper. The usual caveat applies.

ABSTRACT

The incentive contracts that managed care organizations write with physicians have generated considerable controversy. Critics fear that if informational asymmetries inhibit patients from directly assessing the quality of care provided by their physician, competition will lead to a "race to the bottom" in which managed care plans induce physicians to offer only minimal levels of care.

To analyze this issue we propose a model of competition between managed care organizations. The model serves for both physician incentive contracts and HMO product market strategies in an environment of extreme information asymmetry—physicians perceive quality of care perfectly, and patients don't perceive it at all.

We find that even in this stark setting, managed care organizations need not race to the bottom. Rather, the combination of product differentiation and physician practice norms causes managed care organizations to race to differing market niches, with some providing high levels of care as a means of assembling large physician networks. We also find that relative physician practice norms, defined endogenously by the standards of medical care prevailing in a market, exert a "pull to the top" that raises the quality of care provided by all managed care organizations in the market.

We conclude by considering the implications of our model for public policies designed to limit the influence of HMO incentive systems.

INTRODUCTION

Dating back at least to Cournot (1838), students of industrial organization have built formal models to gain insight about product market competition among firms. A similarly rich tool-kit of formal models has been developed for understanding incentive systems in organizations.[1] For the most part, these two literatures have evolved independently. Models of product market competition typically ignore the inner workings of firms, while models of incentives in organizations pay scant attention to the firm's competitive environment. This disconnection is unfortunate because, in many settings, the choices of product market and incentive strategies are inextricably linked. One such setting, of particular importance to economics and public policy, is the delivery of medical services via managed care organizations. In this paper we combine a model of a differentiated product oligopoly with a model of physician incentives to study how competition between managed care organizations shapes the style of medical care physicians practice and, as a consequence, the welfare of their members.

Since Congress passed the HMO Act of 1973, managed care has become the dominant form of health insurance in the United States.[2] Firms in this industry adopt a bewildering variety of organizational forms and acronyms such as: health maintenance organizations (HMOs); preferred provider organizations (PPOs); and point-of-service plans (POSs).[3] The common characteristic of these plans and organizations is the combination of an insurance function with systems for managing the actions physicians take on behalf of their patients. To help fix ideas, we will focus the subsequent discussion on HMOs; in particular we model the network (or independent practice) HMOs that have become an important part of the managed care industry.[4] In contrast to the "staff model," HMOs in which physicians are hired as employees, network HMOs typically do *not* employ physicians. Rather they contract with individual physician

---

[1] For recent reviews of the incentives literature see Gibbons and Waldman (1999); and Prendergast (1999). For a comprehensive discussion of the literature on physician incentives see McGuire (2000).

[2] Ma and McGuire (forthcoming) cite evidence that roughly 75% of the privately insured in the U.S. receive health care under some form of management by their health plan.

[3] See Fox (1997) and Robinson (1999) for institutional background.

[4] These entities comprise one of the largest segments of the managed care market, roughly 40 percent of total HMO enrollment in 1998 (InterStudy, 1999).

practices (or with associations of such practices) to provide medical services. The contracts that network HMOs write are usually non-exclusive, meaning that individual physicians in the network can see patients who are not members of the HMO. The contracts typically also include monetary rewards for successfully controlling medical utilization costs. These cost control incentives are the source of intense public controversy and high stakes litigation (Gosfield, 1997).

The core issue in the public debate stems from the information asymmetries that prevent patients from directly evaluating the quality of care provided by their physician. If patients cannot adequately assess care quality, critics fear that competition will trigger an HMO led "race to the bottom" in which physicians operate under severe cost controls and managed care plans offer only minimal levels of care.[5]

To address this issue, we propose a model of the managed care market place that solves for both physician incentive contracts and HMO product market strategies in an environment of extreme information asymmetry—physicians perceive the quality of care they offer perfectly and their patients do not perceive it at all. Even under these stark conditions, we find that competition between HMOs for patients and for physicians constrains "racing to the bottom."

At the center of our analysis is the process by which HMOs assemble physician networks and the role these networks play in the competition for customers.[6] Consistent with our assumption of extremely asymmetric information, we do not directly include quality of care in consumers' preferences. In this set-up direct monitoring by consumers cannot play any role in preventing HMOs from cutting costs by inducing physicians to provide low quality care. We instead model potential HMO members as having preferences over the cost of insurance and the size of the HMO's physician network. All potential members prefer larger networks because they are more likely to find a physician they like in larger networks, but differ in their willingness to

---

[5] While our exposition focuses on network HMOs, our key conclusions can extend to any type of managed care organization. For this reason, we drop the term "network HMO" in favor of the more parsimonious term HMO.

[6] In the market for commercial health insurance, the "customers" for whom the HMOs compete are typically employers who purchase insurance on behalf of their employees. Competition for employees will compel employers to choose policies with features their employees desire. Indeed, in a perfectly competitive labor market, employers would choose insurance plans with the same combination of cost and quality that the marginal employee would choose for herself. Thus, we simplify our discussion by leaving the employer out of the picture. We treat each individual employee as if they were the HMO's direct customer.

pay for this access. HMOs therefore can try to attract consumers by offering access to a large network of physicians and/or by providing low costs.

Unlike consumers, physicians are assumed to understand the quality of care being delivered. In our model, physicians dislike working under high powered, cost-containment incentives because these systems force them to adopt lower cost, possibly less effective, practice styles. If physicians differ in how averse they are to medical cost-containment (and the resulting changes in their practice style), then attracting large numbers of physicians to a network will require that the HMO write relatively low-powered incentive contracts. An essential element of competitive strategy for HMOs will be the resulting tradeoff between providing large physician networks and maintaining low prices.

The result, as we will demonstrate below, is differentiated product market competition. Customers who place a relatively high value on physician choice and a relatively low value on the cost of insurance will choose HMOs with large networks and high premiums. Conversely, customers who place a relatively high value on low cost insurance will choose HMOs with low premiums and small physician networks. Because assembling a large network of physicians requires relatively lax cost controls, the HMOs will also provide differing practice styles, with the "up-market" HMO providing a relatively generous practice to its members. The interaction between incentives and product market competition—consumers demand large physician networks, and HMOs can only provide these networks by weakening cost-containment incentives—acts to prevent a "race to the bottom."

Once we derive equilibrium incentive contracts and medical practice styles, we use our model to examine the efficacy of two commonly proposed regulatory strategies for limiting the influence of physician incentive contracts: (1) capping the proportion of "at risk" income in physician contracts; and (2) making HMOs legally liable for the adverse medical consequences attributed to their cost-containment systems. We find that both these policies have the intended effect of reducing cost-pressures on physicians, but at the price of increasing the cost of medical services and decreasing the proportion of the population with insurance coverage. Thus, the welfare implications of these policies are ambiguous. Some members, particularly those with a high willingness to pay for physician choice, are made better off by these policy interventions while others, particularly the newly uninsured, are made worse off.

All of our conclusions regarding quality of care depend on physicians having appropriate preferences about the style of medicine they practice. In our model, we refer to these preferences as "norms of medical practice" and treat them as having both an exogenous and endogenous component. [7] The exogenous component of norms specifies the minimum cost practice style a physician will tolerate. This minimum acceptable level of care is determined by exogenous factors such as the state of medical knowledge, medical ethics, and the threat of malpractice suits. The endogenous component of norms is based on relative, rather than absolute, levels of care. We posit that physicians are averse to offering less generous care to some patients solely on the basis of these patients' HMO affiliation. Similarly, physicians dislike adopting a practice style that affords less generous care than is provided by physicians elsewhere in the marketplace. These relative norms may be driven by doctors' desire for social status, or simply by a fear of malpractice suits.[8] In either case, the endogenous component of norms implies that the willingness of physicians to accept an HMO's incentive contract will depend on the incentive contracts prevailing elsewhere in the market place. The presence of these endogenous norms has important economic implications: they amplify forces that move the market *away* from a "race to the bottom" while, at the same time, reducing the scope for product differentiation between HMOs. Instead of racing to the bottom in search of lower costs, HMOs may well find themselves pulled to the top by their need to attract physicians to their networks.

The plan of the paper is as follows. In the next section, we briefly review the empirical evidence concerning physician incentives. Section 3 presents the model and solves for an equilibrium. Section 4 applies the model to regulatory strategies proposed for the managed care industry. We conclude by discussing directions for future research.

---

[7] We use the term "norms" to highlight the fact that physicians derive utility directly from the delivery of appropriate medical care (in addition to the utility derived from the income generated by the provision of medical services). For a discussions of norm based decision making in economic transactions see Kandel and Lazear (1992) and March (1994).

[8] The social comparisons in our model of relative practice norms are similar to Encinosa, Gaynor, and Rebtizer (2001), Gaynor and Rebitzer's (2001) model of physician medical practices, and to Kranton's model of social identity (Akerlof and Kranton 2000).

PHYSICIAL INCENTIVES AND MANAGED CARE

Our model relies heavily on three stylized facts about physicians' responses to incentive schemes: (1) physicians will respond to financial incentives by changing their practice style, (2) physicians' responses to financial incentives are shaped by absolute and relative practice norms, and (3) HMOs take these norms into account when writing incentive contracts. In this section we summarize evidence relating to these stylized facts.

Response to Incentives

A number of recent econometric studies suggest that the practice style of physicians is influenced by the explicit and implicit financial incentives under which they operate. Kessler and McClellan (1996), for example, find that reforms in state malpractice laws have an economically and statistically significant effect on patient expenditures for the treatment of heart disease.[9] Barro and Beaulieu (2000) study the effect of a switch from fixed salary to profit sharing at a set of physician practices owned by a hospital chain. They find that the introduction of a performance-based pay plan increased profitability significantly, primarily because physicians increased the number of patients they saw.[10] Barro and Beaulieu's study looked at compensation practices in a fee for service setting. In contrast, Gaynor, Rebitzer, and Taylor (2001) examine the effect of cost-containment incentives within the type of HMO network we model. The HMO they studied wrote incentive contracts with each of the physicians in their network. These contracts provided substantial monetary rewards to those primary care physicians who kept the medical utilization charges of their patients below target levels. The study found that the incentive contracts led to reduced medical utilization costs, with physicians sharing in the financial benefits of these cost reductions. The common conclusion of these studies is that physicians' choice of practice style *does* respond to financial incentives.[11]

---

[9] This result also informs our analysis of public policies that alter HMO legal liability for malpractice claims.

[10] Barro and Beaulieu also found that the terms of incentive contracts can influence a physician's decision to affiliate with a plan or practice. If incentive arrangements influence physicians' exit and entry patterns, then HMOs need to take preferences regarding incentives into account when constructing physician networks.

[11] Robinson (2001) reaches a similar conclusion based on his review of the medical literature.

Incentives and Physician Practice Norms

In making the case for the importance of physician practice norms in the analysis of incentives, we rely on both direct and indirect evidence. Direct evidence that physicians experience disutility when incentives are perceived to influence the level of care quality comes from a recent survey of physician attitudes published in the *New England Journal of Medicine*. This paper states: "Our findings suggest that bonuses based on limitation of referrals and on productivity heighten physicians' "performance anxiety" and their perceptions that care may be compromised in these areas…" (Grumbach et.al. 1998; p. 1520). The same study also reports that when physicians perceive pressure to limit referrals or improve productivity in ways that compromise care, their satisfaction with their practice declines.

Where one might interpret the survey results as indicating the presence of absolute practice norms, there is also indirect evidence suggesting the importance of relative practice norms. It is well established that physicians' practice styles have a local flavor, varying in persistent and meaningful ways from one location to another. [12] The source of these "small area variations" remains mysterious—they are not accounted for by variations in underlying clinical conditions, cost of treatment, or patient incomes. Some analysts have suggested that these geographic practice patterns are the result of physicians learning by observing the practice style and clinical decisions of other physicians in the vicinity (Phelps 1992). In other words, physicians' norms for acceptable practices are not necessarily based on some absolute external standard, but instead are determined endogenously by the practices of other physicians in the area. To the extent that physicians are responsive to financial incentives, this implies that the efficacy of one HMO's incentive plan will depend on the incentive plans being used by other HMOs in the area.

Another piece of evidence supporting the existence of physician norms, as well as HMOs' sensitivity to these norms, comes from studying the incentive contracts HMOs offer their physicians. If physicians dislike incentives that force them to compromise care, one might expect HMOs to write incentive contracts in ways that mitigate incentive pressure for patients most in need of care. This is what Gaynor *et al* (2001) found in their case study described above. Specifically, they found that the HMO relied on incentive contracts with built-in safeguards to

[12] For an excellent discussion of this large literature see Phelps (1992).

protect seriously ill patients. For the purposes of calculating cost-containment bonuses, the primary care physicians in the HMO's network were only held responsible for the first $15,000 of costs per year generated by each patient. This "stop-loss" provision was intended to remove cost-containment pressures for seriously ill patients and the statistical evidence presented by Gaynor *et al* suggests that it had the intended effect. That this provision was aimed at physicians and not consumers is evident in how the HMO viewed it within their overall strategy. The general impression at the HMO was that purchasers were much more responsive to premium levels and the number of physicians in the network than to assertions regarding quality. The physician incentive contracts were not advertised and even if the contracts had become common knowledge, they were so complex that only the most sophisticated buyers would have been able to understand the significance of the stop-loss provisions. In other words, the incentive contract was weakened not to attract patients through a more permissive style of medical practice, but rather to attract physicians.

This evidence on HMO incentive contracts comes from a case study of a single organization. Ideally we would like to know more generally whether HMOs shape incentives to accommodate physician practice norms. No general database of HMO contracts exists, but we can infer something more about these incentive contracts by examining the effect of HMOs on care quality. If HMOs did not try to shape incentives in accordance with physician norms, one would expect to see corresponding effects on the quality of clinical outcomes. However, the few econometric studies that have directly examined the issue have generally found no HMO effect on care outcomes. Cutler, McClellan, and Newhouse (2000) compare the treatment for HMO members with others insured by traditional indemnity plans. They find that HMOs have 30% to 40% lower expenditures than traditional plans, but that actual treatments and health outcomes differ little across types of plans. Similar results are found in a study of cost and treatment patterns for Massachusetts state and local government employees by Altman, Cutler, and Zeckhauser (2000). They studied costs and treatment intensity for eight serious medical conditions. The authors found that average HMO costs were 40 percent lower than those of an indemnity plan offering insurance to the same pool of employees. However, these lower costs are due to selection and the HMO's ability to negotiate low prices for treatments, not to any difference in the treatments used.

We interpret the absence of an HMO effect on clinical outcomes for serious illnesses as reflecting the reluctance of HMOs to push physicians into treatment decisions that they might find objectionable for either moral or legal reasons.[13]

A MODEL OF PHYSICIAN INCENTIVES AND HMO COMPETITION

We model competition among HMOs as an extensive form game with three stages. The players in this game are two HMOs and the population of doctors that might treat patients insured by these HMOs.[14] In the first stage of the game, the HMOs simultaneously set the number of doctors they want in their network and the quantity of HMO members they intend to service. Prices for each HMO are then set to clear the market.[15] In the second stage of the game, the HMOs write incentive contracts for the physicians in their network. HMOs are constrained to write contracts that yield the promised number of physicians for their network. In the final stage of the game, doctors make two related decisions: which HMO (or HMOs) to join and what style of medical practice to adopt. Both of these decisions are shaped by the incentive contracts HMOs offer. We use subgame perfection as a solution concept and solve the model via backward induction. Our exposition of the model therefore begins with the final stage and works backward in time to reach the first stage.

Stage 3: Physician Choice of Network Affiliation and Practice Style

Consider an HMO whose network is composed entirely of primary care physicians. In this HMO, PCPs are "responsible" for the care of their panel of HMO members in both a clinical and

---

[13] Fears of malpractice suits are likely a central component in the formation of physician norms, and we discuss these in subsequent sections.

[14] We could extend our model to include any number of HMOs as well as old-style indemnity plans, but the key results are easiest to communicate in a model with only two HMOs. To further simplify things we assume that each HMO offers only one plan. Allowing HMOs to offer multiple plans would complicate the model without altering its basic conclusions.

[15] This model of product market competition in the first stage builds off of Gal-Or's (1985) model of differentiated product oligopoly. In the context of HMO competition, the Cournot assumption means that HMOs set target market shares and then set prices to achieve those targets. This focus on market share is roughly consistent with informal discussions about strategy the authors have had with a local HMO.

economic sense. Clinically, primary care physicians must approve any actions that incur medical utilization costs, e.g. drug prescriptions, referrals to specialists etc. Economically, primary care physicians are also held "responsible", via incentive contracts, for the medical costs incurred by their patients. Managing care by making the primary care physician the "gatekeeper" to resources is common in the managed care industry.

The HMO writes incentive contracts with the primary care physicians in its network. For simplicity we focus on linear contracts having two parameters, a capitation rate and a cost share.[16] The capitation rate for incentive contracts offered by HMO i, represented by k_i, is a flat fee that the HMO pays the physician for each HMO member in the physician's patient panel. The cost share parameter for HMO i, represented by d_i, is the fraction of incurred medical costs that the physician must bear. Without loss of generality, we assume that it is HMO 1 that will have the relatively low powered incentives and HMO 2 that has the high powered incentives, i.e. that $d_2 > d_1$.

If incentives are to matter, physicians must be free to adopt different styles of medical practice in response to different levels of cost sharing. Think of these medical styles as short-hand descriptions of the strategies primary care physicians use to treat the patients that arrive in their office. For example, a primary care physician may decide to send every case of acne to a dermatologist and every ankle sprain to a sports medicine specialist. This style of medical practice would typically generate more medical expenses than one in which the primary care physician tried to treat the acne or the sprains themselves. We index practice styles by s and think of s as increasing with the costliness of a practice style. More specifically we write the cost per patient of a physician adopting practice style s as:

$$c(s) = \beta s^2 \qquad \textbf{(1)}$$

Physicians make choices about their practice style based on a combination of clinical and financial considerations. In writing down physicians' preferences over practice styles, we assume that a higher practice style is, ceteris paribus, preferable. As professionals, physicians have a

[16] In principle the model can be extended to include direct and intrusive monitoring, the informal 'steering' of patients to low-cost providers and more complex contracts, but these features would greatly complicate the analysis. See Ma and McGuire (forthcoming) for a discussion of other incentive instruments available to HMOs.

sense of what resources patients *themselves* would choose to have spent on their medical treatment if patients had the knowledge and information to make these decisions. In addition, physicians may have intellectual or scientific motives to use the latest and best technology on their patients. Physicians may also practice "defensive medicine" and use tests and procedures to preempt future malpractice suits. Each of these factors will cause physicians to prefer more expensive over less expensive styles. Indeed, if physicians did not generally prefer more expensive practice styles, there would be little need for HMOs to write contracts with incentives for controlling costs.

We incorporate absolute physician norms into the model by assuming there exists a minimum acceptable practice style α for each physician. This norm acts largely through doctors' participation decisions for the HMOs.[17] We stipulate that a physician will join a network if this action meets two criteria. First, being a member of the network must generate non-negative utility when operating with the utility-maximizing practice style (as derived below in (3) – (5)). Second, this utility maximizing practice style must be greater than α, the minimal acceptable practice style.[18] Under our maintained assumption that patients do not have the information or expertise to adequately assess physician practice styles, it is a physician's judgment about what is minimally acceptable rather than the marginal patient's preferences, which limits the physician's choice of styles. We capture heterogeneity in physician judgments by allowing α to be uniformly distributed on the interval 0 to A, where $A > 0$.

To incorporate relative practice norms, we assume that physicians observe the practice style adopted by other physicians in the market and experience a reduction in utility when they adopt a practice style that is lower than the most expensive style prevailing in the market. There are at least three reasons why physicians may compare their practice style to the maximum style prevailing in the market. First, they may themselves have patients in both HMOs in the market and they may dislike offering less expensive treatment to patients in HMO 2 (where incentive

---

[17] We also incorporate α directly into the physicians' objective function, but this is a matter of analytical convenience rather than an economically significant aspect of the model.

[18] More specifically, we treat physicians as making a two-stage decision, first choosing whether to join HMO i and then choosing a style, s_i, for patients from HMO i. At the second stage, doctors choose the unconstrained maximum, ignoring the absolute practice norm. In the first stage they anticipate their second stage decision and only join if it yields an acceptable style ($s_i \geq \alpha$).

contracts are higher-powered) than HMO 1. Second, there may be status-seeking behavior under which physicians aim to "keep up with the Jones" in terms of the medical treatments they offer their patients. Finally, it may be that physicians who practice a style of medicine that is relatively less expensive than other physicians feel more vulnerable to malpractice claims.

Formally, let $\hat{s}$ denote the maximum style present in the local market and $\lambda s_i(\hat{s} - s_i)$ the disutility derived from offering a style below the maximum.[19] The variable λ is positive, implying that as long as $s_i > \hat{s}/2$, marginal reductions in practice style reduce physician utility. When $s_i = \hat{s}/2$ the disutility of choosing an inexpensive practice style is maximized, so beyond this point ($s_i < \hat{s}/2$) we assume the marginal effect of further reductions in s_i is 0.[20]

Letting q_i be the number of patients the physician has from HMO i, we combine physician preferences for income, their preferences for more expensive styles of practice, the disutility generated by relative practice norms, and the constraints imposed by the incentive contracts into the following objective function:

$$u_D^i = q_i\left((k_i - d\beta s_i^2) + \gamma(s_i - \alpha) - \lambda s_i(\hat{s} - s_i)\right) \text{ if } s_i > \frac{\hat{s}}{2} \quad \textbf{(2a)}$$

or

$$u_D^i = q_i\left((k_i - d\beta s_i^2) + \gamma(s_i - \alpha) - \lambda\frac{\hat{s}^2}{4}\right) \text{ if } s_i \leq \frac{\hat{s}}{2} \quad \textbf{(2b)}$$

---

[19] In assessing $\hat{s}$, the physician is assumed to *not* include his own style. In other words, $\hat{s}$ is the highest practice style used by *another* physician. While $\hat{s}$ is endogenous to the model, it is exogenous from the point of view of a physician. This assumption allows us to avoid having a discontinuity at $s_i = \hat{s}$. The value of $\hat{s}$ is assumed to be known by all physicians.

[20] We chose this specific functional form because it allows us to obtain a simple closed form solution for the equilibrium and expresses the key idea that physician utility drops the further away the chosen style is from the maximum. A notable feature of this functional form is that the marginal disutility of lowering one's style below the maximum decreases the farther one gets from the maximum. This form would be sensible if the saliency of the comparisons decreased as the difference between practice styles grew. In an unpublished appendix, we present an alternative specification that has the opposite property, i.e. that the marginal disutility of comparisons increases with distance from the maximum. Specifically we model the disutility from practicing relatively low cost medicine as $\lambda(\hat{s} - s_i)^2$ for $\hat{s} < s_i$ and zero otherwise. We derive a closed form solution for this alternative function and show using simulations that it has the same basic properties as the version of the model presented here. We reserve these results for an unpublished appendix because of the very complex equations that characterize the closed-form solution.

where $0 \leq d_i \leq 0$ and $\gamma > 0$. The first term of (2a) is the income earned for each member served from HMO i, the second term is the per-member utility derived from the direct returns to adopting a costly practice style, and the final term is the disutility from choosing a practice style below the maximum.

Having fully specified the utility function, we can now determine the style physicians adopt after choosing to join an HMO. Conditional on having joined, we get the following first order condition for doctors in HMO i's network.[21]

$$-2d_i\beta s_i + \gamma - \lambda\hat{s} + 2\lambda s_i = 0 \qquad i \in \{1,2\} \qquad \textbf{(3)}$$

To close the model, (4) adds in the equilibrium condition that $\hat{s}$ equals the higher of the two practice styles chosen by physicians.

$$\hat{s} = \max[s_1, s_2] \qquad \textbf{(4)}$$

Noting that $s_2 < s_1$ because $d_2 > d_1$ by construction, doctors will choose the following styles in HMO1 and HMO 2:

$$s_2 = \frac{\gamma - \lambda s_1}{2(d_2\beta - \lambda)} \text{ and } s_1 = \frac{\gamma}{2d_1\beta - \lambda} \qquad \textbf{(5)}$$

The important point to take away from (5) is that increases in incentives to control costs (represented by increases in d_1 or d_2) have the effect of reducing the costliness of the practice styles physicians adopt.

Stage 2: Incentive Contracts and HMO Cost Functions

In this stage of the game, HMOs have committed to attracting a certain number of doctors, δ_i, to their network. In the network HMOs we analyze here, a doctor "joins" a network by agreeing to receive HMO members as patients, thereby accepting the HMO's incentive contract. Specifically, the doctor agrees to receive the capitation rate, k, and the cost share parameter, d.

---

[21] Note that α does not appear in the first order condition. This implies that all physicians, subject to joining an HMO's network, select the same practice style, a property that greatly simplifies the modeling to follow. It does not imply that all physicians appear identical to consumers, since physicians can differ along many dimensions (e.g. location, bedside manners, availability of extended office hours) other than practice style.

The second order condition for the physician's maximization problem holds so long as $d_i\beta > \lambda$. Since the optimal style s_i goes to infinity as d_i goes to λ/β from above, this condition always holds for any incentive contract the HMO would wish to write.

HMO 1 chooses values of k and d to minimize per patient costs subject to the participation constraints that must be met to build a network with δ_1 doctors.[22] Recall that by assumption $d_1 < d_2$, implying that physicians in HMO 1's network will employ a higher (or more costly) practice style than those in HMO 2's network. Let D be the total number of available physicians. HMO 1's cost minimization problem is as follows, where s_1 is a function of d_1.

$$\min_{k_1,d_1} \; (k_1 + (1 - d_1)\beta s_1^2) \qquad \textbf{(6)}$$

$$\text{subject to } k_1 - d_1\beta s_1^2 + \gamma\left(s_1 - \frac{A\delta_1}{D}\right) \geq 0 \text{ and } s_1 \geq \frac{A\delta_1}{D} \qquad \textbf{(7)}$$

In (7) we give the two physician participation constraints. The first constraint, that physicians must always have non-negative utility under HMO 1's incentive contract, holds with equality. Otherwise, the HMO could always reduce costs by cutting k_i.

The second participation constraint states that if HMO 1 wishes to attract δ_1 physicians to its network, it must write a contract such that the optimal style chosen by physicians working under these contracts equals or exceeds what the marginal physician judges to be the minimum acceptable practice style.[23] Cost minimization requires that this participation constraint must also hold with equality. If it did not, the HMO could always reduce costs further by increasing the share parameter, d_1.[24] Thus it is the marginal physician's judgment of the minimum acceptable style, rather than the judgment of the marginal HMO member, that determines the style that prevails in the HMO. Substituting (7) into (6) we derive HMO 1's cost function.

$$c_1(q_1,\delta_1) = q_1\beta K\delta_1^2, \text{ where } K \equiv \left(\frac{A}{D}\right)^2 \qquad \textbf{(8)}$$

We derive HMO 2's cost function analogously. HMO 2 chooses d_2 and k_2 to minimize per patient costs while attracting δ_2 physicians into its network.

---

[22] Because costs are linear in the number of patients, minimizing the HMO's cost per patient is equivalent to minimizing costs. Cost minimization is implied by profit maximization.

[23] Since α is distributed uniformly between 0 and A, the marginal physician in HMO 1's network of size δ_1 will have $\alpha = (\delta_1/D)A$. Rearranging this equation gives the second constraint in (7).

[24] To see this, it is sufficient to note that $\partial c/\partial d_1 < 0$ for all $d_1 < 1$. Substituting the first constraint (which holds with equality) from (7) into (6) and differentiating, $\partial c/\partial d_1 = dc/ds_1 * \partial s_1/\partial d_1$. Given that s_1 is a decreasing function of d_1, as shown by differentiating (5), and costs are an increasing function of s_1, the derivative $\partial c/\partial d_1$ must be negative.

$$\min_{k_2,d_2}(k_2 + (1 - d_2)\beta s_2^2) \qquad \textbf{(9)}$$

$$\text{subject to } k_2 - d_2\beta s\tfrac{2}{2} + \gamma\left(s_2 - \frac{A\delta_2}{D}\right) - \lambda s_2(s_1 - s_2) \geq 0 \text{ and } s_2 \geq \frac{A\delta_2}{D} \qquad \textbf{(10)}$$

The constraints in (10) are the same as those for HMO 1 except for the final term of the first constraint, $\lambda s_2(s_1 - s_2)$. This term reflects payments that HMO 2 must give physicians to compensate them for the disutility of practicing at less than the highest style. Comparing the participation constraints in (7) and (10), it is clear that the physicians who join HMO 2's network will be a strict subset of those who join HMO 1's physician network.[25]

In HMO 2, as in HMO 1, cost minimization ensures that both participation constraints hold with equality. Substituting (10) into (9) we derive the cost function given by (11). Note that HMO 2's costs depend on the size of HMO 1's network. The size of these cost spillovers depends on the strength of physicians' relative practice norms: the more potent the relative norms, the greater the spillover [26]

$$c_2(q_2,\delta_1,\delta_2) = q_2\left(\beta\delta_2^2 + \lambda(\delta_1 - \delta_2)\delta_2\right)K, \text{ where } K \equiv \left(\frac{A}{D}\right)^2 \qquad \textbf{(11)}$$

Stage 1: Market Equilibrium

We begin by considering the preferences of the consumers in the market. Each consumer can decide to purchase membership in one of the two HMOs or to purchase no health insurance at all. To highlight the information asymmetries that raise fears of a "race to the bottom," we adopt the strong assumption that *consumers cannot perceive the style of care provided by physicians in the network nor can they infer this style from the size of HMO's networks*. Consumers in our model do, however, have preferences over physicians and anticipate that they are more likely to

[25] It may also be that some physicians do not join either network. We can think of these physicians as working exclusively with patients having traditional indemnity insurance. To simplify the exposition, we do not incorporate this sector into our formal model.

[26] The possibility of spillover effects between HMO networks was first noted and analyzed by Beaulieu (2000).

find a physician they like in larger networks.[27]

For these reasons, we posit that, ceteris paribus, consumers prefer HMOs with large physician networks. More formally, we assign each consumer a parameter, x, that determines the strength of preferences for physician choice and hence network size. This parameter is uniformly distributed over the interval [0, X] with X indicating the greatest possible preference for access to a larger network of doctors. We write a consumer's utility in terms of premium costs, p, and network size, δ. Since all consumers prefer larger networks ceteris paribus, b is strictly positive.

$$u(\delta,x:p) = bx\delta - p \qquad \textbf{(12)}$$

By identifying the value of x for the marginal members of HMOs 1 and 2, we can use the consumer's utility function to derive inverse demand functions. For convenience, we assume that the total population size is also X. We maintain our notational convention that HMO 1 has relatively low-powered incentives, which is equivalent to assuming that HMO 1 chooses to provide a larger network of physicians ($\delta_1 \geq \delta_2$). Therefore HMO 2 will set a price such that the marginal member will be indifferent between joining HMO 2 and going without insurance.[28]

$$p_2 = \delta_2 b(X - q_1 - q_2) \quad \textbf{(13)}$$

Likewise, HMO 1 will select a price so that the marginal member of HMO 1 is indifferent between joining either HMO 1 or HMO 2.

$$p_1 = \delta_1 b(X - q_1) - b\delta_2 q_2 \qquad \textbf{(14)}$$

Using these inverse demand functions, we can write the HMOs' profit functions as:

$$\pi_1 = q_1\left(\delta_1 b(X - q_1) - b\delta_2 q_2 - \beta K\delta_1^2\right) \quad \textbf{(15)}$$

$$\pi_2 = q_2\left(\delta_2 b(X - q_1 - q_2) - \beta K\delta_2^2 - \lambda K(\delta_1 - \delta_2)\delta_2\right) \qquad \textbf{(16)}$$

---

[27] Even if patients cannot evaluate the clinical practice styles of physicians, they can still have preferences over physicians because they can evaluate non-clinical attributes of doctors such as location of offices, office hours, age, gender, and communication style. Factors such as recommendations from peers may also lead to patients having references over physicians. Such factors could be built explicitly into the model by using a Hotelling type framework. However, this would complicate the model while gaining us little insight.

[28] Since $\delta_1 \geq \delta_2$, $p_1 > p_2$ or no consumers would join HMO 2. Since utility is an increasing function of x, the segment of available customers [0,X] must be divided into two segments with the upper segment joining one of the HMOs and the lower segment joining neither. Finally, the condition for joining HMO 1 over HMO 2 is $bx\delta_1 - p_1 \geq bx\delta_2 - p_2$. Rearranging, this becomes $x \geq (p_1 - p_2)/b(\delta_1 - \delta_2)$. Thus, the segment of insured consumers can further be subdivided into two segments, with the upper segment joining HMO 1 and the lower joining HMO 2.

The two HMOs simultaneously choose their respective levels of q and δ. In the Nash equilibrium, each HMO maximizes its profits taking q and δ for the other HMO as fixed. Thus the equilibrium must satisfy the four following first-order conditions:[29]

$$\frac{\partial \pi_1}{\partial q_1} = b(X\delta_1 - 2\delta_1 q_1 - \delta_2 q_2) - K\beta\delta_1^2 = 0 \qquad \textbf{(17)}$$

$$\frac{\partial \pi_1}{\partial \delta_1} = bq_1(X - q_1) - 2\beta K\delta_1 q_1 = 0 \qquad \textbf{(18)}$$

$$\frac{\partial \pi_2}{\partial q_2} = b\delta_2(X - q_1 - 2q_2) - K\beta\delta_2^2 - K\lambda(\delta_1 - \delta_2)\delta_2 = 0 \quad \textbf{(19)}$$

$$\frac{\partial \pi_2}{\partial \delta_2} = bq_2(X - q_1 - q_2) - 2\beta K\delta_2 q_2 - \lambda K\delta_1 q_2 + 2\lambda K\delta_2 q_2 = 0 \qquad \textbf{(20)}$$

It is straight-forward, if tedious, to solve this system of first-order conditions for closed form expressions of each HMO's equilibrium values of q and δ:

$$q_1 = \left(\frac{5\beta^2 - 5\beta\lambda - \lambda^2}{23\beta^2 - 23\beta\lambda - \lambda^2}\right)X \qquad \textbf{(21)}$$

$$\delta_1 = \left(\frac{9b(\beta - \lambda)}{23\beta^2 - 23\beta\lambda - \lambda^2}\right)\frac{X}{K} \qquad \textbf{(22)}$$

$$q_2 = \left(\frac{6\beta^2 - 9\beta\lambda + 3\lambda^2}{23\beta^2 - 23\beta\lambda - \lambda^2}\right)X \qquad \textbf{(23)}$$

$$\delta_2 = \left(\frac{3b(2\beta - \lambda)}{23\beta^2 - 23\beta\lambda - \lambda^2}\right)\frac{X}{K} \qquad \textbf{(24)}$$

Fulfilling the first-order conditions given by (17) through (20), along with the second-order conditions, guarantees that each HMO's equilibrium strategy is a local maximum of the payoff function given the other's choice. Because the payoff functions are not quasi-concave, this is a necessary but not sufficient condition for Nash equilibrium. In technical notes available

[29] We are assuming that parameter values are such that both HMOs will employ interior solutions for δ. For simplicity, we are assuming that each physician can serve an unlimited number of patients. For greater verisimilitude, we could add a constraint on the ratio of q_i/δ_i. If this constraint were to be binding, it would provide additional push against a race to the bottom.

from the authors we prove that equations (21) through (24) represent a global profit maximum for each firm as well.[30] We also show that the second-order conditions hold.

Combining (22) and (24), we get the following relationship between the equilibrium network sizes of the two HMOs.

$$\delta_2 = \left(\frac{2\beta - \lambda}{3(\beta - \lambda)} \right) \delta_1 \qquad \textbf{(25)}$$

It follows that our assumption of product differentiation ($\delta_2 < \delta_1$), requires that that $\lambda \leq 0.5\beta$. Thus HMOs will engage in product differentiation provided that relative practice norms are not "too strong" relative to the costs incurred by more generous practice styles.

Having solved for equilibrium values of q and δ for each HMO, it is trivial to solve for equilibrium prices, profits and costs for each HMO. Given that the marginal physician's participation constraints hold with equality for both HMOs, it is similarly easy to derive the equilibrium incentive pay parameters, k and d, and physician practice style, s, for each HMO. Increases in δ *must* be accompanied by a reduction in d and an increase in s.

Discussion of the Market Equilibrium

In the introduction to this paper we raised the possibility of a "race to the bottom" in the quality of care provided by HMOs. Under this scenario, HMOs, responding to competitive pressures, slash costs by employing powerful incentive contracts with physicians. Consumers, unable to judge the quality of care they are getting and thereby unable to vote with their feet, end up receiving increasingly poor medical care. The occurrence of such a downward spiral in care quality seems plausible, but examining the equilibrium described by equations (21)–(24) suggests that such fears may be unfounded.

---

[30] To prove that equations (21) - (24) are a global maximum we must prove that neither HMO can improve its payoff by switching roles, i.e. that, given HMO 1's network, HMO 2 can't do better by building a larger network than HMO 1. Analogously we must also demonstrate that given HMO 2's network, HMO 1 can't do better by building a smaller network than HMO 2. To establish this we derive each HMO's optimal strategy if it switched roles with the other. We then we compare each HMO's maximum payoff after switching roles with the payoff from remaining in the current role. We find that so long as $\lambda \leq 0.5\beta$, the equilibrium in (21)–(24) represents a global profit maximum. Since this is also the condition for the HMOs providing differentiated products in equilibrium, it holds for all cases of interest.

To make this discussion more concrete we need to define what is meant by the market hitting "bottom." An obvious but extreme definition of the bottom is when HMOs uniformly offer the lowest feasible level of care. For our model, this would be equivalent to saying all HMOs induce a practice style, s, equal to zero, the minimum acceptable to any physician in the market place. By examination of (22) and (24), we can see that no such "race to the bottom" occurs. This result is driven by the presence of absolute practice norms. Even if we eliminated all competition from our model and analyzed a monopoly, an HMO without any doctors is like a car without an engine. The need to get at least some physicians to join its network will force all HMOs to induce a style above the minimum.

Even with less extreme definitions, no race to the bottom is observed in our model. For example, we might say competition causes a race to the bottom if all HMOs are pushed down to the lowest practice style provided within the market. Once again, this does not occur. As long as λ is not too large, the market will offer differentiated products and, by extension, differing practice styles.[31] HMO 1 charges consumers a relatively high price for the combination of weak physician incentives and large physician networks that yield a more generous style of medicine. HMO 2, in contrast, offers lower premiums and induces its physicians to adopt a more cost-conscious medical style by writing high-powered incentive contracts. HMO 1 is better off differentiating itself from HMO 2 and enjoying market power in its chosen niche than trying to race HMO 2 for the lowest costs of providing medical care. Once again, physicians' absolute practice norms are playing a critical role. Unlike many industrial organization models, cutting costs does not necessarily make you a stronger competitor here. Tougher cost control incentives come at the price of offering a less attractive product to consumers—specifically, an HMO will have to offer a smaller physician network. For an HMO that has specialized in providing its customers with a plethora of physician options, this is an undesirable tradeoff. In a world with product differentiation, HMOs don't race to the bottom, they race to their market niches.

---

[31] This result is not likely to be undone by adverse selection. In numerical simulations available from the authors, we demonstrate that in settings where x, the parameter governing an individual consumer's willingness to pay for large networks, is positively correlated with the HMO member's expected medical costs, the HMOs still engage in differentiated product market competition in which HMO 1 offers larger physician networks and lower powered physician incentives than HMO 2.

If the combination of product differentiation and absolute physician norms vitiates a "race to the bottom", relative practice norms introduce a "pull to the top". To see this, consider the effect on equilibrium strategies of increasing the potency of physicians' relative practice norms, i.e. of increasing the parameter λ. Differentiating (21) through (24) with respect to λ and maintaining the necessary and sufficient condition for product differentiation ($\lambda \leq 0.5\beta$) we get the following:

$$\frac{dq_1}{d\lambda} = \frac{-18\lambda\beta(2\beta - \lambda)X}{\left(23\beta^2 - 23\beta\lambda - \lambda^2\right)^2 K} < 0 \qquad \textbf{(26)}$$

$$\frac{d\delta_1}{d\lambda} = \frac{9b\lambda(2\beta - \lambda)X}{\left(23\beta^2 - 23\beta\lambda - \lambda^2\right)^2 K^2} > 0 \qquad \textbf{(27)}$$

$$\frac{dq_2}{d\lambda} = \frac{-3\beta\left(23\beta^2 - 50\beta\lambda + 26\lambda^2\right)X}{\left(23\beta^2 - 23\beta\lambda - \lambda^2\right)^2 K} < 0 \qquad \textbf{(28)}$$

$$\frac{d\delta_2}{d\lambda} = \frac{3b\left(23\beta^2 + 4\beta\lambda - \lambda^2\right)X}{\left(23\beta^2 - 23\beta\lambda - \lambda^2\right)^2 K^2} > 0 \qquad \textbf{(29)}$$

From (29), we find increasing the importance of relative practice norms causes HMO 2 to increase the size of its physician network. This is to be expected as heightened relative practice norms increase the cost of getting physicians to adopt a practice style less generous than HMO 1's. It therefore becomes less advantageous for HMO 2 to strongly differentiate itself from HMO 1 by using tight incentives to keep costs (and hence prices) down for its plan. As HMO 2's network size increases, cost-containment incentives are reduced and premiums rise. HMO 1 responds to HMO 2's incursion into the "up scale" insurance market by increasing the size of its own network as shown by (27). The presence of relative practice norms acts as a brake on any race to the bottom, pulling the practice styles of both HMOs up above what would otherwise prevail.[32]

A final comparative static result can be found by differentiating (25):

[32] The pull to the top generated by relative practice norms depends on the existence of differentiated products. The effect exists because physicians can compare their work for HMOs with tight cost controls with the work they (or others) are doing for HMOs with looser controls.

$$\frac{d(\delta_2/\delta_1)}{d\lambda} = \frac{\beta}{(\beta - \lambda)^2} > 0 \qquad \textbf{(30)}$$

When relative practice norms increase in importance, product differentiation falls as the low cost HMO becomes more similar to the high cost HMO. Thus the "pull from the top" due to relative practice norms causes the low-cost HMO to more closely approximate the high-cost HMO's incentive and product market strategies.

To summarize, absolute physician norms and product differentiation combine to prevent a race to the bottom in quality of care, even in the presence of severe informational asymmetries. Relative norms about practice styles act as a pull from the top, moving the quality of care in all market niches to even higher levels.

This conclusion raises a natural question: do the relative physician norms that undermine low-cost/high incentive HMOs also increase consumer welfare? [33] The answer turns out to be ambiguous. From (26) and (28) we see that an increase in λ reduces the number of consumers choosing either HMO 1 or HMO 2. Since an increase in λ reduces membership in *both* HMOs, there must occur a corresponding increase in the number of uninsured. These newly uninsured are, by revealed preference, worse off than before the increase in λ. After all, they had the opportunity to opt out of insurance before the increase in premiums and preferred to purchase insurance. However, it can also be shown that utility increases for consumers with a sufficiently strong willingness to pay for physician choice.[34] Other consumers with less willingness to pay for large physician networks are made worse by increasing relative practice norms, even though they remain insured. For these consumers, the increase in premiums resulting from an increase in λ will lead to a reduction in utility even though they gain access to more physicians. To summarize, an increase in physician practice norms helps consumers who care most about access to large networks and harms those who are most sensitive to prices.

PUBLIC POLICY

Concern over the adverse consequences of "managed care" has grown with the increased importance of HMOs in the U.S. healthcare system. Although the managed care industry has always been subject to regulation at the federal and state level (see Robinson 1999), there is growing interest in public policy that more directly influences HMO incentive systems (Gosfield 1997).

In this section, we consider two broad regulatory strategies for shaping physician incentives: (1) imposing caps on the proportion of "at risk" income allowed in physician contracts; and (2) making HMOs legally liable for the adverse medical consequences attributed to their cost-containment systems. The first strategy is embodied in physician incentive plan (PIP) regulations implemented in 1997 by the Health Care Financing Administration (Gosfield 1997). These regulations required that incentive contracts could not place more than 25% of physician income "at risk," i.e. no more than 25% of income could be linked to performance objectives.[35] The second strategy is embodied in proposals to modify the Employee Retirement and Income Security Act (ERISA) in order to make HMOs liable for damages linked to their cost-containment systems (Havighurst 2000). Some of these proposed changes to ERISA have, in recent years, been included in various proposals for "Patients' Bill of Rights" or "Patient Protection Act" legislation (Studdert, Sage, Gresenz, and Hensler 1999).[36]

---

[33] We do not place much emphasis on profits in our analysis, because the results are sensitive to functional form assumptions. In the context of these assumptions, however, we can show profits of HMO 1 increase as λ increases. The profits of HMO 2, in contrast, are not a monotonic function of λ.

[34] The welfare implications of increasing λ for those who remain insured are worked out in Appendix 1.

[35] Gaynor, Rebitzer and Taylor (2001) document a case in which these regulations substantially weakened an HMO's incentive contracts. They also present some evidence that these weaker contracts increased the HMOs medical utilization costs.

[36] A third regulatory strategy would be to require that HMOs reveal the details of their physician incentive contracts to patients. Meaningful patient disclosure is limited by the complexity of incentive plans (Gaynor, Rebitzer, and Taylor 2001) and the limited ability of patients to understand even basic information about incentives (Miller and Horowitz 2000; and Hall, Kid, and Dugan 2000)

Analyzing Caps on Physician Incentives

In terms of our model, we can analyze the impact of caps on incentive contracts by considering the effect of a regulation that compels HMO 2 (the high incentive/low cost HMO) to move the incentive parameter, d, below its equilibrium level. Since cost minimization requires that the two physician participation constraints hold with equality—each HMO selects the highest level of d consistent with providing the promised number of physicians—forcing HMO 2 to decrease d implies that it will increase the size of its physician network to the maximum feasible given the legally mandated cap on d. Intuitively, if HMO 2 has to carry the higher medical utilization costs that come with relaxing physician incentives, it will also seek to exploit the chief benefit of lax incentives—the ability to attract more physicians to its network.

Thus we can treat incentive caps as if they were rules mandating an increase in the size of HMO 2's network to the maximum feasible under the legislatively determined level of d. Formally, let δ_2^* be the mandated size for HMO 2's network. We assume this constraint is binding—in other words, HMO 2 is forced to use weaker incentives than it would in the absence of the constraint. To get equilibrium values for q_1, δ_1, and q_2, we would like to solve the system of equations given by (17) through (20) but with (20) replaced by the constraint $\delta_2 = \delta_2^*$. Closed form solutions to this problem are too complex to present in this paper, but for policy purposes all we need are derivatives for q_1, δ_1, and q_2 with respect to δ_2^*. These derivatives turn out to be relatively simple:[37]

$$\frac{dq_1}{d\delta_2^*} = \frac{-2\beta K\left(2bq_2 - (\beta - \lambda)K\delta_2^*\right)}{b\left(2b(2X - 3q_1) - \delta_2^* K(2\beta - \lambda)\right)} < 0 \quad \textbf{(32)}$$

$$\frac{d\delta_1}{d\delta_2^*} = \frac{2bq_2 - (\beta - \lambda)K\delta_2^*}{2b(2X - 3q_1) - \delta_2^* K(2\beta - \lambda)} > 0 \quad \textbf{(33)}$$

$$\frac{dq_2}{d\delta_2^*} = \frac{-K\left((\beta - \lambda) - (2X - 3q_1) - (2\beta - \lambda)q_2\right)}{2b(2X - 3q_1) - \delta_2^* K(2\beta - \lambda)} < 0 \quad \textbf{(34)}$$

[37] In signing the derivatives, we maintain the assumption that we are in the region that generates strict product differentiation in the absence of any constraint; e.g. $\lambda \leq 0.5\beta$.

Intuitively, the positive relationship between δ_1 and δ_2^* occurs because an increase in the size of HMO 2's physician network reduces HMO 1's advantage with patients who highly value large networks. To regain its comparative advantage, HMO 1 must broaden its own network. For both HMOs, increases in their network size lead to increases in the marginal cost of serving an additional patient. Thus, imposing a binding regulation on HMO 2 has the effect of reducing the total number of members who join any HMO.

The welfare effects of capping incentives are similar to those described in Section 3 for increases in the strength of relative practice norms. Since both q_1 and q_2 are decreasing functions of δ_2^*, the number of uninsured individuals must increase as caps on physician incentives become more binding. Those who become uninsured as a result of incentive caps are, by revealed preference, made worse off by the regulations. Of the remaining insured, those with sufficiently high willingness to pay for physician choice will be made better off while others, with a smaller willingness to pay for choice, will experience a decline in utility. Even though these individuals benefit from access to larger networks, this gain is offset by having to pay increased premiums.

Changes In HMO's Legal Liability

We have so far focused on interventions that cap the intensity of physician incentives. We now consider an alternative regulatory strategy: making HMOs liable for medical malpractice linked to their cost-containment systems.

Malpractice suits have become a major cost of practicing medicine, with physicians typically paying substantial amounts of money for medical malpractice insurance. Indeed, this is why we argued above that the threat of malpractice suits is one of the factors shaping physicians' absolute and relative practice norms. In designing incentive contracts with doctors, HMOs must account for all the costs generated by serving a patient, including any costs due to malpractice. The incentive contract serves as a means of assigning these costs either to the physician or to the HMO. Under the current regulatory regime, HMOs already bear some of the expected costs of malpractice suits against physicians even though nominally all of these costs fall on physicians. Simply altering ERISA to allow plaintiffs to name HMOs as defendants in malpractice suits will, ceteris paribus, likely have *no* effect on the equilibrium practice styles that characterize HMO networks. Legally reassigning some portion of malpractice costs to the HMO doesn't generally

affect the equilibrium incentive contracts, since other costs can be reassigned to the physicians to yield the original incentive structure.

ERISA reforms *will* have an effect, however, if the naming of HMOs as defendants has influences the magnitude of damages awarded by juries. A recent series of papers examining jury behavior have found that punitive damage awards are influenced not only by the facts of the case but also by the identity of the defendants. Large organizations with deep pockets are typically hit with higher punitive damage awards than smaller organizations (Kahneman, Schkade, and Sunstein 1998). If, by virtue of their size and deep pockets, HMO defendants in medical malpractice suits incur larger punitive damage awards than physician defendants do, we can expect changes in ERISA to affect the behavior of HMOs. In the following analysis we discuss two channels through which HMO malpractice liability can influence equilibrium outcomes: (1) if increased jury awards increase fixed costs per patient; and (2) if increased jury awards raise the incidence of various forms of "defensive" medical practices. As we will discuss below, "defensive medicine" can be represented in our model in three different ways. In the interest of brevity, we state the likely effect of these changes without providing formal proof (see Table 1 for a summary). To the extent that these results don't follow directly from what has been shown above, proofs are available from the authors upon request.

Increases in Fixed Cost per Patient

By assumption, we have set the fixed cost of serving an additional patient equal to zero. We can expect, however, that changes in ERISA leading to higher malpractice damages will lead to an increase in the fixed cost of taking on a patient. Technically, fixed costs appear as a constant on the right hand side of (17) and (19), the first order conditions for HMO 1 and 2 with respect to quantity. It can be shown that an increase in this fixed cost increases network size and decreases quantity of patients served for HMO 2. The intuition behind this result is straight forward—because quantity has become relatively more expensive, the HMO substitutes from quantity towards larger network size. For HMO 1 the results are reversed, with quantity rising

and network size falling.[38] Although the number of members of HMO 1 increases, the *overall* effect of the increase in fixed costs is a decrease in the total number of patients insured by HMOs. This follows because the decrease in quantity for HMO 2 is larger than the increase for HMO 1. As we discussed above, these newly uninsured are *always* made worse off by the change. The welfare effects for those individuals remaining insured is ambiguous.

Increase in the Incidence of Defensive Medicine

Increased malpractice damages due to changes in ERISA can be expected to make "defensive medicine" more attractive as a means of containing legal costs. By defensive medicine we mean the spending of resources on medical care that has little clinical value, but helps limit liability in the event of a law suit. In terms of our model, an increase in defensive medical practices can be represented in three ways: by a decrease in the effect that costly practice styles have on costs (β); by an increase in the minimum acceptable practice style, α; and/or by an increase in the cost of practicing a relatively inexpensive style of practice, λ. We discuss each of these in turn.

We have already indicated that part of the relationship between choice of practice style, s, and medical costs is due to the expected costs of malpractice suits. If the expected awards in malpractice suits increase, then costly defensive medical practices can actually reduce the marginal cost of adopting these expensive practice styles. In terms of our model, this means that β falls as expected jury awards increase.[39] This parameter change causes the equilibrium quantity of patients insured to (weakly) decrease for both HMOs and network size to increase for both HMOs. Intuitively, the decrease in β makes expensive practice styles more attractive for HMOs. This effect is sufficiently strong to overwhelm the direct effect on quantity from decreasing costs per patient (due to decreasing β).

---

[38] This may seem surprising, but note that the direct effect of increasing fixed costs per patient are smaller for HMO 1 than for HMO 2 since the former serves a relatively small number of patients. This cost advantage for HMO 1 is sufficiently large to reverse the direct effects due to an increase in fixed costs.

[39] Medical costs can be treated as having two components, direct costs of medical care and malpractice damages. An increase in practice style can be assumed to increase the direct costs but decrease the malpractice damages. The larger, certeris paribus, malpractice damages are, the more important the second component becomes. If shifting responsibility for malpractice damages onto HMOs increases the size of these damages, the marginal cost of increased practice style is lowered by the greater savings generated from avoiding malpractice claims.

A second plausible effect of changes in ERISA that increase the size of malpractice awards is to make HMOs desire a more generous practice style than physicians would otherwise choose in equilibrium. To induce more generous practice styles, HMOs need to weaken the incentive intensity of their physician contracts with resulting effects on network sizes and quantities of patients served. Formally, this change is equivalent to a cap on incentive parameter *d* in the low-cost HMO. We know the effects of such a change from the preceding section: both HMOs in the market will adopt lower-powered incentives and build larger physician networks. As a result, each HMO will attract fewer members and the number of uninsured will rise.

The third channel by which changes in ERISA could influence HMO management is via the *relative* generosity of the practice style prevailing in the network. Consider a malpractice case in which a doctor is charged with malpractice for treating a patient in HMO 2 of our model. Imagine that the question at issue is that the physician did not recommend a controversial and expensive procedure that physicians in HMO 2's network do not offer but which physicians in HMO 1's network routinely offer. If the jury finds in favor of the plaintiff, Kahneman et. al's results suggest that the damage awards will be greater if the HMO and the physician are plaintiffs than if the physician alone were the plaintiff. From this motivating example, it is not hard to imagine that changes in ERISA can be especially hard on the low-cost HMO in a market. By increasing the costs of using a practice style below that of its competitor, this policy change puts pressure on HMO 2 to weaken its cost-control incentives. We can capture this effect in our model as an increase in λ, the cost of adopting a *relatively* inexpensive practice style. From (26)–(29), we know what the effects of an increase in λ will be. The low-cost HMO will reduce the incentive intensity of its physician contracts and increase the size of its physician network. HMO 1 will respond to HMO 2's entry into its market segment by doing the same thing: reducing incentives and increasing network size. The net result is that membership in both HMOs fall while the number of uninsured increase.

To sum up, an ERISA induced increase in defensive medicine (seen as decreases in β; increases in δ^* and λ), will increase the number of uninsured. We know by revealed preference that individuals who become uninsured must be made worse off. For individuals who remain insured, those with high willingness to pay for access to large physician networks will benefit while those who are more price sensitive will be harmed.

Discussion of Policy Results

We have considered the likely effects of two broad strategies for limiting an HMO's ability to impose cost containment incentives on physicians. Our results indicate that both of these strategies will have the likely effect of increasing medical costs and increasing the number of uninsured. The newly uninsured are always made worse off by these interventions. The welfare effects on those remaining insured are murky—with some consumers being made worse off and others better off.

Should we conclude that regulating physician incentive systems is a bad policy idea? Not necessarily. There are important externalities to health care that may not be fully captured in the factors that constrain HMO behavior: physician practice norms, customer preferences for physician choice, and expected awards for malpractice. Some of these externalities involve health outcomes—if more expensive treatments are also more effective then financing these treatments may improve the welfare of care givers and family members who are not directly involved in the purchase of the health care insurance. Other externalities involve the physician-patient relationship—the use of high-powered financial incentives anywhere may undermine patients' willingness to trust in and listen to their doctors' advice. If the net social value of more expensive practice styles or restrictions on contracting exceeds their private value, there remains a strong case for interventions limiting the ways that HMO's regulate care.

The lesson of our analysis is *not* that policy interventions are necessarily a bad idea, but rather that they must be undertaken with an understanding of their cost—especially the increase in the number of uninsured.[40] Policies regulating HMO incentive systems will be more effective if implemented in conjunction with policies that increase access to care for the uninsured.

[40] Ultimately, the magnitude of the increase in uninsured is an empirical question that cannot be assessed without calibrating our model to real world parameters. Studdert et. al. (1999) suggest the increase in costs from increased HMO liability is quite uncertain.

CONCLUSION

In this paper, we have studied the interaction between the incentive contracts that Health Maintenance Organizations (HMOs) write with physicians and product market competition among HMOs. The central economic and policy question is whether, in an environment where consumers cannot assess the quality of care they receive, competition among HMOs will lead to a "race to the bottom" in terms of care quality. The results of our model suggest that, even in a setting with extreme asymmetric information, this concern may be unfounded. HMOs are constrained from "racing to the bottom" by an interaction between absolute physician practice norms and product market competition. In fact, relative practice norms may act as a force "pulling to the top."

We also consider the implications of our model for public policy designed to limit the intensity of HMO incentive systems. We analyze two such policies: caps on "at risk" physician income and legal changes that make HMOs directly liable for damages caused by their incentive systems. Our results suggest that medical costs and premia will rise as a result of these policies causing the number of uninsured to increase. The newly uninsured are clearly made worse off by these interventions while the welfare effect on those remaining in the insurance system is complex, with some consumers gaining and others losing. It follows that policies aimed at regulating HMO incentive systems could be made more efficacious if they also involved actions to improve access to health care for the uninsured.

Our model highlights a number of still unresolved empirical and theoretical issues that are worthy of further study. Of these, perhaps the most important is understanding the determinants and scope of physician practice norms. At present these norms are only dimly understood by economists. Further research into the formation and operation of norms is critical for understanding the impact of incentives in healthcare and in the many other settings where critical outcomes hinge on the decisions and actions of highly skilled professionals.

Table 1:

Sign of Predicted Effects of Making HMOs Liable in Malpractice Suits

"+" Positive Predicted Effect "–" Negative Predicted Effect

Possible Effect of Shifting Liability to HMOs	**Fixed Cost Per Patient Increases**	**Increase in the Incidence of "Defensive Medicine" Due to Higher Malpractice Awards.**		
Representation of Policy Change in Formal Model	Fixed Cost$\uparrow$	$\beta\downarrow$	$\delta_2^*\uparrow$	$\lambda\uparrow$
Insured by HMO 1 (q_1)	–	–	+	–
HMO 1 Network Size (d_1)	+	+	–	+
Insured by HMO 2 (q_2)	–	–	–	–
HMO 2 Network Size (d_2)	+	+	+	+
Total Insured (q_1+q_2)	–	–	–	–

REFERENCES

Akerlof, George A., and Rachel E. Kranton. 2000. "Economics and Identity." *Quarterly Journal of Economics* 115: 3: 715–53.

Altman, Daniel, David M. Cutler, and Richard J. Zeckhauser. 2000. "Enrollee Mix, Treatment Intensity, and Cost in Competing Indemnity and HMO Plans." NBER Working Paper. No. 7832. August.

Barro, Jason R., and Nancy Dean Beaulieu. 2000. "Selection and Improvement: Physician Responses to Financial Incentives." Harvard Business School Working Paper. July.

Beaulieu, Nancy Dean. 2000. "Externalities in Overlapping Supplier Networks." Harvard University Working Paper. November.

Cournot, A. 1838. *Recherches sur les Principes Mathématiques de la Théorie de Richesess.* English edition (ed. N. Bacon). 1897. *Researches into the Mathematical Principles of the Theory of Wealth.* New York: MacMillan.

Cutler, David M., Mark McClellan, and Joseph P. Newhouse. 2000. "How Does Managed Care Do It?" *Rand Journal of Economics* 31: 3: 526–548.

Encinosa, W., M. Gaynor, and J. Rebitzer. 2001. "The Sociology of Groups and the Economics of Incentives: Theory and Evidence on Incentives in Medical Groups." Working Paper. Case Western Reserve University. August.

Fox, Peter D. 1997. "An Overview of Managed Care." In Peter R. Konstvedt, ed., *Essentials of Managed Health Care.* Gatihersburg, Maryland: Aspen Publishers, Inc.

Gal-Or, Esther. 1985. "Differentiated Industries Without Entry Barriers." *Journal of Economic Theory* 37: 310–339.

Gaynor, Martin, James B. Rebitzer, and Lowell J. Taylor. 2001. "Incentives in HMOs." No. 85222. October.

Gibbons, Robert, and Michael Waldman. 1999. "Careers in Organizations: Theory and Evidence." In Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics 3B.* Oxford: Elsevier Science, North-Holland. 2373–2437.

Gosfield, Alice G. 1997. "Who is Holding Whom Accountable for Quality?" *Health Affairs* 16: 3: 26–40.

Grumbach, K., D. Osmond, K. Vranizan, D. Jaffe, and A. Bindman. 1998. "Primary Care Physicians' Experience of Financial Incentives in Managed-Care Systems." *New England Journal of Medicine* 339: 21: 1516–21.

Hall, Mark A., Kristin E. Kidd, and Elizabeth Dugan. 2000. "Disclosure of Physician Incentives: Do Practices Satisfy Purposes?" *Health Affairs* 19: 4: 156–64.

Havighurst, Clark C. 2000. “American Health Care and The Law—We Need to Talk!” *Health Affairs* 19: 4: 84–106.

InterStudy. 1999. “The InterStudy Competitive Edge, Part II: The HMO Industry.” *Interstudy Publications.* Report 9.1.

Kahneman, Daniel, David Schkade, and Cass R. Sunstein. 1998. “Shared Outrage and Erratic Awards: The Psychology of Punitive Damages.” *Journal of Risk and Uncertainty:*1–53.

Kandel, Eugene and Edward P. Lazear. 1992. “Peer Pressure and Partnerships.” *Journal of Political Economy* 100: 4: 801–817.

Kessler, Daniel and Mark McClellan. 1996. “Do Doctors Practice Defensive Medicine?” *Quarterly Journal of Economics* 111: 2: 353–90.

Ma, Albert Ching-to and Thomas G. McGuire. Forthcoming. “Network Incentives in Managed Care.” *Journal of Economics and Management Strategy.*

March, James G. 1994. *A Primer on Decision Making: How Decisions Happen.* New York: The Free Press.

McGuire, Thomas G. 2000. “Physician Agency.” In Anthonly J. Culver and Joseph P. Newhouse, eds., *Handbook of Health Economic* 1A. Amsterdam: Elsevier.

Miller, Tracy E. and Carol R. Horowitz. 2000. “Disclosing Doctors' Incentive: Will Consumers Understand And Value the Information?” *Health Affairs* 19: 4: 149–55.

Phelps, Charles E. 1992. “Diffusion of Information in Medical Care.” *Journal of Economic Perspectives* 6: 3: 23–42.

Prendergast, Canice. 1999. “The Provision of Incentives in Firms.” *Journal Of Economic Literature* 37: 1: 1–63.

Robinson, J. C. 2001. “Theory and Practice in the Design of Physician Payment Incentives.” *Milbank Quarterly* 79: 2: 149–77.

Robinson, James C. 1999. *The Corporate Practice of Medicine: Competition and Innovation in Health Care.* Berkeley, Calif.: University of California Press.

Studdert, David M., William Sage, Carole Roan Gresenz, and Deborah R. Hensler. 1999. “Expanded Managed Care Liability: What Impact on Employer Coverage?” *Health Affairs* 18: 6: 7–27.

Appendix 1: Relative Norms and Consumer Welfare: We consider the effect of an increase in the strength of relative norms on consumer's welfare. Differentiating the consumer's utility function (12) we get the following:

$$\frac{du}{d\lambda} = bx\left(\frac{d\delta}{d\lambda}\right) - \left(\frac{dp}{d\lambda}\right) \qquad \textbf{(A1)}$$

We have already shown that $d\delta/d\lambda > 0$ for members of either HMO and that $dp/d\lambda$ is independent of *x.* From this it follows that $du/d\lambda$ is an increasing function of x. We can prove that $du/d\lambda > 0$ for a segment of the market by identifying the consumer with the minimum value of *x* for which $du/d\lambda > 0$.[41]

To determine the welfare effects of increasing λ for consumers originally in HMO 1, we evaluate $du/d\lambda$ for the consumer in HMO 1 with the weakest willingness to pay for physician choice. This marginal consumer has $x = X - q_1$ and it is easy to show the following:[42]

$$\left.\frac{du}{d\lambda}\right|_{x=X-q_1} = \frac{9b^2X^2\left(26\beta^3\lambda - 39\beta^2\lambda^2 + 15\lambda\beta^3 - \lambda^4\right)}{\left(23\beta^2 - 23\beta\lambda - \lambda^2\right)^3 K^2} > 0 \qquad \textbf{(A2)}$$

Therefore, an increase in relative practice norms benefits all consumers with a willingness to pay strong enough to choose HMO 1.

The story is more complex for consumers that initially choose HMO 2. We know that some consumers in HMO 2, those with relatively low values of x, are made worse off by increased λ ($du/d\lambda < 0$) because some of them choose to become uninsured. Similarly, we know from (A2) that some consumers in HMO 2 with relatively high values of x are made better off by

[41] On the margin, the worst off consumer in an HMO either switches HMOs or moves to being uninsured as λ increases. For our calculations, we treat this consumer as if she stayed with the same HMO. If this consumer would have been better off staying with the same HMO than she was prior to the change in λ (although not as well off as by switching), all consumers in the same HMO with higher values of x must also be made better off by the increase in λ .

[42] This derivative is positive if we maintain the necessary and sufficient condition for product differentiation ($0 < \lambda < 0.5\beta$).

increased λ ($du/d\lambda>0$).[43] Since $du/d\lambda$ is a monotonic function of x, there must exist some value of x such that $du/d\lambda\big|_x = 0$.

Rather than solving for this breakpoint directly, it is simpler to solve for the proportion of consumers using HMO 2 who are made worse off by an increase in λ. Define the function $\hat{x}(x)$ as follows: $\hat{x}(x) \equiv x - (X - q_1 - q_2)$. This function gives the number of patients in HMO 2 whose willingness to pay for larger networks is less than x. Combining this definition with the utility function given by (12) and the price function for HMO 2 given by (13), we can rewrite the consumer's utility function as $u(\delta_2, \hat{x}) = b\delta_2 \hat{x}$ for any consumer in HMO 2. Let $\hat{x}^*$ be the value of $\hat{x}$ where consumers are indifferent about an increase in λ.. Differentiating u by λ and solving for $du/d\lambda = 0$ yields (A3).

$$\hat{x}^* = \left(\frac{\frac{d(q_1 + q_2)}{d\lambda}}{\frac{d\delta_2}{d\lambda}} \right) = \delta_2 \left(\frac{3\beta(2\beta - \lambda)(23\beta^2 - 38\beta\lambda + 20\lambda^2)}{(23\beta^2 + 4\beta\lambda - \lambda^2)(23\beta^2 - 23\beta\lambda - \lambda^2)} \right) X \qquad \textbf{(A3)}$$

Dividing $\hat{x}^*$ by q_2, we can solve for the proportion of consumers in HMO 2 who are harmed by an increase in relative practice norms:

$$\frac{\hat{x}^*}{q_2} = \left(\frac{\beta(23\beta^2 - 38\beta\lambda + 20\lambda^2)}{(\beta - \lambda)(23\beta^2 + 4\beta\lambda - \lambda^2)} \right) \textbf{(A4)}$$

This ratio is positive as long as the condition for product differentiation ($\lambda < 0.5\beta$) holds. Thus some positive proportion of HMO 2 members are hurt by an increase in relative practice norms.

[43] This follows because some consumers in HMO 1 switch to HMO 2 when λ increases, even though their utility would increase if they stayed in HMO 1. It follows that HMO 2 consumers with values of *x* close to those who switch from HMO 1 to HMO 2 will also experience an increase in utility.