Volume Title: Doctors and Their Workshops: Economic Models of Physician Behavior

Volume Author/Editor: Mark Pauly

Volume Publisher: University of Chicago Press

Volume ISBN: 0-226-65044-8

Volume URL: http://www.nber.org/books/paul80-1

Publication Date: 1980

Chapter Title: Physician Information and the Consumer's Demand for Care

Chapter Author: Mark Pauly

Chapter URL: http://www.nber.org/chapters/c11524

Chapter pages in book: (p. 43 - 64)

4 Physician Information and the Consumer's Demand for Care

In chapter 1, I developed the rudiments of a simple model of a consumer's demand for medical care, conditional on the level of health physicians had led him to expect to result. In this chapter, I shall examine in much more detail how consumer expectations are formed.

If consumers always believed all that physicians told them, and accepted advice unquestioningly, then the only constraint on movement to that level of demand which would maximize physician income—probably a very high level of demand indeed—would be the moral scrupulousness of physicians. The physician's attitude toward truth-telling, or "accuracy" as I shall call it hereafter, can be shown to be an important influence on a consumer's use of care and on health levels, but it may not be the only influence. In particular, consumers can control the effect of physician-provided information on their behavior by deciding, within some limits, both how to react to advice and which physicians to patronize for advice. In some emergencies, the consumer does not of course have these options. But the bulk of medical encounters are not of this sort, and even in many situations labeled "emergency" the consumer has in principle a considerable amount of power over what can be done to him (including whether or not he chooses to be an "emergency" case) and which physician he chooses in order to obtain advice.

In view of recent questioning of the appropriateness of medical advice and medical decisions from such diverse parties as Ivan Illich[1] and the U.S. House of Representatives Subcommittee on Oversight and Investigations,[2] it seems appropriate to determine whether a more analytical approach to the question can make a contribution. As will soon become apparent, even a designedly simple approach to modeling the problem soon becomes quite complicated.

What will determine how a consumer will react to physician advice? Intuitively, one might suppose that the stronger a person's prior beliefs, the less he will respond to the information provided. Unfortunately, this conjecture is incorrect in general for reasonable measures of "strength of beliefs." But it is still possible to show that for some sufficiently high level of prior certainty, this conclusion will follow. Thus there is a theoretical basis for the empirical expectation that those spending units characterized by a sufficiently high level of a priori information will, other things being equal, be less responsive to changes in information obtained.

The model is one in which the consumer's demand for medical care is conditional on the content of the advice. There are two critical questions to be addressed: (1) What determines from which physician he will seek advice? (2) What determines how the consumer will respond to the advice? Corresponding to each of these aspects of demand, there is an appropriate supply response: first, the content of the advice each physician decides to provide; and second, the overall content of advice physicians choose. In what follows I will examine each of these decisions.

The Consumer's Demand for Medical Care

The approach here is first to develop a simple model of the consumer's demand for medical care under certainty, then introduce uncertainty but with information unavailable, and finally to show the consequences of permitting information to become available. One possible aspect of behavior that will not be incorporated here is a possible consumer suspicion, based solely on the content of the advice received, that the physician is willfully not providing accurate information. In the model to be discussed, the physician may lie, and the consumer may not believe him, but consumers do not believe that physicians individually or collectively lie on purpose; physicians are only supposed to make honest mistakes.

It is assumed that the consumer has a single period utility function in health H and other goods x. The intertemporal aspects of the choice of health levels and of the production process for health, which have been treated extensively by Grossman,[3] will be ignored here. Likewise the time cost of obtaining health will be ignored. As in chapter 1, the utility function is

$$U = U(x, H) \tag{1}$$

The composite good x is available at a price of unity, but health must be produced. The production function for health is

$$H = g(M, H_o) \tag{2}$$

The marginal health product of medical care g_1 is positive, the effect on final health of H_o (or g_2) is positive, and g_1 is larger the smaller the value of H_o, or $g_{12} < 0$ (i.e., medical care benefits people more the sicker they are). For simplicity, it is assumed that, given H_o, g_1 is a constant, i.e., there is a constant marginal and average health product. If the consumer knows with certainty that the value of H_o is $\bar{H}_o$ and the value of g_1 for that H_o is $\bar{g}_1$, then his problem is to maximize (1) subject to

$$(2') \qquad H = \bar{H}_o + \bar{g}_1 \cdot M$$

and

$$(3) \qquad Y = x + pM.$$

Given his endowment (Y,H_o) and the shadow price of health, g_1/p, the consumer chooses the amount of health he wishes to buy.

The consumer might be uncertain either because he does not know H_o or because he does not know g_1 for a given H_o. That is, he might be uncertain either about what is wrong with him, which determines H_o, or how effective medical care is in dealing with his condition, which determines g_1. (In a more complex but realistic model, g_1 might depend not just upon H_o, but upon the particular disease the person has. But for simplicity I will continue to assume that conditions are only classified by severity.)

The first case to be considered is that in which g_1 is uncertain but H_o is known. Suppose g_1 (given H_o) has the (subjective) distribution $f(g_1)$. The problem then is to maximize expected utility:

$$(4) \qquad EU = \int U\,(x - pM, H_o + g_1 M)\, f(g_1)\, dg_1$$

subject to constraint (3). Solution of this problem implies that M is chosen so that $\pi_i u'\,(g^i_1) = \pi_j u'\,(g^j_1)$, where π_i and π_j are the probabilities attached to two alternate values of g_1. In effect, the consumer chooses a level of M that would be somewhat appropriate for all possible states, but ideally appropriate for almost none.

Now assume that the consumer can obtain information on H_o or g_1 from the physician. In order to explain how the consumer will respond to any given information, it is necessary to explain how he judges the accuracy (or diagnostic and prescriptive skill) of a physician. One way the consumer can tell whether a physician is giving him accurate advice is to observe the results of experiments. Those experiments would take the following form: suppose g_1 is uncertain but H_o is known. Suppose a physician asserts that the value of g_1 is $\bar{g}_1$. The consumer would then observe whether $H = H_o + \bar{g}_1\bar{M}$ when he uses $\bar{M}$ units of care.

The result of a single such experiment will ordinarily not be conclusive. H might differ from $H_o + \bar{g}_1\bar{M}$ for a number of reasons. For exam-

ple, the true production process might be $H = H_o + g_1 M + u$, $E(u) = 0$, $\sigma^2_u > 0$. Then additional observations on sample values of g_1 would be needed to get an estimate of the population mean of (actual) g_1. Given some a priori distribution, the consumer can make an estimate of the accuracy with which this physician predicts g_1. A similar argument holds for H_o. Note that the physician's motivation is irrelevant here.

Suppose then a person is trying to determine whether or not he would buy diagnostic information, and from which physician he would purchase it. His "information about the accuracy of the information" will be relevant. What will determine the amount of such information he possesses? It is the number of experiments he has observed, or the price of additional experiments. These "experiments" will represent his own encounters with the physician, or those of his friends. Moreover, his own skill in evaluating observed results may also affect his perceptions of informational accuracy. Given his own and others' experiences, the consumer can come to subjective estimates of the value of g_1, and of the accuracy of physician conjectures about the level of g_1.

Now consider the effect of such estimates on the consumer's choices. Suppose $g^{\tau}{}_1$ is the actual value of g_1. Suppose the physician can determine $g^{\tau}{}_1$. Finally, suppose the person is given information by a physician on what the value of g_1 is; this physician advice is represented by $g^{\phi}{}_1$, and is not necessarily equal to $g^{\tau}{}_1$. If the person's prior distribution were $f(g_1)$, his posterior distribution $f(g_1|g^{\phi}{}_1)$ is given by Bayes' rule as

$$f(g_1|g^{\phi}{}_1) = \frac{f(g^{\phi}{}_1|g_1)}{f(g^{\phi}{}_1)} \cdot f(g_1)$$

where

$$f(g^{\phi}{}_1) = \int f(g^{\phi}{}_1|g_1)\, f(g_1)$$

In each case, the value of $f(g_1)$ is adjusted by the ratio of the conditional to the unconditional distribution of $f(g^{\phi}{}_1)$. We can immediately distinguish two special cases. First, suppose the person knows the truth with certainty. Consequently, the prior and posterior distributions are the same. Information does not affect this person's demand for care at all.

The alternative case is one in which the person puts complete confidence in the physician's opinion, so that his posterior estimate of g_1 is identical to what the physician tells him.

If physicians tell the truth ($g^{\phi}{}_1 = g^{\tau}{}_1$), then M will be set at M_{τ} for both kinds of persons. If physicians are not always accurate, then the person who is certain of the truth still chooses M_{τ}, but the person who believes the physician will choose some quantity other than M_{τ}. If persons are of either of these two extreme types, an empirical measure that distinguishes them permits one to make predictions about the possible

responsiveness of demand to physician information which has varying degrees of accuracy.

Difficulties arise when either the a priori distribution or the likelihood function are not of the degenerate forms discussed here. While it is easy to show that the change in probability π attached to same value of g_1 in response to information, given some likelihood function, is smaller the larger is $|\pi - 1/2|$, it does not follow that the change in the preferred level of M will be smaller. That depends on how the preferred M changes with changes in π. If M changes very rapidly with changes in π for a π in excess of $1/2$, then it is possible that information may make a bigger difference in the use of such persons as compared to the use of more "uncertain" persons, with π closer to $1/2$. One can say that the response to new information of the individual's use of care is likely to be different for persons with different prior beliefs or prior stocks of information. But it does not appear that one can make any general a priori conjectures about the direction of this relationship.

However, theoretical determinateness can be salvaged if the extremes are considered. It can be said that one can find some π sufficiently close to 1 or zero that the effect on preferred M of the message $g_1 = g^{\phi}{}_1$ is smaller than the effect of that message on use at any π further away from these extremes. Since the effect of any change in π on M must be finite, if we can find some $\Delta\pi$ (as a result of receiving information) that is sufficiently small, we can get as small an effect on M as we want. All we need to show is that, in the neighborhood of $\pi = 1$, we can find a π such that $\Delta\pi$ as a result of the message $g_1 = g^{\phi}{}_1$ is as small as we want. Consider some value of $\pi = (1 - \epsilon)$. Since the physician's advice has no effect when $\pi = 1$, by continuity it follows that, by selecting some value of ϵ sufficiently small, we can make $f(g_1|g^{\phi}{}_1 \neq g^{\tau}{}_1)$ as close to $f(g_1)$ as we want for any given likelihood function.[4] That is, we can make the posterior probability as close to the prior probability as we want, even if the physician provides incorrect information.

This discussion of the effect of information on use suggests that, however indeterminate the relationship in general, one can find sufficiently extreme values of information and ignorance such that the informed are less responsive to information than the ignorant. This proposition will be the basis of the empirical analysis.

A Censoring Problem

This proposition applies only if the set of conditions for which informed and uninformed persons receive physician advice is the same. Such an assumption may not be plausible in general. One problem is that the incidence of conditions may differ according to the level of information, but this problem is not likely to be serious. A more serious

problem arises from a kind of censoring. The person who is virtually certain of the truth will ignore erroneous physician advice, as described in the preceding section. But he will also have no incentive to seek that advice in the first place. Those persons in the "well-informed" set who actually meet with physicians will tend to be precisely those whose behavior is easier to change; the unresponsive persons will have been "censored" out initially, independent of the eventual content of physician advice. Those persons classified as poorly informed who seek advice may therefore be no more responsive than those persons classified as well informed who seek advice.

This difficulty will not be important if persons who seek medical therapy usually must first see a physician and get his advice (or at least his diagnosis), whether they demand that advice or not. Suppose, for example, a child in a well-informed family has tonsilitis. The parents know that tonsillectomy is not warranted for his condition. Nevertheless, they must go through a physician in order to obtain a prescription for antibiotics, and are therefore potentially exposed to the content of physician advice.

It seems reasonable to suppose that many conditions for which demand creation is likely are of this sort; at least one physician contact is often needed for therapy, no matter what the state of patient information. As long as those persons in the well-informed set who really do seek advice (i.e., are not virtually certain) are not *more* responsive than those in the less well-informed set, the elasticity, and probably the magnitude, of the response of the well-informed who use positive amounts of care will be smaller than that of the less well-informed. This occurs because the well-informed set will include some persons who really are virtually certain, but are compelled to go through at least one physician in order to obtain any care at all. That is, as long as some of those persons who are truly well-informed are persons who seek care (even if they do not seek advice), the overall response of persons in the well-informed set will be smaller, other things being equal.

The Level of Accuracy, the Demand for Care, and Physician Availability

The empirical finding for which we seek a theoretical framework is that, *ceteris paribus*, the demand curve for physicians' services appears to shift when the stock of physicians per capita changes, because accuracy of physician advice decreases. The purpose of this section is to construct models which are consistent with a negative relationship between physicians per capita and accuracy. It is not my intent to argue that a negative relationship *must* hold. A model is useful if it *permits* a

negative relationship to hold; as will be shown, certain otherwise attractive and plausible models do not permit a negative relationship to occur, and so those models must be discarded.

For the purpose of distinguishing among models, the classification in chapter 1 is useful. To begin with, there is no point in discussing competitive market clearing models. If the seller takes price as given and supposes that he can sell as much as he wants at that price, there is no reason for him to alter accuracy in order to sell more. Consequently, the discussion will primarily be concerned with noncompetitive models.

Model 1: The Physician as a Real-income Maximizing Monopolist

The usual assumption in studies of labor supply is that the agents have two arguments in their utility functions: money income and leisure. In order to simplify the exposition, it will be assumed that physician time is available at a constant opportunity cost. This assumption avoids the necessity of including leisure in the utility function, and makes maximization of utility equivalent to maximization of the difference between total revenue and total opportunity costs, including the opportunity cost of leisure foregone.

The physician may be thought of as selling two products: diagnostic information and therapeutic care. These markets are not separate. The amount of information about a product that consumers will want to buy will depend upon both the price of information *and* the price of the product, in this case therapeutic care. Likewise, the amount of therapeutic care that a person will eventually buy at a given price will depend upon the price of information. Although the quantity of each type of care demanded is inversely related to its own price, it is not possible to establish definitive comparative statics results for the cross-price effects. The problem is further complicated in practice by the nonmarginal nature of many information purchases. Often in order to receive any therapeutic care at all the individual is required to seek diagnosis; in principle the price of diagnostic information could absorb all of the consumers' surplus from therapeutic care.

But the concern here is not primarily with these price and quantity effects. Rather, we wish to determine the accuracy or the content of a *given* amount of information purchased. Assume initially that the accuracy of the diagnosis does not affect a physician's information demand curve. Holding other things constant, including leisure time and the physician time devoted to diagnosis and therapeutic care, the real-income maximizing physician will then adjust the level of accuracy A to that level at which the increase in his net income from changing accuracy

is zero. This conclusion implies that, for any quantity of therapeutic care demanded, accuracy is set at that level which maximizes the unit price paid.

If total demand for therapeutic care Q_D is given by $Q_D = Q(P, A)$, where P is the user price of care and A is the level of accuracy among a set of identical physicians, then the individual physician demand Q^i_D, assuming pro-rata sharing among N identical physicians, is $Q^i_D = 1/N\ Q_D(P, A)$. The individual physician demand curve when A is set at the level of true information, or A_T, is shown in figure 4.1 as $D^i_T = 1/N\ Q_D(P, A_T)$. But the maximizing physician will not choose to confront this curve; instead he will probably choose that level of accuracy which

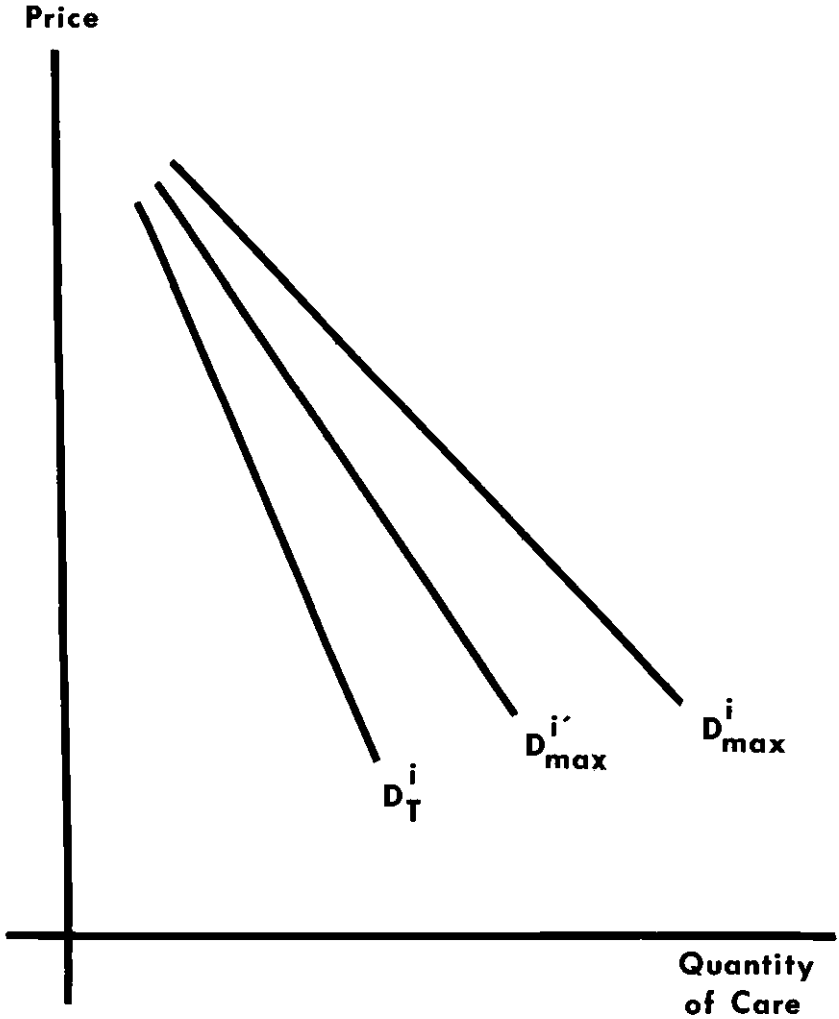

Figure 4.1

maximizes price at every quantity. This choice yields a maximum curve, $D^i_{max} = 1/N\, Q_D(P, A_{max})$, where A_{max} is the level of accuracy which maximizes P at every Q_D. He will then choose the profit-maximizing level of P given this curve.

The effect of increasing the supply of physicians can be thought of as a pro-rata decrease in the demand facing each physician. (It might also change the extent of competitiveness, but that will be ignored here.) The D^i_{max} curve will shift to $D^{i'}_{max}$, the result being smaller output per physician at any price, and probably a lower price. However, this change will not necessarily affect the level of accuracy, and it will have no effect on the market demand curve D^i_{max}. That level of accuracy which maximizes price at a given quantity demanded from a particular seller also maximizes price at any pro-rata share of that quantity. There would, in this situation, be *no* detectable availability effect resulting from an increase in the number of physicians (even though demand is, of course, different from D_T). Quantity demanded would rise as price falls along the D^i_{max} demand curve, but this would be wholly captured in an accurate measure of user price. Any observed availability effect must be due to changes in nonmonetary rationing. If we are to postulate a measurable availability effect arising from information, we must enlarge the set of arguments in the physician's utility function.

Model 2: The Physician as a Partially Benevolent Oligopolist

In this section we shall consider a model of demand creation among physicians who take price as given but do not expect to be able to sell unlimited amounts at that price. In addition to being realistic, especially for situations in which third parties set fee levels, this model permits us to throw into sharp focus the physician's incentive to alter accuracy.

It may not be unreasonable to assume that alterations in accuracy operate mainly on quantity, not price, at least as far as the individual physician is concerned. Oligopolistic features of the industry may lead to a reluctance to raise price, as may a kind of altruism that recognizes that higher prices give no benefit to consumers, while higher quantities may provide some benefit to both consumers and physicians.

It is assumed that physicians are partially benevolent, in the sense that the physician's maximand also includes a measure of accuracy. Other things being equal, physicians would rather tell the truth, but they would be willing to surrender some accuracy for some amount of money income. The physician obtains utility from real income Y, which is total revenue minus total opportunity costs, and from accuracy A. Both real income and accuracy are normal goods.

With P fixed, the physician who does not get utility from accuracy will choose that level of accuracy at which the quantity demanded from

him equals or gets as close as possible (given demand) to that quantity at which price equals "marginal cost," where the latter includes the value of sacrificed leisure. However, when accuracy yields utility, the level of accuracy is the one which satisfies

$$\frac{U_A}{U_Y} = (P - \partial C/\partial Q)\ \partial Q/\partial A \tag{5}$$

where C is total opportunity cost, including the opportunity cost of physician time, and U_A and U_Y are the marginal utilities of accuracy and income respectively.

These effects are shown in figure 4.2. Suppose price is initially at P, and "MC" measures both money marginal cost and the money value of sacrificed leisure. At any P, income would obviously be maximized by setting Q either at that Q at which $P =$ "MC" or at the Q on the D^i_{max}

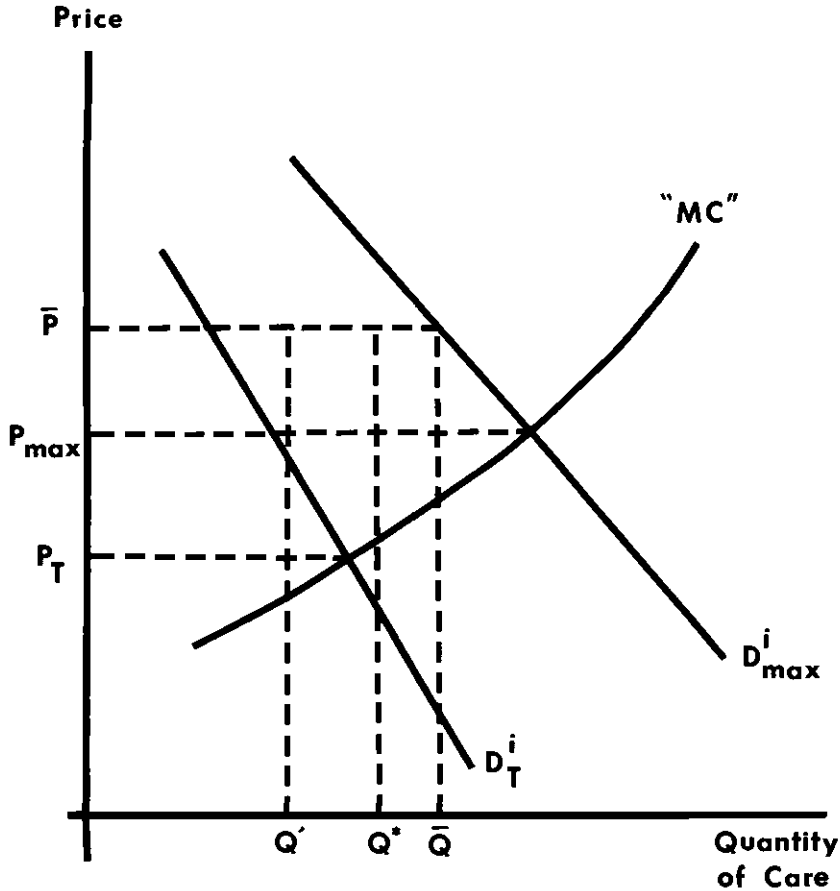

Figure 4.2

demand curve, whichever is less. If $P = \bar{P}$, for example, Q would be set at $\bar{Q}$. But if the physician values accuracy, he may not be willing to expand Q (and reduce A) to this level. He may instead set accuracy at the level given by equation (5), and consequently set Q at, say, Q^* less than $\bar{Q}$.

Now let the number of physicians be increased. Each physician's Q will fall below Q^*, say to Q', if A stays at its initial level. At this point, the physician's income is less so he may be willing to sacrifice some accuracy to recoup part of the loss in income. (Ordinarily he would not wish to reduce accuracy so much as to restore his original level of income.) Moreover, at Q', there will also be a bigger gain from increasing Q by one unit, since P — "MC" is greater. If $\partial Q/\partial A$ is constant or increases as Q decreases, then the net return from changing A will also have increased.

As with most oligopoly models, this model does not make specific predictions about the equilibrium level of price. It is likely that price would decline below $\bar{P}$, though by how much, and to what extent the decline would be related to physicians per capita, is impossible to say. But whatever the final price, at that price there would generally be some incentive to create demand. Moreover, the incentive to create demand would generally be greater at a given price the larger the number of physicians.

There would, however, be no incentive to create demand if the price settled to P_T or below. Note that if P equals P_T, a physician, even an income-maximizing one, would tell the truth; an incentive to create demand requires excess supply at P_T. This is an important result, and it will be discussed in more detail later in this chapter. Note also that P_T falls as physician stock increases.

If price is given but is initially below P_{max}, then use will increase with physician stock *even if* physicians only value income, not accuracy. In effect, use is constrained by supply, and follows the outward-shifting aggregate supply curve. The more elastic supply is, the less likely that price is below P_{max}. For example, if there are constant returns to scale in the supply of some type of physician service, the quantity given by the D_{max} demand curve will be supplied for any price in excess of average cost, and no availability effect will be observed.

Figure 4.2 also indicates that, *given* the number of physicians, the extent of demand creation will vary positively with the price, unless income effects are very strong. The reward from creating demand, or $P - MC$, is obviously larger at any Q the higher P is. Of course, income will also be higher if P is higher, and this "income effect" may to some extent offset the "substitution effect" in the direction of reduced accuracy.

Model 3: The Physician as a Partially Benevolent Monopolist

If the physician has some control over price, it is still true that in equilibrium the marginal effect of changes in accuracy on his income must compensate him for the disutility of reduced accuracy. The effect that changes in accuracy have on the level of income depends upon *how* the demand curve is shifted when A changes. Whatever price is finally charged, the marginal condition (5) must obviously still be satisfied. In addition, the usual $MR = MC$ condition, given the level of accuracy, must hold as well. By the same arguments as in the preceding section, the level of accuracy will decrease with the physician stock at any *given* level of price if $\partial Q/\partial A$ is constant or increases with decreases in Q. Consequently, one would expect to observe an availability effect on consumer demands.

This appears to be the only unambiguous prediction one can obtain from "demand-creation" models. As Feldman and Sloan[5] have shown, and as Reinhardt[6] has emphasized, literally any relationship between equilibrium market price and physician stock is consistent with such models. Depending upon how the demand curve pivots when it shifts, the new equilibrium price could be above, equal, or below the price that prevailed with a smaller stock of physicians.

This result has led both Evans[7] and Reinhardt to conclude that it is not possible to refute the demand creation hypothesis by looking at the relationship between physician stock and price. The most that such an investigation could do is to cast doubts on the neoclassical competitive hypothesis of no demand creation. Even this weak conclusion is really not very useful. If the market is monopolistic rather than competitive, and individual physician demand curves become less elastic as the physician stock rises, then price can vary directly with that stock even in a neoclassical, income-maximizing, but noncompetitive world. While one usually does not assume that firm demand curves become less elastic when more competitors enter, Satterthwaite[8] has recently shown that the effect of numbers of sellers on information about prices or quality levels of providers can lead to such a result without having consumers pushed off their "true" demand curve by changes in accuracy. Moreover, Pauly and Satterthwaite[9] have found that variation in the level of prices across cities can be explained without recourse to the modified target income model.

The relationship between price and availability will therefore not settle the question of the availability effect, but a study of the relationship between availability and use, *given* price, and between price and use, *given* availability, will provide some answers. These answers will be valid even if prices are not fixed, as long as price is exogenous to the

individual demander. It is precisely this kind of test that will be presented in the next chapter.

Information about Information

Up to this point, we have taken the consumer's estimate of the accuracy of physician advice as depending only on the total number of experiments or experiences he has observed. We have also assumed that the consumer does not change physicians in response to the content of information provided by the physician; the consumer's only decision is whether to follow that advice or not. But consumers may well perceive differences among physicians with regard to the accuracy of their advice, and act on those perceptions in choosing which physician to patronize. In other words, the individual physician firm's demand curves for its output of advice and for its output of therapeutic care may increase if it is perceived to give more accurate advice. Such a response may provide an incentive to the physician to offer accurate advice; it constrains his ability to generate net increases in demand. In terms of the geometric presentation, competition may affect the position of the D^i_{max} curve; whether or not it actually pushes the curve in as far as D^i_T, competition may still affect the observed level of accuracy.

We now need to ask how competition, as measured by the number of physicians (in total or per capita), affects the equilibrium level of accuracy. While a complete characterization of this kind of equilibrium with information and goods as joint products has not yet been fully developed, the following simple model possesses what appear to be critical insights.

Suppose a market area has N identical patients (customers) and M identical physicians (sellers). Suppose that the selling price of therapeutic care is fixed at $\bar{P}$, which is above "MC" at the output of therapeutic care demanded when the truth is told. Suppose profits are only earned on therapeutic care. At any level of accuracy A, let demand per patient be $Q(A)$; demand per physician is $\frac{NQ}{M}$. Suppose that each physician takes other physicians' levels of accuracy as given. Finally, suppose that any one physician expects that, if all other physicians maintain their (identical) levels of accuracy, and he raises his, he will obtain some fraction k of all other physicians' customers.

Given some initial level of accuracy, a physician's income will then be increased by increases in his level of accuracy if

$$k \cdot \frac{NQ}{M} \cdot (M-1) > \frac{\partial Q}{\partial A} \frac{N}{M}$$

The term on the left is the gain to the physician of k percent of the $\left(\frac{NQ}{M}\right)$ services provided by each of the $M - 1$ other physicians. The term on the right is the loss to the physician resulting from less demand among his current set of customers. Simplification of the inequality indicates that an increase in accuracy will increase income if $kQ\,(M - 1) > \partial Q/\partial A$.

If k and $\partial Q/\partial A$ are assumed to be constant, the expression obviously implies that accuracy will tend to increase as the number of physicians (or competitors) increases. But while there is no obvious reason why $\partial Q/\partial A$ should change as M changes, there are some reasons why k might vary with M. On the one hand, standard search models would suggest that, if accuracy can be measured with certainty by just one search, k would be likely to increase with M. The more sellers, the lower the average cost to the consumer of getting an "accuracy quotation" from another seller, and consequently the more attractive will be a higher accuracy seller. There is a second possibility, however, if the level of accuracy is not measured perfectly. Given some total number of experiments or experiences, the consumer's estimate of the accuracy of a *particular* physician's diagnosis will be likely to depend on the total number of physicians in his market area. The more physicians, the lower the average number of experiences the consumer or his friends are likely to have had with any given physician, and hence the less precisely he can estimate the quality or accuracy of any individual physician's diagnosis. If there is only one physician in town, all of the consumer's past experience will have been with that physician. If there are hundreds of physicians, the *average* accuracy of his estimate will be less (although he could have a very accurate estimate of the diagnostic capability of any individual physician). If the consumer is risk averse, or if he has difficulty remembering the identities of multiple sellers, he may be less willing to leave his present seller (whose accuracy he probably knows fairly well) for another physician whose accuracy he can estimate only imperfectly. To take two extremes: in a town with only a few physicians, differential levels of accuracy will quickly become well known. But in a large metropolitan area, it may be very difficult to find several friends who use the same physician, so that it will be impossible to get an accurate measure of the level of accuracy of any alternative seller. Consequently, it is possible that k will tend to be larger in small towns than in large metropolitan areas; k may well decrease as the number of physicians increases.

An availability effect will be nonexistent, or small and difficult to detect, if D_{max} is close to D_T. What do the observations above imply about measurement of the availability effect? They suggest that the position of the D_{max} demand curve is likely to vary with the number of phy-

sicians (*not* physicians per capita) in the market area. The relationship between number of physicians and D_{max} cannot, however, be predicted a priori: D_{max} can increase, decrease, or remain constant as M changes. The relationship depends upon the sign and magnitude of the rate of change in k with respect to M. Moreover, precise measurement of the extent of market area is probably impossible, although there have been some fairly successful attempts to use proxy measures.[10] It does seem possible, nevertheless, to characterize extreme cases. In rural areas, at one extreme, the level of consumer information is likely to be fairly high, so it is possible that D_{max} will be fairly close to D_T—little or no availability effect will be observed, because it will not pay physicians to reduce their level of accuracy when some consumers are able to detect the reduction.[11] In large metropolitan areas, at the other extreme, it is likely that k will be low, possibly near zero, because few of the consumer's friends are likely to be able to provide information on the same physician's accuracy. Consequently, D_{max} will lie far to the right of D_T, and the "standard" availability effect analysis developed above will apply.

While these statements are obviously highly conjectural, they do suggest that, in the absence of accurate measures of the number of sellers in the consumers' market area, it may be desirable to estimate separate availability effects for rural areas, large metropolitan areas, and other areas. Such a procedure is followed in the empirical work described in the next chapter.

Incentives Under a Fee-for-service System

Before going on to discuss the outcomes produced by the physician incentives present in the existing fee-for-service system, I will digress to examine fee-for-service under alternative fee schedules. The goal here is to see whether there is something inherent in the fee-for-service concept which leads to departures from accurate advice. One of the results from the analysis earlier in this chapter that deserves special emphasis is the following: if the physician's notional supply curve intersects that consumer's demand curve which corresponds to true and accurate information, then the price at which this intersection occurs is a price which will induce the physician to tell the truth, and act as the consumer's agent. In this section I wish to expand on this notion, to extend it to a model in which there are multiple kinds of outputs supplied by the physician, and to use it to analyze the possible superiority of prepaid group practice or Health Maintenance Organizations (HMOs) over fee-for-service medicine.

With respect to the choice among competing methods of treating a given illness, it is often argued that the fee-for-service physician has an

incentive *not* to provide true information and *not* to recommend to the patient the utility-maximizing course of treatment. In particular, the fee-for-service physician, it is argued, will provide too much care as well as unnecessarily expensive forms of care.[12] I explored the issues of cost-minimization for a given level of health in chapter 1, and I argued that (1) the income-maximizing physician may give distorted advice on the level of health to be achieved and (2) if fees are not free to vary, the mix of treatments may not be the one which minimizes the cost of whatever level of health is produced. Here I shall examine the possibility of choosing a schedule of fees to minimize both kinds of distortion.

With respect to fee-for-service versus HMOs, there has been much discussion of the apparent fact that the fee-for-service system of payment tends to encourage additional use of medical care, especially hospitalization, as compared to capitation-salary schemes. Often this finding is extended to a comparison of fee-for-service in general with capitation in general, and often, too, it is argued that the additional use is unnecessary in some normative sense. What I wish to suggest here, however, is that these characteristics of the present fee-for-service system are not, in themselves, evidence against the concept of paying for medical care (or any other good) on a fee-for-service basis. If there are failings in the present system, they may stem *not* from the fee-for-service system as such, but primarily from the present level of actual fees. A possible misallocation of manpower is a secondary (and related) cause. It is not fee-for-service as a concept that is faulty, but rather some changeable characteristics of the present fee-for-service system. It is therefore wrong-headed to argue for a general preference for capitation or salary as opposed to fee-for-service.

In fact, it is possible to establish a stronger result: not only can fee-for-service be made as "good" as capitation, from a theoretical viewpoint, it can be better than any feasible form of capitation. Put another way, there exists some set of fees for specific services which will achieve a pattern of incentives for physicians that, at worst, provides him no incentive to depart from an agency role. Any preference for patient welfare, however slight, would then be sufficient to cause him to act as a perfect agent. These theoretical considerations do not, of course, bear directly on the question of the merits of existing systems, since each may be at different places in the best-worst continuum. They do suggest, however, that even if the present fee-for-service system is not perfect, it is at least perfectible. Accordingly, I shall go on to consider (1) why the present system might not have achieved the ideal pattern of fees and (2) what alterations might improve matters.

Initially, I will assume that the prices (fees) received by physicians are given. This assumption is useful analytically, and may not be too unrealistic when third party payers are involved. Suppose that the total

hours physicians will work is given, and suppose physician utility depends on income received and patient well-being. (Physicians do not prefer to do some tasks more than others.) The concern for patient well-being can be very small, in the sense that the physician would be willing to sacrifice a great deal of patient well-being for a very small increase in income. All that is required is that, given a set of actions which yields equal money income, the physician will choose that course of action which most benefits his patients. The question now is: What is the set of prices for individual services the physician prescribes which will cause him to behave in a way which maximizes patient welfare?

The intuitive basis for the answer to this question is clear: if the physician makes the same amount of income no matter how he spends his time with patients, he might as well choose the way which most benefits his patients. That is, a set of prices which equalizes physician net income per hour worked will permit physician utility maximization and patient welfare maximization to coincide.[13]

There is still the question of how total physician hours worked are determined. If physician utility depends on income and leisure, and the other usual conditions are satisfied, then there will be a supply curve of total hours worked for each physician. Given the number of physicians in practice, there will be some overall level of net income per hour worked which will maximize patient welfare.

Contrast the pattern of output which physicians would want to supply under this scheme with that under a scheme in which net income per hour is not equalized. The income-maximizing physician will choose to produce that type of output which yields the highest returns per hour worked. If there were no other constraints, this would be the only type of output he would produce. But there may be limits on the amount of this type of output he can persuade people to take. Then he may produce the next most profitable output, or he may even produce some complementary lower-yield outputs (e.g., initial office visits) in order to be able to persuade consumers to take the higher yield output.

If the production process for each output displays constant returns to scale, then the level of net income from any output will depend on the price of output and the price of inputs, but not on the particular amount of output. If there are increasing costs for a given output, price can be used to select any particular quantity or mix of output. In the first case, the most that can be accomplished is what might be called "incentive neutrality": the physician would be indifferent toward producing the ideal mix of output and some other combination of those outputs which are to be produced in positive quantities. However, if the physician has the slightest preference for doing what is in the patient's best interest, or if there is the slightest chance that he will lose patients in whose interests he does not act, then he will behave as a perfect agent. In the

second case, the total quantity of each type of output produced by each physician is determined by the price, but not the mix provided to each patient. Again, a slight degree of concern for patient welfare or a slight extent of competition is all that is needed for the optimal outcome to be an equilibrium outcome.

Compare these ideal price incentives to those under a capitation (HMO) system. Total revenue is fixed, so real income maximization provides incentives to minimize total inputs, whether physician or nonphysician. For the income-maximizing (or cost-minimizing HMO), there are incentives to provide a small amount of total output, even when that would not be in the patient's own interests, and to select cheaper mixes of outputs. Competition may constrain the ability to underprovide somewhat, but it is unlikely that it will be perfect enough to prevent it entirely. This is not to say that an HMO equilibrium may not be preferable to existing fee-for-service prices. However, there exists a set of fee-for-service prices, not necessarily existing ones, which can always do as well or better, in terms of providing incentives for agency, than any capitation scheme. For a capitation scheme to achieve even as good an outcome as fee-for-service, perfect competition is required. Perfect competition is *not* required under the ideal set of fees-for-service.[14]

Comparisons between existing fee-for-service and HMO systems are made difficult because of an ambiguity in the notion of agency, which has been noted above. With insurance that is not individually experience rated, the individual patient is best off by using care as long as the value he places on the care exceeds the user cost he pays. Since all patients must pay collectively the full cost of care, all may be better off if they keep use below this level. But which of the two levels of output is the one the physician should choose in his role as agent? One of the advantages of the HMO may be precisely that it does cause the physician to behave *not* as the agent of his own patients, or of the fraction of the membership he treats, but rather as agent for the entire membership group.

Can this same kind of "group agency" be attained under fee-for-service? If the physician does completely control what happens to the patient, there is still some set of fees that can be found to induce him to supply exactly what he would supply if he were acting as agent for all patients. There is a stronger result, however. Given such a set of prices, the physician will provide this ideal amount of output to patients regardless of what individual patients do. Patients may want their physician to provide more output, but income-maximizing physicians will refuse to do so. Rather than levying user charges—copayments or deductibles —to discourage moral hazard, there is a scheme of reimbursement under fee-for-service that will have exactly the same effect as long as physicians are concerned, even if only slightly, about the aggregate well-

being of their patients. If the fee structure is incentive neutral, physicians may even take differences in patient preferences into account to some extent, since doing so will improve patient well-being.

This abstract discussion can be illustrated with some examples. Consider those commonly cited examples of "overprovision" under fee-for-service: surgery and hospitalization in general. The physician's net income per hour worked is likely to be greater if he performs surgery than if he does not. It is therefore not surprising that there may be "excessive surgery" under fee-for-service, or that this rate can be reduced by capitation. What is not generally recognized, however, is that with a sufficiently low relative fee for surgery, this incentive could also be made to disappear. The emphasis here is on *relative* fees; an alternative to cutting fees for performing surgery would be to raise the fee for other procedures, such as consultation.

A second way of reducing marginal net real income (MNRI) is to raise opportunity costs; in addition to taxing the use of inputs, this could be done by increasing the number of hours worked per physician. Fewer surgeons should lead to incentives to perform less surgery, because the opportunity cost of each surgeon's time—the value of lost leisure—will be greater if he has less leisure. It is interesting to note that HMOs typically follow both of these strategies: they pay salaries or profit shares, which may imply zero or negative real net income from additional surgery, and they hire fewer surgeons relative to population served.

If a physician hospitalizes a patient for treatment of a given condition, he usually benefits in three ways: (1) he may be able to charge (and have insurance cover) a higher fee for procedures performed in hospital; (2) it may take less of his time to see patients in the hospital rather than in the office or at home, and yet his fee for a follow-up visit may be the same or greater; (3) as indicated in chapter 3, he may be able to substitute insured hospital inputs for uninsured inputs which he would provide. All of these reasons lead to a higher MNRI from hospitalization. The cure, again, is some reduction in MNRI—paying the same for a procedure regardless of where performed (or even possibly less if the procedure is performed in the hospital), paying less for visits which take less time, and perhaps some method of charging the physician for additional hospital inputs.[15] Where the user price of hospital treatment is below that of ambulatory treatment because the insurance only covers inpatient expenses, then the patient will still have an incentive to seek overprovision, even if physician incentives are rationalized. In such a case, as noted above, the notion of agency is ambiguous, and conflict between physician and patient may arise. Here moral hazard cannot be controlled directly. But at least for those outputs equally covered by insurance, restructuring of fees to yield, on average, approximately

equal MNRI might go a long way toward reducing some of the abuses under fee-for-service.

Why Are Fees Out of Line?

In order for such a policy suggestion to succeed, one needs to have some idea of the forces which lead to the initial pattern of fees. Why *are* the MNRI per hour different for different outputs? One would have expected that the return to any input in different uses would tend to be equalized. Suppliers would devote time to the higher-yielding output, causing its price to fall and the price of other outputs to rise. Why hasn't this happened?

For those outputs whose physician fees are covered by insurance, it is easy to see why prices do not fall. Cutting price would not increase a surgeon's volume of business, even if he wanted to produce more output. This would still not prevent equalization if there were an unlimited amount of insured business. But there clearly is not enough surgery to permit all physicians who are licensed to perform it to do so.

Why do insurance plans not permit the fees they pay to fall? Perhaps it is because some physician insurances are provided by physician-dominated firms. To provide a complete explanation here would take us far from the main line of our argument. The only purpose here is to point out that insurance fee schedules may be the ultimate cause of overprovision of many physician services.

At a policy level, there are three broad strategies for remedying overprovision under fee-for-service. MNRI can be reduced either by cutting price or by raising unit cost. A cut in fees paid by third parties would accomplish the former, while a reduction in the number of physicians in specialties in which there is too much output would accomplish the latter. A third strategy involves making the demand for physician services used to produce insured outputs more elastic. Payment of indemnities, for example, makes price cutting pay, and so may lead to lower prices.

In summary, it is not fee-for-service as such that yields incentives for overprovision; rather, it is the level of existing fees for some procedures. Change the fees, either by lowering some or by raising others, and the overprovision will disappear. An additional corrective is that the total stock of physician input must be efficiently utilized; otherwise the specific examples of overuse we now see may be converted to a general problem of overuse of all physicians' services. Somewhat surprisingly, writers on the subject of incentives have generally ignored the flexibility available under a fee-for-service system. Monsma has given perhaps the most extensive treatment of the effect of what he calls "marginal revenue" on physician choice of output, but he has only a confused footnote

exploring the possibility of how the MNRI under fee-for-service might be reduced.[16] He correctly says that, since marginal gross revenue is zero under capitation, MNRI there must be lower than under fees which yield positive MNRI. This yields the prediction that output should be less under capitation, but it does not answer the objection that output under capitation may be too low. In the absence of perfect competition, fee-for-service payment can still provide ideal incentives.

Changing the Fee Structure

There are two related objections to the notion that fee-for-service, with radically restructured fees, could be used to provide optimal incentives. It may be argued that the changes in fees may be so large as to (1) be infeasible or (2) if feasible, impractical because of the large reductions in some specialists' incomes they imply. For example, given the current level of MNRI for surgery, equalizing returns from surgery and consultation would probably imply enormous fees for consultation —perhaps $200 for a half-hour visit. Up to a point, this change might result not in less surgery, but just more consultation, if excess surgical capacity is sufficiently large. Conversely, cutting surgical fees to equalize MNRI might involve hefty cuts. Even if these were feasible, the consequent reduction in surgeons' incomes might, again up to a point, induce them to desire to do *more*, rather than less surgery than at present in order to keep their incomes up to their accustomed levels. A still further reduction in fees could reduce this incentive, but at the cost of a still less palatable reduction in income.

There are two kinds of responses to these objections. One is to note that they only represent the obvious consequences of erroneous manpower policy. Paying surgeons to do consultation may look expensive, but it may be less costly than having them use their time performing surgery, given that too many surgical specialists have been trained. The second is to note that in the long run, "appropriate" levels of physician-specialist incomes can be combined with "appropriate" incentives only if "appropriate" numbers of specialists are trained. In the interim, one might be able to obtain consent to a substantial reduction in fees by proposing to supplement fees with lump-sum payments to those physicians in specialties thought to be in excess supply. In order to get from a "wrong" fee-for-service system to a proper system, it may be necessary to use partial salary-capitation reimbursement for a time, but always as a supplement to a fee-for-service system.