DEMAND INDUCEMENT AND THE PHYSICIAN/PATIENT RELATIONSHIP

DAVID DRANOVE*

The physician/patient relationship is a paradigm for any expert/client relationship. The physician both diagnoses the patient's illness and recommends a treatment. This dual role gives the physician incentive to recommend treatments whose costs outweigh their medical benefits. These socially inefficient treatments correspond to the notion of "physician-induced demand." The level of inducement chosen by the physician is shown to depend on the price and potential medical benefits of treatment and the relative diagnostic skills of physician and patient. This model offers several testable hypotheses, some of which are confirmed by related studies.

I. INTRODUCTION

An important characteristic of the physician/patient relationship is that the physician and the patient jointly determine what treatments will be performed. The theory of "physician-induced demand" suggests that the physician, as joint decision maker, can influence the demand for her own services. The theory was embraced in the 1970s by several authors, including Anderson et al. [1981], Evans [1974], Fuchs [1978], and Fuchs and Kramer [1973], but none proposed a theoretical model explicitly to describe inducement. Recently, authors such as Pauly [1980], Coyte [1982], and Wilensky and Rossiter [1983] have begun to explore the individual decisions that lead to inducement, but the precise causes of inducement, the factors which may limit inducement, and the ways in which market forces may affect inducement remain largely unexplained. In addition, models of demand inducement have yielded few testable hypotheses. This paper addresses all of these topics by providing a testable model of rational economic behavior by physician and patient in which inducement can be understood.

The phenomenon of demand inducement need not be confined to the physician/patient relationship. This paper is really about any situation in which an expert both advises a client about needed services and offers to supply the recommended services for a fee. Auto mechanics, plumbers, salespeople, and their clients, all face the incentive problems posed here. Models of such problems of information transmission have been called "games of persuasion."[1]

* Associate Professor, University of Chicago Graduate School of Business. I am grateful to Tim Bresnahan, Jeremy Bulow, Alain Enthoven, John Roberts, and Mark Satterthwaite for their suggestions. They are absolved from any responsibility for errors that may appear.

[1] See Green and Stokey [1980], Milgrom [1981], Crawford and Sobel [1982], and Pitchik and Schotter [1984] for various models of games of persuasion.

These models conclude that utility-maximizing experts may or may not use their information in a socially desirable way, depending on the specifics of the particular game. This paper builds on the game theoretic structure of these models to explore, in a stylized way, the game between the physician and the patient. The model pays close attention to those exogenous factors that will influence their strategic choices to shed light on the nature of the inducement problem.

In this model the physician induces demand by recommending treatment even though clinical indications suggest that the costs exceed the benefits. If the patient consents, the physician performs the treatment and receives a fee. However, if the physician has a reputation for being an "overprescriber," the patient may not consent. The physician determines the optimal level of inducement by balancing the gains from recommending and performing expensive treatments and the losses from possessing a "bad" reputation. The level of inducement she chooses will depend on the price of the treatment, the value of treatment, the relative diagnostic skills of the physician and patient, and the expected duration of the physician/patient relationship. The supply of physicians is shown to affect the level of inducement indirectly, through the effect of supply on price.

The idea that a physician can use her superior information to her advantage first attracted attention in the seventies. This argument was first presented formally by Evans [1974] who suggested that there is some underlying demand for physicians' services which the physician shifts out through unspecified "inducing activities." This argument was repeated by Fuchs [1978]. Their models, which resemble simple advertising models, do not make much sense unless there is a limit to the shift in demand, or else physicians could be infinitely wealthy.[2] One limit might be patient credulity—if physicians continually induce demand then patients might no longer believe that treatment is worthwhile.[3] The models also fail to provide a micro-theoretic exposition of the nature of inducing activities.

More recent discussions of inducement benefit from careful descriptions of inducing activities, and precise analyses of the informational asymmetries. Wilensky and Rossiter [1983] draw a parallel between "patient-initiated" versus "physician-initiated" visits on the one hand, and "regular demand" versus "induced demand" on the other. They use survey data on the relative numbers of patient and physician-initiated visits to test for the magnitude of inducement, with ambiguous results. Pauly [1980] uses an agency approach to define the informational problem confronting patients. Patients have imperfect information about the value of treatment and must rely on diagnostic information provided by the physician. The physician's problem is to choose the accuracy of the information he provides to patients. Patients are Bayesians and use past

[2] Reinhardt [1985] presents an interesting discussion of how economists have traditionally tackled this problem.

[3] McCarthy [1985] finds evidence that there are, indeed, market constraints on the demand facing individual physicians.

experience to judge the degree of physician accuracy. Pauly finds that physicians with utility for income and leisure will choose to be partially inaccurate. Pauly's model demonstrates that physicians may use their private information to their own advantage, even if consumers have information that this demand creation is occurring. Pauly's model has few empirical implications, however, suggesting that unless physicians have benevolent motives, the level of accuracy may be invariant to exogenous supply and demand conditions.

The model in this paper draws the same distinction between patient and physician-initiated visits as do Wilensky and Rossiter. The model is also similar to Pauly's in that patients rationally combine information provided by the physician with their own private information to estimate the value of treatment. This paper more thoroughly investigates the nature of the informational asymmetry in this particular "game of persuasion" than do these authors, and shows how the level of inducement depends, in predictable ways, on several key factors such as the price of treatment, the nature of the illness and the diagnostic skills of the patient.

As a historical note, the inducement hypothesis also attracted attention as a possible means of explaining two alleged anomalies in the data on the demand for physician's services. Earlier researchers, including Evans [1974], Feldstein [1970] and Fuchs [1978], seemed to suggest that demand was increasing in price and that price was increasing in the supply of physicians. It turns out that within the context of utility-maximization models, demand inducement is not sufficient to explain these alleged anomalies. However, recent work has called into question the methods and conclusions of this early inducement-related research.

Ramsey and Wasow [1981] point out serious flaws in the models and empirical specifications used by these early researchers, stating that no theoretical conclusions about the demand for medical care can be drawn from this literature. Satterthwaite [1979; 1982] and Pauly and Satterthwaite [1981] further show that the latter of the suggested "anomalies" may be a natural consequence of utility maximizing behavior when demand elasticities differ across markets. Auster and Oaxaca [1981] consider the methodology used to test the "shifting demand curve" hypothesis. They conclude that the data requirements of such a test are so severe as to render it impossible. More recent inducement models (e.g., Pauly, Wilensky and Rossiter, Coyte and this model) have been sensitive to this criticism, and develop alternative methodologies for testing the inducement hypothesis.

The paper is organized as follows. Sections II through IV present a simple model of the medical process and describe how the physician can induce demand. Factors that affect the level of inducement are considered in section V. Section VI explores the effects of competition on demand inducement. Section VII proposes several tests of the model. Section VIII concludes.

II. A STYLIZED MEDICAL PROCESS

This section presents a model of the medical process that will be used to describe demand inducement. Assume initially that the physician is a monopolist. It may help to think of her as the only physician in town. Suppose only one kind of illness exists and the physician can perform only one kind of treatment. Let the objective value of treatment be an increasing function of some unobservable, underlying variable Z. Z is distributed in the population according to the density $\phi(Z)$. It may help to think of Z as a measure of the severity of the illness. Neither the patient nor the physician knows Z. Instead, each observes symptoms which are estimates of Z, enabling them to diagnose the illness. Let the doctor's and patient's estimates of Z be S_d and S_p respectively. S_d may be thought of as summarizing all of the clinical information obtained by the physician, including lab tests, radiology reports, etc. S_p reflects the symptoms observable to the patient, including pain, discomfort and inconvenience. For any severity Z, the probability that the physician will observe S_d is given by the density $f(S_d \mid Z)$. The probability that the patient will observe S_p is given by the density $g(S_p \mid Z)$. The idea captured by these densities is that the patient and physician have private information about the severity of the illness, causing each to estimate the value of treatment independently.

Upon observing his estimate of severity, S_p, the patient must decide whether or not to visit the physician. For simplicity, assume that the patient only pays a fee when he receives treatment. The patient receives treatment if and only if the physician recommends it and he consents. Patients are assumed to be rational, making use of all available information before deciding to consent. In particular, the physician's treatment recommendations reveal information about her estimate of severity, S_d. This paper examines the case in which this is the only way that a physician can reveal information about S_d.[4] The patient uses the information revealed by the physician to update his own estimate of Z and then makes his consent decision accordingly.

In this formulation, patients unwilling to consent to treatment will have no reason to visit the physician. It may be more appropriate to assume that the physician can perform more than one type of treatment (e.g., write a prescription versus operate) and induces demand by recommending the more costly treatment. This addition would not alter the principal results developed here.[5]

III. THE PATIENT'S PROBLEM

Each patient has utility for health and wealth. Assume that utility, U^p, is additively separable:

$$U^p = M + \mu(E - P).$$

M = health status (detailed below).
$\mu(\ )$ = utility for wealth. Assume that $\mu' > 0$ and $\mu'' \leq 0$.

[4] In reality, the physician can and does reveal some limited diagnostic information (such as the patient's blood pressure). The main concern of this paper, however, is the information conveyed by the treatment recommendations. This is the information that is affected by demand inducement.

[5] Dranove [1983] derives a much more general version of this model, with essentially the same results.

E = initial endowment of wealth.
P = price of treatment.

The patient's health status, M, depends on the severity of his illness, Z, and whether or not he receives treatment. The medical value of treatment is $\pi(Z)$. This includes any time and inconvenience costs of treatment, and may be negative. Z is defined so that $\pi(Z)$ is an increasing function of Z. The health status of a patient with illness of severity Z is as follows: $M = -\pi(Z)$ if the patient does not receive treatment; $M = 0$ if the patient receives treatment.[6] Each patient has an initial wealth endowment of E. If he receives treatment he pays a fee of P, leaving him with wealth of $E - P$. Thus, if a patient with illness of severity Z does not receive treatment his utility is

$$-\pi(Z) + \mu(E). \tag{1}$$

If a patient with illness of severity Z receives treatment his utility is

$$\mu(E - P). \tag{2}$$

The patient who falls ill must decide whether or not to visit the physician, and whether or not to consent to treatment. Let $j(Z; S_p)$ *denote the patient's posterior (revised) density over the possible values of* Z, conditional on the patient having observed S_p and the physician having recommended treatment. The patient estimates that if he does not receive treatment his expected utility will be

$$-\int_{\infty}^{\infty} \pi(Z)j(Z; S_p)\, dZ + \mu(E). \tag{3}$$

A comparison of (2) with (3) indicates that an expected utility maximizer will consent to treatment if and only if

$$-\int_{-\infty}^{\infty} \pi(Z)j(Z; S_p)\, dZ + \mu(E) \leq \mu(E - P). \tag{4}$$

The value of S_p *which satisfies expression (4) (with equality) is called the patient's cutoff and is denoted by* $\bar{S}_p$. If the density $j(Z; S_p)$ has the Monotone Likelihood Ratio Property (defined in the appendix), then the left-hand side of (4) is decreasing in S_p.[7] Intuitively, as S_p increases the patient believes the illness to be more severe. As a consequence, the patient is more willing to consent. Therefore, the patient consents if his symptoms are as serious or more serious than symptom level $\bar{S}_p$.

To further analyze the patient's behavior we must specify how he processes

[6] This normalization makes the mathematics less cumbersome. One obtains the same results if health status is a function of Z.

[7] This is easily proved using the properties described in the appendix. If j has the MLRP, and $S_p^* > S_p'$, then $j(Z; S_p^*)$ is first-order stochastic dominant over $j(Z; S_p')$. Because $\pi(Z)$ is non-decreasing in Z, $\int_{-\infty}^{\infty} \pi(Z)j(Z; S_p^*)\, dZ > \int_{-\infty}^{\infty} \pi(Z)j(Z; S_p')\, dZ$.

the information conveyed by a physician who recommends treatment. The patient must infer from the recommendation the seriousness of his illness. The patient bases this inference on his beliefs about the recommendation strategy of the physician. For example, if he believes that the physician only prescribes treatment when absolutely necessary, the patient would correctly infer that treatment is justified. The patient would give less credence to the recommendation if the physician is known to prescribe treatment often. Thus, the same recommendation can lead to different inferences, depending on the physician's reputation.

When there is only one physician, her recommendation strategy may be reasonably well known. The patient can draw on many sources, such as his own medical history and the histories of others to determine the physician's strategy. A useful model must employ an equilibrium concept that recognizes that the patient can make use of available information to form a "sensible" conjecture of the physician's strategy. Initially, I will make the strong assumption that patients always know the physician's recommendation strategy.[8] This assumption serves three purposes. First, it allows us to show that even under the most generous conditions of patient informedness, the physician has incentive to recommend unnecessary treatments. Second, it serves as a useful benchmark for considering how the physician's incentives change when patients have only limited information about her strategy. Third, it stands as a reasonable assumption if the medical profession enforces an industry-wide recommendation strategy, a possibility discussed in section VI.

We can formally define the equilibrium as follows. Let R denote the set of symptoms for which the physician recommends treatment. Let C denote the set of symptoms for which the patient is willing to consent. Holding price constant, let $V^d(P, R, C)$ denote the physician's indirect utility when price P and strategies R and C are chosen. $V^p(P, R, C)$ is the patient's indirect utility. In equilibrium, the following conditions hold.

1. For any P and R chosen by the physician, the patient chooses C to maximize $V^p(P, R, C)$.
2. The physician chooses P and R to maximize $V^d(P, R, C)$ given that patients behave according to condition 1.

Condition 1 states that patients behave rationally, choosing the consent rule which maximizes their utility. Condition 2 states that the physician chooses her price and recommendation strategy to maximize her utility, bearing in mind that her choices affect the patient's consent rule. This means that the physician acts as a Stackelberg leader.

We have already shown that the patient's consent strategy is a cutoff rule, consenting only if $S_p \geq \bar{S}_p$. It can be shown that a utility-maximizing physician (with income/leisure preferences) also uses a cutoff rule, recommending treat-

[8] In this equilibrium, it is assumed that the patients know the incentive guiding the behavior of the physician and can correctly anticipate her choice of recommendation strategy.

ment if and only if $S_d \geq \bar{S}_d$.[9] This gives us a natural description of demand inducement. The physician induces demand by lowering her cutoff $\bar{S}_d$, thereby recommending treatments for progressively less severe clinical indications.

If the physician recommends treatment, then the patient knows that $S_d \geq \bar{S}_d$. The patient uses Bayes' rule to combine this information with his own observation S_p to form $j(Z; S_p)$. He then solves equation (4) to determine his cutoff, $\bar{S}_p$. It is straightforward to show that $\partial\bar{S}_p/\partial\bar{S}_d < 0$. That is, if the physician lowers her cutoff, patients raise theirs. This is in accord with our notion that the physician who prescribes too many costly treatments will establish a "bad" reputation, evoking skepticism on the part of her patients. When choosing her cutoff, the physician must be cognizant of this reputation effect. This is more clearly established in the next section.

IV. THE UTILITY-MAXIMIZING PHYSICIAN'S PROBLEM

This section develops the first-order conditions for the utility-maximizing physician's choices of price and cutoff. The physician has utility for income, Y, and disutility for work, W. Her utility function is

$$U^d = U^d(Y, W) \quad \text{where} \quad U_y > 0 \quad \text{and} \quad U_W < 0.$$

The physician's income and workload depend on the proportion of patients who receive treatment, denoted by ρ. ρ is a function of the physician's and patient's cutoffs:

$$\rho = \int_{-\infty}^{\infty} \int_{\bar{S}_d}^{\infty} \int_{\bar{S}_p}^{\infty} g(S_p \mid Z) f(S_d \mid Z) \phi(Z) \, dS_d \, dS_p \, dZ.^{10} \tag{5}$$

Because $\bar{S}_p$ is a function of P and $\bar{S}_d$, ρ is a function of P and $\bar{S}_d$, $\rho(P, \bar{S}_d)$.

Suppose that each treatment takes one unit of time. Diagnoses are assumed to be costless and take zero time. The implication of this latter assumption will be made clear shortly. If the population size is Q, the physician's income is $Q\rho P$ and her workload is $Q\rho$. Her utility is

$$U^d = U[Q\rho(P, \bar{S}_d)P, Q\rho(P, \bar{S}_d)].$$

The physician chooses P and $\bar{S}_d$ to maximize U^d. The first-order conditions for utility maximization are

$$(1 + 1/\epsilon)P = -(U_W/U_Y) \quad \text{where} \quad \epsilon = (\partial\rho/\partial P)(P/\rho), \tag{6a}$$

and

$$(PU_Y + U_W)(\partial\rho/\partial\bar{S}_d) = 0. \tag{6b}$$

[9] The proof of this argument turns out to be fairly long and tedious, without adding to further understanding; see Dranove [1983].

[10] Equation (5) gives the expected value of ρ. Assume that the physician treats sufficiently many patients so that the expected value of ρ roughly equals the actual value. Thus, her expected utility will equal her actual utility.

Equation (6a) states that the marginal revenue from an additional treatment must equal the marginal cost. Equation (6a) also implies that unless demand is perfectly elastic (unless $\epsilon = -\infty$), $PU_Y > -U_W$. We conclude from (6b) that $\partial \rho/\partial \bar{S}_d = 0$. The physician's cutoff maximizes the number of treatments she performs at the simultaneously chosen price. This important result has a strong intuition behind it. If $\partial \rho/\partial \bar{S}_d \neq 0$, the physician can perform additional treatments by changing her cutoff. The marginal revenue from changing the cutoff equals the price, and therefore exceeds marginal cost. Consequently, the physician's utility increases. This result is consistent with Pauly's [1980] observation that a utility-maximizing physician induces demand "to the maximum." As I will show, however, this does not imply that the physician always chooses the same cutoff, independent of market conditions.

The result that the physician induces to the maximum depends on the assumption that the diagnosis is free. We have already argued that as the physician lowers her cutoff, fewer patients will be willing to consent. These patients will not purchase diagnoses if the physician charges for them. The physician's loss from doing fewer diagnoses will offset somewhat any gain from doing more treatments. This would be further disincentive to demand inducement, beyond those discussed below.

The physician, in choosing $\bar{S}_d$, is interested in knowing how a change in this cutoff will affect her workload. This is given by (7).

$$\frac{\partial \rho}{\partial \bar{S}_d} = -\int_{-\infty}^{\infty} \int_{\bar{S}_p}^{\infty} g(\bar{S}_p | Z)\, \mathrm{d}S_d\, f(\bar{S}_d | Z) \phi(Z)\, dZ$$

$$- \int_{-\infty}^{\infty} \int_{\bar{S}_d}^{\infty} f(\bar{S}_d | Z)\, \mathrm{d}S_d g(\bar{S}_p | Z)(\mathrm{d}\bar{S}_p/\mathrm{d}\bar{S}_d)\, \mathrm{d}Z. \qquad (7)$$

To develop comparative statics it is useful to express $\bar{S}_p$ as a function of $\bar{S}_d$ and a demand shifter, $X : \bar{S}_p = \bar{S}_p(\bar{S}_d, X)$, where $\partial \bar{S}_p/\partial X < 0$. Totally differentiating (7), allowing $\bar{S}_d$ and X to vary yields

1. $\mathrm{d}\bar{S}_d/\mathrm{d}X < 0$. Ceteris paribus, if an exogenous factor causes patients to lower their cutoff, the physician will lower his cutoff.

Totally differentiating (7), allowing $\bar{S}_d$ and the slope of the response function $\partial \bar{S}_p/\partial \bar{S}_d$ to vary, yields

2. $\mathrm{d}\bar{S}_d/\mathrm{d}(\partial \bar{S}_p/\partial \bar{S}_d) > 0$. If $\partial \bar{S}_p/\partial \bar{S}_d$ becomes bigger in magnitude (more negative), the physician will raise her cutoff and induce less demand. We can interpret this as meaning that as patients become more resistant to inducement, the physician induces less demand (where bigger values of $\partial \bar{S}_p/\partial \bar{S}_d$ connote greater resistance).

These comparative statics results will be useful in determining how various characteristics of treatment affect the level of inducement.

Equations (5) and (7) and the subsequent comparative statics results are best interpreted with the help of Figure 1. The horizontal axis is the range of

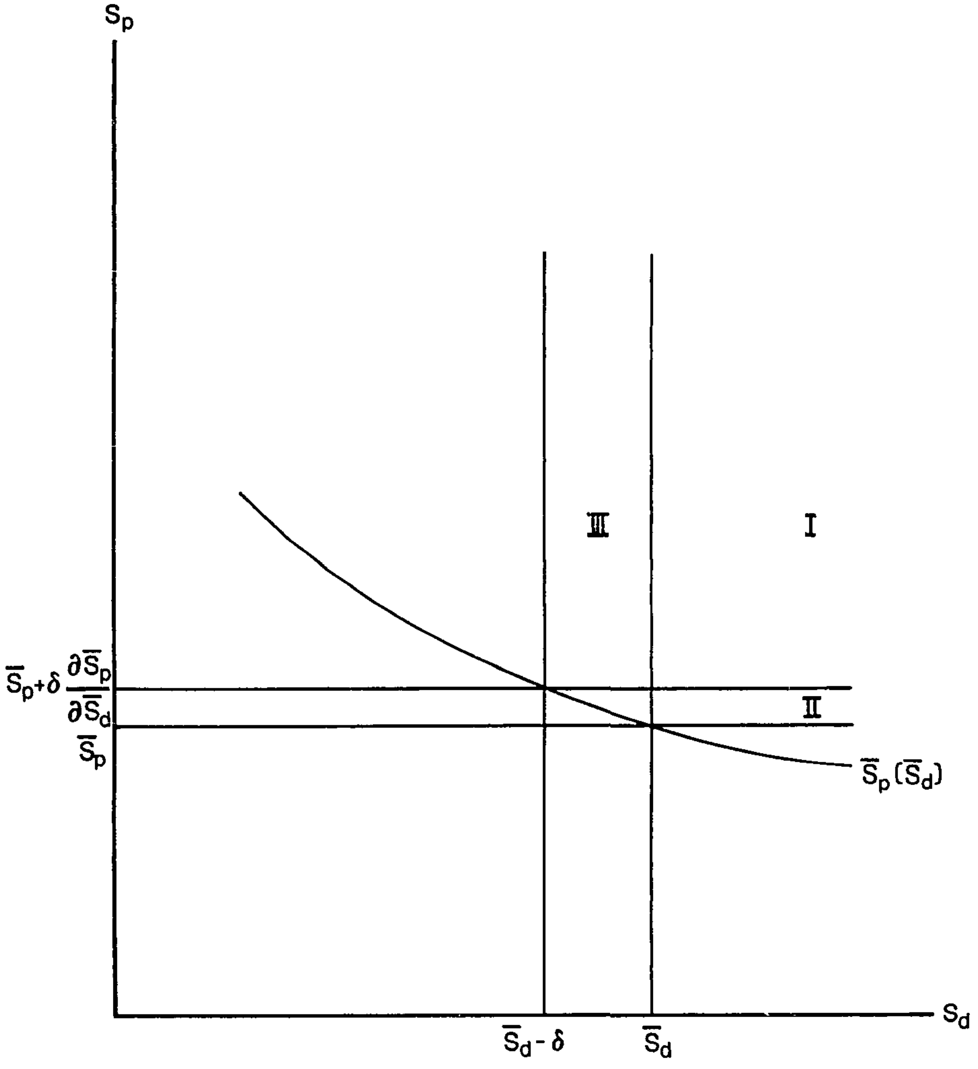

FIGURE 1
The Gains and Losses from Inducement

possible physician observations. The vertical axis is the range of possible patient observations. The curve $\bar{S}_p(\bar{S}_d)$ plots the optimal patient cutoff as a function of the physician's cutoff. A steeper curve would suggest greater resistance to inducement. Suppose the physician's cutoff is $\bar{S}_d$ and the patient's is $\bar{S}_p$. Then a patient receives treatment only if the pair (S_d, S_p) lies in regions I or II. The joint probability of these events is given in equation (5). Now suppose that the physician lowers her cutoff to $\bar{S}_d - \delta$. The patients raise theirs to $\bar{S}_p + \delta(\partial \bar{S}_p/\partial \bar{S}_d)$, and patients receive treatment only if (S_d, S_p) lies in regions I or III.

Region II corresponds to the loss in workload as the physician induces demand. This region consists of patients who refuse to consent because of the physician's enhanced reputation as an overprescriber. It is this reputation effect that limits incentives to induce. The probability of (S_d, S_p) falling in region II is given by the second term in (7). The magnitude of the loss is larger if $\bar{S}_d$ is low or if $\partial\bar{S}_p/\partial\bar{S}_d$ is large in magnitude (if patients are very resistant to inducement).

Region III corresponds to the gain in workload as the physician induces demand. This region consists of consenting patients for whom the physician has changed her recommendation from "no treatment" to "treatment." These patients know that the physician has lowered her cutoff but do not know that they personally are the victims of inducement, so they continue to consent. The probability of (S_d, S_p) falling in region III is given by the negative of the first term in (7). The magnitude of the gain from inducement is bigger if $\bar{S}_p$ is smaller (if many patients are willing to consent).

V. FACTORS AFFECTING PHYSICIAN AND PATIENT BEHAVIOR

Even if the physician always induces "to the maximum," she does not always choose the same $\bar{S}_d$. Her choice of $\bar{S}_d$ depends on her patients' willingness to consent, and this in turn depends on many factors. Tractable analysis of these factors requires a good deal of simplification. Dranove [1983] shows that the intuition is quite strong, however, and can be developed using less restrictive assumptions, though much more cumbersome mathematics. The following assumptions hold for the remainder of the section.

1. Z is uniformly distributed over the interval $[-\Theta, \Theta]$.
2. If the patient receives treatment, his utility is $-P$; if he does not receive treatment, his utility is 0 if $Z < \bar{Z}$ and $-M$ if $Z \geq \bar{Z}$ (where $P < M$).
3. $S_d = Z$ (physicians are perfectly accurate diagnosticians).
4. $S_p \sim U[Z - \gamma, Z + \gamma]$.

Under these assumptions, the patient faces only two possible health outcomes. If the patient is ill $(Z \geq \bar{Z})$, then treatment is said to be "necessary," and the gain from treatment, M, exceeds the price, P. If the patient is healthy, treatment is "unnecessary" and provides no benefits. Although there are only two possible outcomes, Zs differ according to the ease of diagnosis. For example, if $Z \gg \bar{Z}$, then the patient will always make the correct inference and consent. To avoid endpoint problems I assume that $\bar{S}_p - \gamma > -\Theta$ and $\bar{S}_p + \gamma < \Theta$. I also assume the physician is a perfect diagnostician, although the results which follow rely only on the assumption that the physician is a *better* diagnostician than patients. Note that when the physician is a perfect diagnostician she knows with certainty whether or not a treatment is necessary. The patient's diagnostic skill is captured by γ. When γ is smaller the patient is better able to determine whether a proposed treatment is necessary. γ may decrease as the patient's experience with the illness increases. For example, a longtime asthma sufferer may be a better judge of the value of hospitalization following an asthma attack than would be a first-time asthma sufferer. γ may also vary with the amount of publicly available information about the illness.

With this specification the average expected utility of patients is maximized when the physician chooses $\bar{S}_d = \bar{Z}$. Physicians only recommend necessary treatments, and patients always consent. Patient utility decreases monotonically as the physician lowers her cutoff to induce demand. This is because more unnecessary treatments are performed and an increasing number of genuinely ill patients refuse to consent because of the physician's tarnished reputation.[11]

Given these assumptions we can readily analyze the physician's and patient's decisions. We begin by solving for $\bar{S}_p$ in terms of $\bar{S}_d$. From (4) we know that $\bar{S}_p$ must satisfy

$$-P \int_{-\theta}^{Z} j(Z; \bar{S}_p)\, dZ + M \int_{Z}^{\theta} j(Z; \bar{S}_p)\, dZ = 0. \tag{8}$$

The density $j(Z; \bar{S}_p)$ depends on the relative values of $\bar{S}_p$ and $\bar{S}_d$. For example, if $\bar{S}_p - \gamma > \bar{S}_d$, then

$$j(Z; \bar{S}_p) \sim U(\bar{S}_p - \gamma, \bar{S}_p + \gamma). \tag{9a}$$

In this situation a marginal change in $\bar{S}_d$ would not affect the patient's posterior distribution so the patient cutoff would not change. Inducement is irrelevant in such situations, and so emphasis is on the case in which $\bar{S}_p - \gamma \leq \bar{S}_d$. In this case

$$j(Z; \bar{S}_p) \sim U(\bar{S}_d, \bar{S}_p + \gamma). \tag{9b}$$

Substituting this distribution into (8) yields

$$-P[(\bar{Z} - \bar{S}_d)/(\bar{S}_p + \gamma - \bar{S}_d)] + M[(\bar{S}_p + \gamma - \bar{Z})/(\bar{S}_p + \gamma - \bar{S}_d)] = 0. \tag{10}$$

By rearranging we can solve for $\bar{S}_p$ in terms of $\bar{S}_d$:

$$\bar{S}_p = (P/M)(\bar{Z} - \bar{S}_d) + \bar{Z} - \gamma. \tag{11}$$

From (11) these results follow:

1. $\partial\bar{S}_p/\partial\bar{S}_d = -P/M$. Resistance to inducement depends on the cost of unnecessary care (P) and the value of necessary care (M). Patients may be much more resistant to inducement in routine care (where P/M is relatively large) than in acute situations (where P/M is apt to be low).
2. $\partial\bar{S}_p/\partial\gamma = -1$. Patients who are poorer diagnosticians are more willing to consent. This is because they are less able to distinguish a necessary treatment from an induced treatment.
3. $\partial\bar{S}_p/\partial P = (\bar{Z} - \bar{S}_d)/M$. This is positive as long as any inducement is occurring (so that $\bar{S}_d < \bar{Z}$). As price increases, patients are less willing to consent.
4. $\partial\bar{S}_p/\partial M = -P(\bar{Z} - \bar{S}_d)/M^2 < 0$. As the value of necessary treatment increases, patients are more willing to consent.

[11] While a perfect diagnostician who recommends treatment when $S_d < \bar{Z}$ may be viewed as unethical, a near-perfect diagnostician who chooses a seemingly low cutoff might only be accused of erring too much on the side of caution. Such caution can still leave patients worse off.

Now recall the comparative statics results from section V. If patients are more resistant to inducement, the utility-maximizing physician will raise her cutoff; likewise if the patient's cutoff is higher. In light of these results, the following conclusions can be drawn about demand inducement.

1. As patient diagnostic skills improve (γ becomes smaller), physicians induce less demand. This is because fewer patients are willing to consent, and as a result the gains from inducement are smaller.
2. If price decreases, or if the value of necessary treatment increases, the physician will induce more demand. In either case, more patients are willing to consent, and they are less resistant to additional inducement.

The first result suggests an important benefit of second-opinion programs. Not only do they allow patients to make more informed consent decisions, they also discourage inducement. Thus, the information conveyed by the physician's recommendation will be more accurate as well. The second result implies that public policies and market forces that lower the physician's price also lead to greater demand inducement. Reductions in price may therefore be accompanied by a disproportionately greater number of unnecessary treatments. Similarly, insurance coverage that lowers out-of-pocket payments may also encourage inducement.

Relaxing the assumption that patients know the physician's cutoff with certainty further highlights the nature of the physician/patient relationship. Suppose that each patient knows the physician's cutoff only up to a probability distribution. Let this distribution be normal with mean $\bar{S}$ and variance σ^2 [N($\bar{S}$, σ^2)]. The patients determine their own cutoffs by solving (4), integrating over all possible values of $\bar{S}_d$. Each time a patient visits the physician, or asks friends and relatives about their visits, he obtains new information about the cutoff.[12] Suppose that the sampling distribution from which this new information is drawn is N($\bar{S}_d$, σ_d^2), where $\bar{S}_d$ is the actual cutoff chosen by the physician.[13] As each patient obtains this new information, he uses Bayes' rule to update his beliefs about the physician's cutoff. The patient's mean beliefs are distributed

$$N\{\bar{S}_d[\sigma^2/(\sigma^2 + \sigma_d^2)] + \bar{S}[\sigma_d^2/(\sigma^2 + \sigma_d^2)],\ \sigma_d^2\sigma^2/(\sigma^2 + \sigma_d^2)\}. \qquad (12)$$

The updated mean belief about the value of the cutoff is just the weighted sum of the prior belief and the new information. It is important to note that as the variance of the sampling distribution declines, patients place greater weight on the sampling information. It is likely that this variance would decrease as the physician/patient relationship endures.

[12] This model implicitly assumes either that the physician maintains her chosen cutoff for some time or that patients act as if she does. It is, of course, possible that the physician continually changes her cutoff or uses a mixed strategy to choose her cutoff. Such situations are beyond the scope of this analysis.

[13] The assumption of normality is mainly for analytic convenience, although if the patient has a lot of independent information then his estimate of the mean physician cutoff will be normal, by the central limit theorem.

If the physician changes her cutoff, this will affect the beliefs of all patients, and consequently affect all future consent decisions. However, the degree to which beliefs and consent decisions change depends on the variance of patient estimates of $\bar{S}_d$. For example, if a physician were to lower her cutoff by δ, all patients would lower their estimates of S_d by $[\sigma^2/(\sigma^2 + \sigma_d^2)]\delta$. If σ_d^2 does decrease as the physician/patient relationship endures, then the inducing physician should expect greater patient resistance as the relationship endures. The long-term relationship thus serves as a disincentive towards inducement. If, as some predict, the long-term physician/patient relationship is being threatened by prevailing market forces, physician incentives to induce demand may increase.[14]

VI. INDUCEMENT WITH MANY PHYSICIANS

The early inducement models of Evans [1974] and Fuchs [1978] predicted that the level of inducement would increase with the supply of physicians. They were unable, however, to generate such predictions without abandoning utility maximization. This prediction does follow from the results in this model.

To illustrate the effect of physician supply suppose that the market for physician services is monopolistically competitive, and that each physician faces an identical price elasticity of demand that is independent of individual quantities.[15] Then all physicians will choose the same price (chosen according to (6a)), and the same level of inducement (chosen according to (6b)). If the number of physicians per capita were to increase and if physicians maintained price and cutoff, then each physician would experience a decrease in workload. This would decrease the marginal cost of production (assuming concave utility for income and leisure). As a result, physicians would lower prices. It was shown in section V that at the lower price the reputation cost of inducement is lower, so physicians would simultaneously induce more demand. The overall effect on patient utility is therefore ambiguous.

The increasing supply of physicians may affect inducement in other ways. Satterthwaite [1979] suggests that as the physician/population ratio increases, each consumer's information about particular physicians will decrease. As discussed in section V, if consumers have limited information about particular physicians, physicians may have added incentive to induce demand.

In an extreme case patients may have a good idea about the average physician, but no information about individuals. In this case each individual physician has incentive always to prescribe treatment, because she suffers no reputation loss by doing so. However, if all physicians do this then patients will correctly perceive it and adjust their consent rules accordingly. The total number of treatments falls. Physicians face a "prisoner's dilemma," as they would have been better off if they had conspired to maintain a common

[14] In a thoughtful piece, Tarlov [1983] cited both the changing structure of the health services industry and the increasing supply of physicians as forces threatening the physician/patient relationship. Such views are shared by Starr [1982].

[15] Satterthwaite [1979] argues pursuasively in support of such assumptions.

recommendation strategy and thereby uphold the reputation of the profession. The collusive strategy which maximizes the total income of the profession is the monopolist's strategy described in section IV, a strategy involving prescription of unnecessary treatments. Since this collusive strategy involves less overprescription than the prisoner's dilemma outcome, patients benefit from this collusion as well! The profession's strong ethical standards and such practices as utilization review and the release by peer review organizations of guidelines for diagnosis and treatment may enable physicians to achieve the Pareto-superior outcome.

A final consideration is that when there are many physicians, patients may leave the practices of overprescribers in favor of physicans whom they believe will have more conservative recommendation strategies. This provides a disincentive to inducement. However, competition should not completely eliminate inducement.

To illustrate this point consider how patients choose their physicians. The physician's reputation for inducement is just one of many factors that may affect the patient's choice. Others include the physician's location, her manner, her perceived quality, etc. This suggests that many patients will be loyal to their physicians, and if a physician lowers her cutoff, she will lose some, but not all, of her patients. Of course, those who remain loyal will revise their consent decisions accordingly. The physician's lost workload from inducing demand therefore includes both the additional consent refusals and patients who no longer consult with the physician at all.

It is simple to translate these arguments into a model. As in section IV, let N be the size of the physician's practice and ρ equal the probability that a patient will receive treatment. Now, both ρ and N are functions of $\bar{S}_d$. The function $\rho(\bar{S}_d)$ was completely described earlier. We assume that $N(\bar{S}_d)$ is increasing in the relevant range (patients prefer physicians who do not induce unnecessary demand).[16] The physician chooses $\bar{S}_d$ to maximize ρN. A necessary condition for maximization is

$$N(\partial\rho/\partial\bar{S}_d) + \rho(\partial N/\partial\bar{S}_d) = 0. \qquad (13)$$

Recall that in the single physician case, $d\rho/d\bar{S}_d = 0$. If $dN/d\bar{S}_d > 0$, then the solution to (13) requires that $\bar{S}_d$ be higher than in the single physician case. That is, if consumers switch from physicians who induce too much demand, the amount of demand inducement will decrease.

VII. TESTING THE INDUCEMENT HYPOTHESIS

As Auster and Oaxaca [1981] point out, it may be impossible to find appropriate data to test the shifting demand version of the inducement hypothesis.

[16] I am precluding a very real possibility that some patients have high values of $\bar{S}_p$ and are so sure that treatment is warranted that they seek out physicians with low values of $\bar{S}_d$. With the exception of these hypochondriacs, however, it will generally be true that most patients prefer physicians with high $\bar{S}_d$. This is especially true if the physician/patient relationship is expected to be long-term and patients anticipate having many illnesses with many different severities.

However, models that specify what they mean by inducement behavior may be testable. To test for inducement behavior one needs to specify the empirical counterparts of the key variables identified in the model. If done satisfactorily, the appropriate tests are readily apparent.[17]

Inducement models, including this one, suggest that physicians induce demand by recommending procedures even though the available clinical information indicates that the expected costs of the procedure (to the patient) exceed the expected benefits. The medical literature documents many procedures for which these unwarranted recommendations are commonplace. As just one example, Kaplan et al. [1985] show that many laboratory tests, including blood cell counts, glucose level, and others, should be ordered only for specific recognizable indications. Otherwise, the cost of false-positive test results is excessive. Kaplan et al. document that roughly 50 percent of these tests are, in fact, unwarranted. The model predicts that the number of treatment recommendations that are unwarranted on the basis of clinical indications will vary systematically (inversely) with the price of the treatment. Moreover, if we compare across treatments, unwarranted recommendations will be more numerous when the medical value of warranted treatments is higher. Testing the prediction that the number of unwarranted recommendations varies inversely with price should be feasible. It may be more difficult to obtain objective evidence about the relative merits of warranted medical treatments. More promising might be to assess the merits of warranted automotive repairs (for which objective outcome measures might be available), and test the predictions in that market.

An alternative way of thinking about induced visits was presented by Wilensky and Rossiter [1983], who examined the relative frequency of physician treatments initiated by patients (initial visits to the physician) and by physicians (follow-up visits and subsequent treatments). The latter were considered induced visits. If price falls then we should expect more patients to initiate visits ($\bar{S}_p$ falls). These additional patients are likely to be healthier than the average patient, so the relative proportion of further treatment recommendations by physicians should fall, ceteris paribus. If, however, when price falls physicians relax the indications for recommending additional treatments, the proportion of follow-up visits may increase. (Note that Pauly [1980] obtains the opposite prediction when physicians choose accuracy and have benevolent motives.) Wilensky and Rossiter, using data from the 1977 National Medical Care Expenditure Survey, confirm this prediction. They found that a 1 percent decrease in the out-of-pocket price of physician visits increased the proportion of visits that were physician-initiated by .17 percent.

Another prediction of the model (and Pauly's as well) is that inducement

[17] Observing that physicians do not make optimal recommendations for their patients does not, by itself, constitute proof of the inducement hypothesis. Dionne and Contandriopoulos [1985] show that prestige and ethical concerns alone are sufficient to drive a wedge between physician actions and patient desires.

should vary inversely with patient diagnostic skill. For example, physicians are expected to tighten the clinical indications for hospitalization of asthma sufferers with each succeeding asthma attack. If one accepts educational attainment as a proxy for diagnostic skill (e.g., more educated people may have greater access to and understanding of the popular medical press), then Wilensky and Rossiter again provide supporting evidence. They report that the percentage of physician initiated visits falls with educational attainment, ceteris paribus.[18]

Finally, perhaps the most obvious prediction of this and other inducement models is that if we eliminated the financial reward from successful inducement, the prescription of expensive treatments would fall. By salarying physicians, health maintenance organizations [HMOs] remove the incentive to induce demand. Luft [1978] documents that hospitalization rates at HMOs are as much as 40 percent below hospitalization rates for fee-for-service patients, while utilization rates for less intensive procedures are roughly equal. This is consistent with my model. My model would also predict that consent rates at HMOs were higher than for fee-for-service physicians.

VIII. DISCUSSION

Asymmetric information in the physician/patient relationship drives a wedge between what patients would like physicians to do and what physicians actually do. One way to eliminate this wedge would be to eliminate the link between diagnosis and treatment. For instance, if there were physicians who only diagnosed patients, receiving the same fee regardless of diagnosis, they would have no incentive to recommend unnecessary treatments. One possible reason we do not generally observe this is that physicians may make too much money performing treatments, to specialize in diagnosis.[19] Additionally, diagnosis and therapy are often intermixed and the cost advantage of their joint production may offset the disadvantages of inducement behavior. These points aside, it is still not clear that such a system could be achieved. Physicians must depend on an intricate referral network to get patients, and if diagnosticians must also hook up to this network, they might still be rewarded for inducing extra treatments. The incentive problems that generate inducement would remain.

Another way to eliminate the wedge may be to eliminate fee-for-service practice. If physicians receive the same fee regardless of the treatment performed, this should also eliminate the incentive to induce demand. This is one of the ideas behind prepaid group practices such as HMOs. Of course, prepayment may introduce incentives to recommend too few treatments, as providers try to reduce costs. Just as the market might not discipline overprescribers, it might also fail to discipline underprescribers.

[18] An alternative explanation of this result is that the less educated consistently underestimate the value of the treatment.

[19] We do see auto diagnostic centers, however.

APPENDIX

This appendix describes two important statistical properties used in the text.

1. The densities $h(a)$ have the strict monotone likelihood ratio property (MLRP) if for every $x > y$ and $S^* > S'$,

$$h(x \mid S^*)h(y \mid S') - h(x \mid S')h(y \mid S^*) > 0.$$

This translates into something akin to "more severe symptoms are associated with more severe illnesses." Alternatively, suppose that A and S are jointly distributed with conditional density $h(a \mid s)$. If h has the MLRP, then one would correctly infer that if S is observed to be high (low), a is also likely to be high (low).

2. We say a distribution F dominates another distribution G in the sense of first-order stochastic dominance ($F \overset{d}{\geq} G$) if and only if, for all a, $F(a) < G(a)$. For any nondecreasing function U, if $F \overset{d}{\geq} G$, then

$$\int_{-\infty}^{\infty} U(a)f(a)\, \mathrm{d}a \geq \int_{-\infty}^{\infty} U(a)g(a)\, \mathrm{d}a.$$

If F is first-order stochastic dominant, then the distribution G has more weight in the lower tail. If $U(a)$ is a nondecreasing function and F and G are distributions with $F \overset{d}{\geq} G$, then the expected value of U is higher when the distribution of a is $F(a)$. This is because when the distribution is F, high values of a (and therefore high values of U) are more likely to occur.

Let Z be a random variable whose densities have the strict MLRP. For any two intervals $[a, b]$ and $[c, d]$, with $a \geq c$ and $b \geq d$, where at least one inequality is strict,

$$F(Z \mid S \epsilon [a, b]) \overset{d}{\geq} F(Z \mid S \epsilon [c, d]).$$

The result holds if $b = d = \infty$. In this case, if S is observed to be in an interval containing high (low) numbers, Z is correctly believed to be high (low) as well. For further discussion of the use of the MLRP in economic theory, see Milgrom [1981].

REFERENCES

Anderson, R., D. House, and M. Ormiston. "A Theory of Physician Behavior with Supplier-Induced Demand." *Southern Economic Journal*, July 1981, 124–33.

Auster, R. and R. Oaxaca. "Identification of Supplier-Induced Demand in the Health Care Sector." *Journal of Human Resources*, Summer 1981, 327–42.

Coyte, P. "The Economics of Medicare: Equilibrium Within The Medical Community." Ph.D. dissertation, University of Western Ontario, 1982.

Crawford, V. and J. Sobel. "Strategic Information Transmission." *Econometrica*, November 1982, 1431–51.

Dionne, G. and A. P. Contandriopoulos. "Doctors and Their Workshops: A Review Article." *Journal of Health Economics*, March 1985, 21–34.

Dranove, D. "Physician-Induced Demand and Home Nursing Care: Two Economic Analyses." Ph.D. dissertation, Stanford University, 1983.

Evans, R. "Supplier-Induced Demand: Some Empirical Evidence and Implications," in *The Economics of Health and Medical Care,* edited by M. Perlman. New York: Wiley, 1974.

Feldstein, M. "The Rising Price of Physicians' Services." *Review of Economics and Statistics,* May 1970, 121–33.

Fuchs, V. "The Supply of Surgeons and the Demand for Operations." *Journal of Human Resources,* Supplement 1978, 35–56.

——— and M. Kramer. "Determinants of Expenditures for Physicians' Services in the United States 1948–1968." Technical Report, National Bureau of Economic Research, 1973.

Green, J. and N. Stokey. "A Two-Person Game of Information Transmission." Harvard University, HIER Discussion Paper 751, March 1980.

Kaplan, E., L. Sheiner, A. Boeckmann, M. Roizen, S. Beal, S. Cohen, and C. Nicoll, "The Usefulness of Preoperative Laboratory Screening." Journal of the American Medical Association, 28 June 1985, 3576-81.

Luft, H. "How Do Health Maintenance Organizations Achieve Their Savings?" *New England Journal of Medicine,* 15 June 1978, 1336–43.

McCarthy, T. "The Competitive Nature of the Primary-Care Physician Services Market." *Journal of Health Economics,* June 1985, 93–118.

Milgrom, P. "Good News and Bad News: Representation Theories and Applications." *Bell Journal of Economics,* Autumn 1981, 380–91.

Pauly, M. *Doctors and Their Workshops.* Chicago: University of Chicago Press, 1980.

——— and M. Satterthwaite. "The Pricing of Primary Care Physicians' Services: A Test of the Role of Consumer Information." *Bell Journal of Economics,* Autumn 1981, 488–506.

Pitchik, C. and A. Schotter. "Internal and External Regulation of Markets with Asymmetric Information." Working Paper, Center for Applied Economics, New York University, 1984.

Ramsey, J. and B. Wasow. "A Re-evaluation of Supply and Demand Concepts in Physician Care." Report to the Department of Health and Human Services, Contract No. HRA-79-0068, 1981.

Reinhardt, U. "The Theory of Physician-Induced Demand: Reflections After a Decade." *Journal of Health Economics,* June 1985, 187–94.

Satterthwaite, M. "Consumer Information, Equilibrium Industry Price and the Number of Seller." *Bell Journal of Economics,* Autumn 1979, 483–502.

———. "Competition and Equilibrium as a Driving Force in the Health Services Sector." Working Paper, Northwestern University, 1982.

Starr, P. *The Social Transformation of American Medicine.* New York: Basic Books, 1982.

Tarlov, A. "The Increasing Supply of Physicians, the Changing Structure of the Health-Services System, and the Future Practice of Medicine." *New England Journal of Medicine,* 19 May 1983, 1235–44.

Wilensky, G. and L. Rossiter. "The Relative Importance of Physician Induced Demand on the Demand for Medical Care." *Milbank Memorial Fund Quarterly,* Spring 1983, 252–77.