Competition Among Physicians, Revisited

Roger Feldman*, University of Minnesota, and* ***Frank Sloan****,*
Vanderbilt University

Abstract. Ten years ago we developed a model of demand inducement in the physician services market and explored the properties of that model. We found that predictions concerning physicians' prices, workloads, and income were ambiguous and in many cases were consistent with those derived from a standard monopoly pricing model. Spurred in part by our work, numerous empirical studies of the demand inducement model have been conducted. These studies found little evidence of demand inducement for primary care physician services. Demand inducement may exist in the market for surgical services, but its extent is less than previously estimated. We disagree with those who say that physicians generate demand to avoid price controls and that national health care spending is proportional to the number of physicians; the evidence does not support these arguments. Substantial uncertainty may surround the physician's choice of diagnosis and treatment mode. However, this does not imply a breakdown of the agency relationship. In this paper we extend our earlier model of demand inducement to include variations in the quantity of services (which was previously assumed to be less than socially ideal). Using the model, we conclude that the major objection to government price setting is not that physicians will get around the controls by inducing demand; rather, price controls result in a quantity and quality of physicians' services that is not ideal and may be inferior to those provided in an unregulated monopoly.

Ten years ago, following Robert Evans, we classified health economists into two groups—Ns (narrow economists) and Bs (broad economists).[1] The key issue dividing these groups was whether the demand curve for physician services is subject to shifts induced by physicians in pursuit of their own interests: the Ns argued that it was not and the Bs argued that it was. The issue is important to the analysis of competition in the health care industry. If doctors can generate demand for their care, they possess far more market power than is usually attributed to the monopolist, whose price-setting ability is constrained by a fixed demand curve.

This argument is still raging; indeed, the volume of papers published on physician-induced demand seems to be growing at an increasing rate. Interestingly, few of the participants in the debate show any sign of changing their positions.

Journal of Health Politics, Policy and Law, Vol. 13, No. 2, Summer 1988.

1. All references to our earlier paper are to Sloan and Feldman (1978).

Uwe Reinhardt, one of the leading Bs, suggests a particularly simple test of the inducement hypothesis: "Indeed, the present author has never yet met a physician who would deny the existence of that power (nor has he ever met one who would admit to exploiting it for personal gain)" (Reinhardt 1985: 189). Aside from their lack of methodological rigor, such simple views are contradicted by other "stylized" facts—organized medicine's traditional opposition to increased physician supply, the movement of physicians to towns previously not served by physicians (Newhouse et al. 1982), and recent "handwringing" by physicians in the face of both increased competition from competitive health plans and the greater aggregate number of U.S. physicians. The issue is not whether physicians have ever induced consumers to purchase a service that they would not have purchased had they been fully informed; rather, it is whether such added care is quantitatively important and whether the amount of induced demand varies systematically with changes in such variables as physician supply.

To make such assessments, it is essential that other influences on demand be held constant. Sophisticated econometric studies can be found to support both sides of the argument. One can look to the work of McCarthy (1985), who found that the market for primary care physician services in large SMSAs is reasonably competitive, and Cromwell and Mitchell (1986), whose findings on the market for surgery "provide definite support for the notion of competitive market failure—particularly in large metropolitan areas" (ibid.: 293).

This debate will not be resolved easily or soon. Therefore, it is appropriate to take stock of the inducement argument and other issues related to competition among physicians. In this paper we summarize recent empirical studies of physician-induced demand. In addition, we evaluate evidence that per capita utilization of services tends to rise when fees are controlled or when the physician/population ratio increases. Finally, we take note of studies that show large differences in utilization of physicians' services among small geographic areas.

Lest we be accused of hidden bias, we both admit to membership in the N camp. This was true ten years ago and is still true today. However, this does not mean that we reject further improvements in the theoretical framework we created ten years ago or in empirical studies of supplier-induced demand. In fact, we will expand the theory of inducement to include variations in the quantity of physicians' services. Both positive and normative aspects of this model will be considered simultaneously. In a sense, the model extends work started in our earlier paper. We conclude that the major objection to government price setting (by a fee schedule or a "physician DRG" system) is not that physicians will get around the controls by inducing demand, but that price controls result in a quantity and quality of physicians' services that are not ideal and that may be inferior to those provided in an unregulated monopoly.

Econometric studies of demand inducement

Ten years ago we developed a model of demand inducement in the physician services market and explored the properties of that model. We found that predictions concerning physician prices, workloads, and income were ambiguous and in many cases consistent with those derived from a "standard" monopoly pricing model. Spurred in part by our work, numerous econometric studies of demand inducement have been conducted in the past decade.

Direct tests of the inducement hypothesis. Suppose there is a change in the level of some variable (e.g., physicians per capita in the market area) that shifts the demand curve facing an individual physician. Standard theory (in the form which recognizes quality differences among physicians) predicts that such shifts will not affect per capita demand for physicians' services directly once the indirect effects operating through money and time prices have been controlled. Inducement theory is ambiguous, but it is usually assumed that an increase in doctors per capita (an inward demand shift) will cause physicians to induce demand. This is the most straightforward test of the inducement hypothesis, since any nonzero partial derivative of the physician/population ratio in the demand equation (once prices have been held constant) is inconsistent with standard theory.

Econometric studies using aggregate data generally support the direct inducement hypothesis. Two examples are found in Fuchs (1978) and Cromwell and Mitchell (1986). Fuchs examined the effects of the supply of surgeons on the demand for surgical operations in 22 geographic areas of the United States in 1963 and 1970. He argued that surgery was well suited for empirical analysis of inducement since the "time price" is relatively unimportant to surgical patients—that is, increased numbers of surgeons may reduce patient travel time and waiting time for surgical care. But travel time is generally a minor consideration relative to the financial price of surgical care, and waiting time for surgical services, at least in the United States, has typically been low. Fuchs found that a 10 percent increase in the surgeon/population ratio results in about a 3 percent increase in per capita utilization. Cromwell and Mitchell estimated a model of demand and equilibrium fees for surgery with data from about 250 SMSAs (or collections of rural counties) from 1969 to 1976. Other things equal, utilization was higher in surgeon-rich areas, although the estimated demand-shift elasticity was about one-third of that found by Fuchs.

Both the Fuchs and the Cromwell/Mitchell studies used an instrumental variable approach in which the physician/population ration was endogenous. The most severe critics of this approach are Auster and Oaxaca (1981), who point out that the underlying model being tested is

$$Q^d = D(P, X^d, Q^s), \tag{1}$$

where Q^d is quantity demanded per capita, P is price, X^d is a vector of exogenous demand-shift variables, and Q^s is quantity supplied. This is in contrast to the standard model, which excludes quantity supplied from the demand equation:

$$Q^d = D(P, X^d). \tag{2}$$

The problem is that these two equations are indistinguishable in equilibrium, where $Q^d = Q^s$. For example, let equation (1) be linear:

$$Q^d = \alpha_1 P + \alpha_2 X^d + \alpha_3 Q^s. \quad (1a)$$

Using the equilibrium condition, equation (1a) can be written as

$$Q = (\alpha_1/[1-\alpha_3])P + (\alpha_2/[1-\alpha_3])X^d. \quad (3)$$

An estimate of equation (3) is indistinguishable from the standard model implied by equation (2).

To circumvent this problem, the Fuchs and Cromwell/Mitchell studies used an input from the supply side (surgeons per capita) to shift demand. However, inputs and outputs are likely to be strongly correlated. Thus the identification problem cannot be solved unless it can be shown that multiple inputs are used to produce output (a condition that is necessary to distinguish the supply of surgeons from the number of surgeries) and that the relative prices of these inputs differ across markets. The surgeon and his surgical assistant are clearly the dominant inputs in the production of surgical care. Little variation in factor price ratios could preclude efficient estimation of the market supply relationships.

Neither of the studies reviewed here addresses the question of how surgical operations are produced. Regarding the identification of market supply, Fuchs's model did not incorporate any factor prices pertinent to the production of surgical operations. Cromwell and Mitchell used a vector of cost and productivity-influencing variables, but this vector was limited to two variables: hospital beds per capita and the hourly retail wage (the latter was assumed to measure the wage of nonphysician inputs). Most of the variables explaining physician density were measures of professional amenities and measures of the area as a place to live. Not surprisingly, when the variable for estimated number of surgeons per capita was used in an equation to determine equilibrium fees, many of the estimated coefficients had implausible signs. For example, age-specific dummy variables indicated that price was inversely correlated with the age of the population, even though age was a positive demand-shift factor in the estimated utilization equations. The authors suggested that part of the problem was due to collinearity between the predicted surgeon supply and the included exogenous variables. This is exactly the problem that Auster and Oaxaca raised in their critique of the instrumental variable method.

In contrast to studies using aggregate data, those based on observations of individual physicians or consumers indicate little evidence of demand inducement. For example, Rossiter and Wilensky (1983) analyzed differences in the use of physicians' ambulatory care services and the extent to which these differences are attributable to the physician, the patient, or both. For services identified as physician-initiated, most of the variation was due to patient characteristics that reflect medical need, such as disability days, perceived health status, and the presence of chronic health conditions. Variables related to the physician's financial interest, such as physician density, had very little effect on demand. In another study, Rossiter and Wilensky (1984) concluded that physician-initiated expenditures for ambulatory care and for all medical services could not readily be explained by the inducement hypothesis. When health insurance and other factors were held constant, the magnitude of physician inducement was very small and was statistically significant only for more discretionary procedures.

More recently, McCarthy (1985) estimated demand curves for individual primary care physicians in large SMSAs. From the demand regressions McCarthy found that consumers are very responsive to prices charged by individual physicians. The elasticity of demand with respect to physician density was very large in absolute value (between −2.61 and −3.64). This contrasts not only with demand inducement theory, which predicts that this elasticity should be less than 1 in absolute value, but also with standard theory, which predicts an elasticity of −1.[2] McCarthy suggested that these results are evidence of a binding demand constraint that, at the margin, limits the inducement activities of individual physicians. But with an elasticity of −3, an annual increase in doctor supply of 3 percent would lead to decreases in per doctor workloads of 9 percent per year—hardly a trivial decrease, especially when repeated annually.

Studies using microdata appear to sidestep Auster and Oaxaca's criticism, since it can be argued that physicians per capita are exogenous in equations explaining the behavior of individual consumers or physicians. However, this escape from the endogeneity problem is more apparent than real. Suppose, for example, that doctors per capita in the jth city ($MDCAP_j$) are a linear function of price (P_j), quality of life in that city (L_j), and an error term (u_j):

$$MDCAP_j = \beta_1 P_j + \beta_2 L_j + u_j. \tag{4}$$

Let average quantity demanded per doctor depend on price and physician density (this is the equation estimated in studies using aggregate data):

$$Q_j^d = \alpha_1 P_j + \alpha_2 MDCAP_j + \epsilon_j. \tag{5}$$

Suppose, further, that quantity demanded from the ith physician in this city depends on the deviation between that physician's price and the average price:

$$\begin{aligned} Q_{ij}^d &= Q_j^d + \alpha_1(P_{ij} - P_j) + \epsilon_{ij} \\ &= \alpha_1 P_{ij} + \alpha_2 MDCAP_j + v_j, \end{aligned} \tag{6}$$

2. In standard theory, when price and demand shift variables are held constant, a 1 percent increase in physicians per capita should reduce the services demanded from each physician by 1 percent, assuming that quality does not change.

where $v_j = \epsilon_j + \epsilon_{ij}$. This is the equation estimated by McCarthy, with $MDCAP_j$ assumed to be exogenous.

However, the model also has an equilibrium condition that demand equals supply. The simplest possible supply assumption is that supply per doctor in the market area is fixed at Q_j^s. Thus the equilibrium condition is $Q_j^d = Q_j^s$. Using this condition and equations (4) and (5), we can solve for doctors per capita as a function of exogenous variables and the city-specific error term:

$$MDCAP_j = \ldots\ldots - (\beta_1 / [\alpha_1 + \alpha_2\beta_1])\epsilon_j. \qquad (7)$$

The last term is positive, indicating that stronger unmeasured community demand conditions lead to stronger demand per doctor and more doctors per capita. As a consequence, the estimated coefficient on the physicians per capita variable in McCarthy's equation (6) is biased toward 0. Therefore, it is surprising that the absolute value of his estimated coefficient is as large as 2 or 3, let alone the value of 1 predicted by standard theory.

Our suggestion for future studies of demand inducement is that physicians per capita cannot be regarded as exogenous, even in studies that use microdata. One way to solve this problem is to use an instrumental variable for $MDCAP_j$, but this throws us back on Auster and Oaxaca's criticism. An easier correction is to use a fixed-effects model in which influences of unspecified pertinent community characteristics are captured by binary explanatory variables for each community. This would have been especially suited to McCarthy's data, which came from a few large cities. The addition of city-specific dummy variables as proxies for the ϵ_j errors would have left physicians per capita uncorrelated with ϵ_{ij}, and thus equation (6) could have been estimated with observed physician supply on the right-hand side.

Physician prices and consumer information. Standard theory predicts that the price of physician services (or, in more sophisticated versions of standard theory, the "quality-adjusted" price) should fall following an increase in the number of physicians per capita. Demand inducement theory, as we pointed out ten years ago, is ambiguous about this prediction, but it usually assumes that price rises as physicians attempt to recoup their lost demand. As we discussed in our earlier paper, some studies found a positive association between physician density and the average level of fees. But we also pointed out that these studies are flawed because they fail to control for one or more of the following variables: the mixture of services provided (complexity, amenity level, etc.), the effect of nonmonetary factors such as travel and waiting time, and the tendency of physicians to migrate to areas with higher fees.

The importance of the first variable is very simple to explain. Producers in large markets are more specialized than those in small markets—for example, highly specialized physicians tend to congregate in large metropolitan areas. It is observed that these areas have high fees and more doctors per capita, but this association is due to the specialized services produced by urban physicians. Attempts to control for product differentiation, such as the distinction between "elective" and "nonelective" surgery used by Cromwell and Mitchell, only begin to capture the extent of endogenous specialization.

The second variable is more subtle. Patients in areas with more doctors should have easier access to physicians in terms of lower travel and waiting times. These nonmonetary factors will encourage them to use more services, which, other things equal, tends to increase the monetary price of services. Some evidence on this effect comes from a study by Feldman and Ballard (1981), who examined office waiting times in the British National Health Service. Feldman and Ballard found that areas with more doctors had significantly lower waiting times than areas with fewer doctors. Using this estimate, they calculated that the entire effect of physicians on demand for services observed in the United States (Fuchs and Kramer 1972) could be explained by easier access to medical care in areas with more doctors, without recourse to demand inducement.

The third variable will lead to higher fees in areas with more doctors per capita. However, it is important to note that the high fees cause doctors to enter the area, not (as inducement theory suggests) that more doctors cause higher fees.

The only serious theoretical attempt to explain the positive association between price and physician density outside the framework of the standard model was made by Pauly and Satterthwaite (1981), who argued that consumers have more difficulty finding information about a particular doctor in physician-rich areas. Consequently, monopoly power is enhanced, theoretically leading to higher prices in such areas. Cromwell and Mitchell found no support for this model, however. Phelps (1986) has also trenchantly criticized the search model implied by Pauly and Satterthwaite's approach to the market. Pauly and Satterthwaite assumed that the patient's search strategy is to select a doctor at random and then ask friends for information about the doctor. The quality of information obtained by this strategy clearly declines as the number of doctors increases. Phelps suggested that a more appropriate and realistic search strategy would be to ask one's friends (or those who have similar tastes for styles of treatment) for the name of their physicians. Under this search strategy, which is at least as plausible as the one specified by Pauly and Satterthwaite, physician density should not affect information-gathering costs.

The lasting contribution of Pauly and Satterthwaite's model may be to focus future studies on physician market power on the critical role played by information. In fact, Pauly (1978) had previously called for studies that would document whether consumers were well informed about medical care. In the same paper he emphasized that consumer information about frequently consumed physicians' services (such as primary care) is likely to be much better than information about infrequently consumed services (such as surgical operations). For some services, such as many forms of cancer therapy, there is considerable uncertainty on the provider side about the proper course of treatment. Too many researchers—and policymakers, for that matter—have overlooked such important distinctions.

A few studies have been conducted on consumer information, both before and after the publication of Pauly's 1978 paper. For example, Bunker and Brown (1974) assessed the extent of excess surgery by comparing the rates of surgery performed on a group of Stanford University Medical School faculty and their spouses with those performed on a group of attorneys, Protestant ministers, and graduates of Stanford Business School and their spouses. Presumably, if physicians give self-serving advice, it should be detected more often and accepted less often by their peers. However, Bunker and Brown found that physician patients and their spouses had a 20–30 percent higher total rate of operations than the general population and a rate of "nonessential" procedures greater than the three professional control groups.

A similar conclusion was reached in a study by Hay and Leahy (1982), who found that medical professionals and their families were as likely, if not more likely, to visit physicians as were other families, controlling for variables related to perceived health status, access to care, and ability to pay.

If the present studies show that the general public is no more likely to be bamboozled by demand inducement than physicians are, then the goal of future research in this area should be to identify market structures that will produce accurate information. Some of the N economists believe that a competitive medical care system would produce more reliable information than the present one (Office of the President 1985). For example, only 29 percent of the participants in the RAND Health Insurance Experiment realized the falsity of the statement: "If you have to go to the hospital, your doctor can get you into any hospital you prefer" (Newhouse et al. 1981). When the same statement was presented to 5,000 employees in Minneapolis, where many employees have a choice among competing health maintenance organizations (HMOs), researchers found an appreciably higher percentage of correct answers (Dowd et al. 1984). This finding suggests that consumers in Minneapolis were aware that choosing a closed-group HMO limits one's ability to choose any hospital.

The target income hypothesis. An extreme form of the demand inducement model combines inducement with the assumption that physicians shift demand until they achieve a satisfactory, or "target," level of income. This model was examined by Sweeney (1982), who showed that more competitors can lead to higher prices (assuming, for the moment, that physicians do not induce demand) only if the demand curve cuts the downward-sloping supply curve at the doctor's chosen equilibrium point. Figure 1 displays two panels with identical demand curves. In the left-hand panel the "supply curve" (that is, all combinations of price and quantity that yield the target income) cuts the demand curve at points A and B. If the doctor chooses point A, then an increase in competitors (an inward

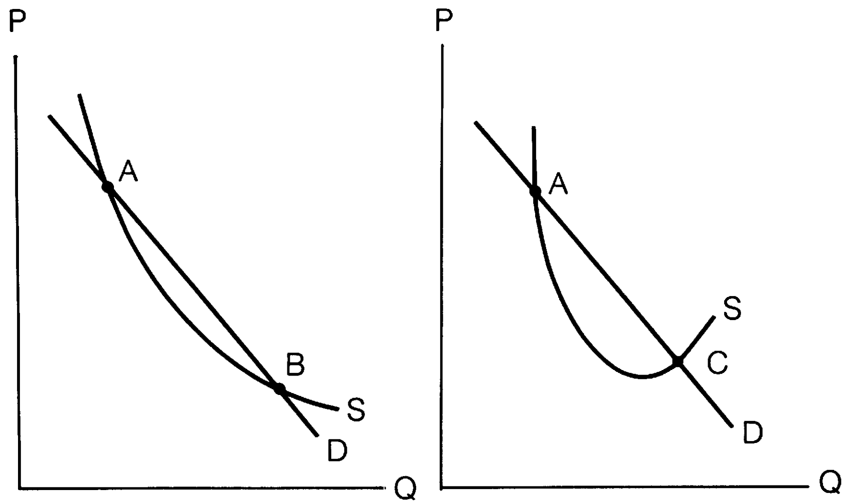

Figure 1. The Target Income Model Without Demand Inducement

shift in demand) will cause price to fall. Likewise, both points A and C in the right-hand panel are consistent with a negative correlation between physician density and price. Only at point B are predictions inconsistent with standard theory. Sweeney was relatively noncommittal regarding which of the possible equilibrium points will be chosen.

When inducement is added to the target income hypothesis, physicians will shift demand as much as they can in order to hit the target with the least possible work. Further exogenous increases in the number of doctors will result in lower prices, which is the standard result from a model without demand inducement. Alternatively, one can assume that demand shifting is constrained, at the margin, by monetary or psychic costs. But this assumption makes the model equivalent to Sloan and Feldman's analysis, which showed that price may rise or fall when the number of physicians increases. Thus the target income model is theoretically sterile: either it cannot be distinguished from the standard monopoly approach to physician pricing or it adds nothing to the monopoly model with inducement analyzed by Sloan and Feldman.

Sweeney (1982) suggested that the problems of interpreting supply or demand equations under conditions of target-income pricing and inducement can be avoided by directly estimating the physician's target income and the probability of hitting that target. He found that physicians hit the target more than half the time in only 9 of 30 markets, the majority of which were likely to be small and rural.

Interesting and novel as Sweeney's estimation approach and results are, they raise more questions than they answer. In particular, his empirical work seems to lack any concept of long-run market equilibrium. If doctors cannot hit their target income even after shifting demand, why do they not move to another area? The only conceivable long-run equilibrium is at the tangency of supply and demand, which must occur on the downward-sloping portion of the supply curve. All changes in physicians per capita should then be viewed as endogenous responses to changes in the given conditions of supply and demand.

Responses to fee controls

The Bs have argued that fee controls are fruitless, since when price is constrained physicians can maintain their incomes by boosting demand and hence increasing the number of billable services. The Bs cite evidence from the United States and Canada to support this view. But as we will demonstrate, the evidence is ambiguous as to whether supplier-induced demand actually occurred. We will examine physician responses to fee controls from a conceptual standpoint below; here we look at the empirical evidence.

The Economic Stabilization Program (ESP), instituted by the Nixon administration and covering the period of August 1971 through April 1974, applied price controls to all sources of patient revenue to physicians. After a short price freeze, ESP allowed prices to rise but limited the annual increase to 2.5 percent. In addition, Medicare placed limits on the growth of reasonable charges.

In a comprehensive evaluation of bills submitted by California physicians to Medicare during ESP, Holahan and Scanlon (1978) found that while price controls succeeded in limiting the rise in actual and reasonable charges per unit of service, there was an appreciable increase in the number of services billed and in the complexity of services. As a consequence, payments to physicians in the California sample increased markedly during ESP.

One explanation for this behavior is that physicians stimulated demand to maintain their incomes. There are alternative explanations, however, several of which are consistent with standard economic models. First, as discussed below, imposing a binding price constraint on a seller with monopoly power may lead to increased output. Since demand for physicians' services increased during this period (Sloan and Schwartz 1983), price probably would have risen at more than the allowed rate of increase. Thus the controls probably kept price lower than it otherwise would have been.

Second, placing a limit on office visit fees encourages physicians to substitute diagnostic tests for their own time so long as the price of such tests covers marginal cost (Munch 1980). Such tests are often performed in the physician's office and are thus billed by the physician.

Third, ESP may have stimulated physicians to game the system by relabeling services provided—e.g., by reporting an "intermediate" visit to the payer rather than a "brief" one. Physicians can always do this, but before ESP, when fee increases were not controlled, it was less costly to raise prices than to spend time

Figure 2. Billings per Physician in British Columbia, 1968–1985 (Constant 1971 Fee Dollars)

Source: Wilson v. Medical Services Commission of B.C., 9 B.C.L.R. 350 (1987), statement of evidence filed by Morris Barer and Robert Evans.

figuring out ways to outsmart insurers. ESP decreased the relative cost of gaming and thus may have encouraged such behavior.

Fourth, Holahan and Scanlon presented price and output data only for Medicare and Medicaid. If physicians shifted output from private to publicly insured patients, it would appear that ESP controls increased output, while all they actually did was influence output shares by payer type.

Evidence from Canada (where the output data apply to all sources of payment) lends support to the importance of examining total output rather than output components. Here the evidence is conclusive: fee controls worked. Fees were controlled under universal health insurance plans introduced by Canadian provinces during the late 1960s and early 1970s. Physician expenditures as a percentage of GNP in Canada were 1.21 percent in 1970, 1.16 percent in 1975, and 1.10 percent in 1980. For British Columbia only, the respective figures were 1.36 percent, 1.26 percent, and 1.28 percent.

Similar evidence can be seen in time series data for British Columbia from 1968 to 1985, which is shown in Figure 2. Real billings per physician have been divided by a fee index so that they are expressed in terms of services per physician. The most noteworthy features of Figure 2 are three dips that occurred in 1972–1973, 1976–1978, and 1983–1985. The first two declines are steep and short, whereas the third represents a more gradual but continuing trend. These three dips all coincided with periods of government efforts to limit physician fees. In 1972–1973 and 1976–1978, the profession was subject to wage and price

Table 1. Percentage Increase in Physicians per Capita and Annual Percentage Change in Physician Income, Canada, 1960–1979

Time Period	Annual Percentage Increase in Physicians per Capita	Annual Percentage Change in Physician's Income Relative to Average Worker's Income
1960s	1.78%	2.75%
1969–1974	3.60	−3.65
1974–1979	1.55	8.88

Source: Brown (1982).

controls. During these periods costs went down. More recently, average fees received in British Columbia rose only 2.8 percent from 1983 to 1984, and not at all in the next year. Fee negotiations in early 1985 placed a cap on total expenditures, which would reduce fees if total billings per capita rose more than 2 percent annually.

These comparisons show that the fee controls *are* an effective means to limit spending on physician services. The problem with fee controls is not that they do not constrain expenditures, but that the prices established by fee controls bear no relation to competitive prices that would allocate resources efficiently, both among the different physician services and between the physician services sector and other sectors of the economy. We will say more about this point later.

Other evidence from foreign countries

Advocates of the inducement hypothesis suggest that evidence from Canada and Europe proves that per capita utilization of physician services tends to rise in step with the physician/population ratio. This statement would imply that physician incomes are unaffected by the supply of physicians. However, this argument is contradicted by the simple ''stylized'' facts from Canada, which show that physicians' relative incomes tend to decrease when their numbers increase rapidly (see Table 1). Analyzing data from ten Canadian provinces from 1957 to 1971 and controlling for the effect of real per capita income, a time trend, and the existence of a provincial hospital insurance program, Schaafsma and Walsh (1981) found that real income per physician starts to fall when the number of physicians per 100,000 population exceeds about 80. In 1973, this number was exceeded in all Canadian provinces.

An even worse argument is that a nation's total health bill is determined in a simple proportionate way by the number of doctors. As shown in Table 2, the financial burden of universal health insurance in Canada was not related to the number of physicians per capita. For example, Manitoba and Nova Scotia both had ratios of 200. In Manitoba, 8.8 percent of GDP was devoted to health care;

Table 2. Physician Density and Health Care Spending, Canada

Province	Physicians per 100,000 Population, 1983	Health Care Spending as a Percentage of Gross Domestic Product, 1981
Alberta	167	5.63%
Saskatchewan	157	6.60
Manitoba	200	8.80
Ontario	205	6.91
Quebec	204	8.49
New Brunswick	131	10.62
Nova Scotia	200	11.89
Prince Edward Island	120	13.41
Newfoundland	168	12.35
British Columbia	204	7.50

Source: Health and Welfare Canada (1984a, 1984b).

in Nova Scotia, the health care share was much larger (11.9 percent). Other factors must be at work.

On a worldwide scale, efforts to correlate total health care spending and the number of physicians per capita are also fruitless. We examined variations in health care spending as a percentage of GNP in 1980 for seventeen OECD countries and the province of British Columbia. The estimated regression equation was:

$$\underset{}{\mathrm{HCC}} = \underset{(2.168)}{6.91} + \underset{(0.428)}{0.00463\mathrm{DOCS}} - \underset{(0.183)}{0.0415\mathrm{BEDS}} \quad r^2 = 0.02 \tag{8}$$

where HCC is total health care costs as a percent of GNP in 1980, DOCS is physicians per 100,000 population, and BEDS is acute care hospital beds (excluding psychiatric beds) per 1,000 population (source: OECD Data Bank, National Statistical Yearbooks, 1980–1982). Estimated t statistics are noted in parentheses below each coefficient. Clearly each country is unique. Doctors and hospital beds per capita have little influence on health care spending as a percentage of GNP.

Evidence from studies of small area variations

Numerous studies indicate that the rates of surgical and medical procedures differ substantially among small areas such as individual communities within a county or state. The study of these variations, now known as small area analysis, has been described by Wennberg et al. (1982: 813):

> In a typical small area analysis the smallest demographic units (such as zip codes of minor civil divisions) are assigned to hospital areas, based on the

location of the hospital with the plurality of hospital admissions from the demographic unit. The resultant populations, which usually range in size from 10,000 to 100,000 persons, are sufficiently large to obtain statistically "stable" procedure rates, particularly for common procedures. The numbers of physicians whose workloads contribute substantially to each area are usually small. Consequently, the physician contribution to the area rate for a particular procedure is determined by the decision of a small group of physicians regarding the indications for the procedure.

When factors which are rough proxies for illness rates (such as age and sex composition) are controlled, physician choices concerning treatment patterns have been said to be the dominant force influencing the use of selected procedures in the population. For example, Wennberg and Gittelsohn (1973) showed that demographically similar populations living in areas served by more general practitioners who do not perform surgery have fewer surgical admissions than areas served by general practitioners who do perform surgery.

Wennberg et al. (1982) interpret this evidence in a way that chides both N and B economists. They dismiss the assumption of rational agency on the part of physicians (at least in the current environment), but believe that the Bs have misinterpreted physician behavior as a "self-serving, deviant professional response to economic incentives" (ibid.: 812). They attribute small area variations largely to physician uncertainty about appropriate diagnostic and treatment patterns. As evidence for this interpretation, they cite examples where physicians have changed their professional behavior when provided with information on geographic variations.

The study of small area variations has made an important contribution by pointing out that provider uncertainty is an important determinant of medical practice patterns. This does not necessarily imply a breakdown in the physician/patient agency relation, however. Decisions made with imperfect information and uncertainty may characterize both patient and physician behavior in most medical markets, even though the physician acts according to his perception of the patient's best interests.

Small area analysis can also be faulted on methodological grounds. First, although it shows that variations exist, it makes very little effort to control for relevant explanatory variables. For example, differences in insurance coverage are not controlled. Second, the statistical interpretation of differences in small area procedure rates is problematic. Diehr (1984) noted that surprisingly large differences in utilization rates can arise by chance alone. For example, if utilization rates are thought of as observations from a normal distribution, the highest and lowest rates will differ (on average) by 2.3 standard deviations if five small areas are being compared and by 3.7 standard deviations if twenty areas are being compared, even if the underlying rate is the same in all areas.

A reconsideration of the demand inducement model

In our earlier paper we stated, "To assess adequately the notion of supplier-induced demand, it is essential to isolate qualitative aspects of physicians' services" (Sloan and Feldman 1978: 60). What appear to be variations in demand induced by physicians' discretionary behavior may be due to systematic increases in quality or amenities. For example, patients may value longer and more thorough visits or a more attractive doctor's office. Physician decisions to provide quality can be examined in a neoclassical framework. Some of the "anomalies" of the market which the Bs say are unique characteristics of this sector may be adequately explained by a neoclassical model in which quality is a decision variable. In our earlier paper, in which we employed such analysis, we suggested that quality, like quantity, may be undersupplied by a monopolist relative to the level that maximizes social welfare.

This insight was important because it suggested that demand shifting involves more than a simple recommendation to use *more* visits. The way that shifting occurs in practice is likely to involve a change in the quality or intensity of treatment. This is particularly true if one observes what a physician does for a patient over the course of treating an episode of illness (the number of visits per episode of illness is one indicator of treatment intensity). Thus, it seems reasonable to write the inverse demand function as $P = P(Q,q)$, where Q is quantity and q is quality of service or amenities.

A shortcoming of our earlier analysis, however, was the assumption that the physician chooses the optimal level of quality at a fixed level of quantity. We accepted the conventional prediction that less quantity is produced under monopoly than under competition. This result may not apply when both quantity and quality are physician decision variables, however. In the following extension of our earlier model, we show that multiple outcomes are possible. In fact, the only case that can be ruled out is overproduction of both quality and quantity by a monopolist. We then use the model to analyze the effect of one popular prescription for controlling the cost of physicians' services—price controls, or a "physician DRG" system—on quality and quantity of service.

As a reference point, we begin with a monopolist that produces a good of fixed quality. This monopolist produces too little of the good, but a unit subsidy can be used to induce it to expand its output to the socially optimal level. The desired subsidy equals the difference between price and marginal revenue at the output where the demand curve cuts the monopolist's marginal cost curve. Alternatively, a regulatory agency could control the monopolist's price at the socially optimal level. This converts the marginal revenue curve into a horizontal line at the optimal price and induces the monopolist to produce the optimal quantity.

It is unlikely that either remedy will work when product quality is also a decision variable. In a nutshell, the regulatory agency has one instrument—price—

to hit two targets: optimal quantity and quality. In order to understand the details of the regulator's dilemma, we analyze a formal economic model in which the physician maximizes profit:

$$\pi = P(Q,q)Q - C(Q,q), \tag{9}$$

where P(Q,q) is the inverse demand function and C(Q,q) is the cost function. First-order conditions for profit maximization are

$$\pi_Q = P + P_Q Q - C_Q = 0 \tag{10a}$$

$$\pi_q = P_q Q - C_q = 0. \tag{10b}$$

In contrast, the social welfare function is

$$W = \int_0^Q P(Q,q)dQ - C(Q,q). \tag{11}$$

Income effects are assumed to be small and are ignored. First-order conditions for welfare maximization are

$$W_Q = P - C_Q = 0 \tag{12a}$$

$$W_q = \int_0^Q P_q dQ - C_q = 0. \tag{12b}$$

In order to compare the profit and welfare-maximizing equilibria, it is useful to draw a graph of the first-order conditions as lines in (Q,q) space. The relevant properties of these lines are their slopes and positions. First, consider the slope between quality and quantity, dq/dQ, when the monopolist is maximizing profit with respect to quality. This can be found by differentiating the equilibrium condition, $\pi_q = 0$, to obtain $dq/dQ = -\pi_{Qq}/\pi_{qq} < 0$ if $\pi_{Qq} < 0$. Similarly, when the monopolist is maximizing profit with respect to quantity, $dq/dQ = -\pi_{QQ}/\pi_{Qq} < O$ if $\pi_{Qq} < O$. The significance of the negative sign of π_{Qq} will be explained below.

When the social welfare function is maximized with respect to quality and quantity, respectively, the derivatives are $dq/dQ = -W_{Qq}/W_{qq}$ and $dq/dQ = -W_{QQ}/W_{Qq}$. These are both negative, provided that $W_{Qq} < O$.

Terms π_{Qq} and W_{Qq} will be negative if $P_{Qq} < O$. In our earlier analysis of the physicians' services market we assumed that consumers who are willing to pay a high price also have high marginal valuations of quality, which implies negative P_{Qq}; furthermore, negative P_{Qq} is consistent with plausible demand functions, such as the Cobb-Douglas function.

The next problem is to determine which of the lines in each pair is steeper—that is, which cuts from above. This question can be answered by taking the difference in the derivatives:

$$dq/dQ_{(\pi_q=0)} - dq/dQ_{(\pi_Q=0)} = \frac{\pi_{QQ}\pi_{qq} - \pi_{Qq}^2}{\pi_{qq}\pi_{Qq}} \quad (13a)$$

$$dq/dQ_{(W_q=0)} - dq/dQ_{(W_Q=0)} = \frac{W_{QQ}q_{qq} - W_{Qq}^2}{W_{qq}W_{Qq}} \quad (13b)$$

The numerators of (13a) and (13b) are positive from the respective second-order conditions, and the denominators are positive if π_{Qq} and W_{Qq} are negative. Provided that the latter conditions are met, (13a) and (13b) are positive; thus, the lines representing maximization with respect to quantity cut from above.

Finally, we can take the differences between the respective first-order conditions. At a given level of quality,

$$\pi_Q - W_Q = P + P_QQ - C_Q - P + C_Q = P_QQ < 0. \quad (14)$$

The implication is that the monopolist tends to produce too little quantity relative to the social optimum. At a given quantity,

$$\pi_q - W_q = P_qQ - C_q - \int_0^Q P_q dQ + C_q < 0, \quad (15)$$

which must hold since the C_q terms cancel and the remaining terms are negative. In other words, the monopolist produces too little quality (at a given quantity) if the average valuation of quality exceeds the marginal valuation, and the sufficient condition is that consumers attracted as the monopolist moves down the demand curve have lower marginal valuations of quality.

These conclusions might seem obvious. If the standard monopolist underproduces quantity, then the variable-quality monopolist should underproduce quantity and quality. However, the inequalities of (14) and (15) are valid only when the other choice variable is held constant. Equilibrium effects depend on the relative strength of the two restrictions. Figure 3 incorporates all the results derived thus far. The lines representing first-order conditions slope downward, the π_Q and W_Q lines are steeper, and there are partial quality and quantity restrictions. Monopoly equilibrium at M involves more quality and less quantity than the social optimum (which is used as the origin of the graph).

This equilibrium corresponds to a market with quantity restrictions and overproduction of quality. Obviously, two other outcomes are possible: the monopolist can underproduce both goods, or the monopolist can produce too much

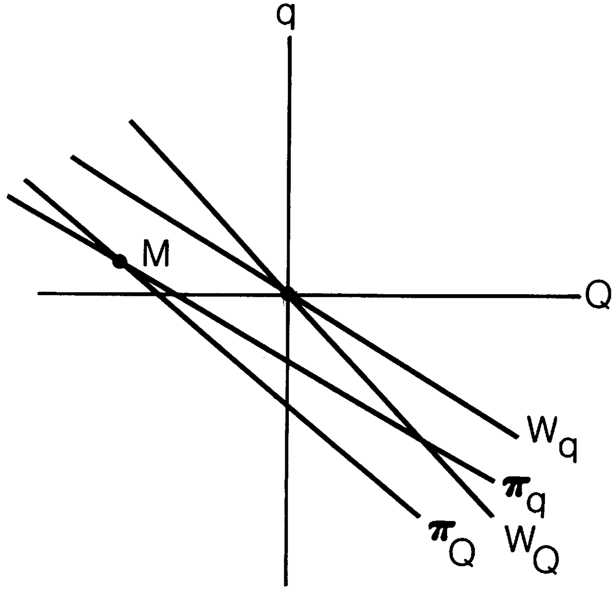

Figure 3. Monopoly Equilibrium With Too Much Quality and Too Little Quantity

quantity and too little quality. The last result is indeed surprising, given the naive supposition that monopoly always produces too little quantity.

Given this basic setup, we can analyze the economics of a physician price control system. First, we totally differentiate the profit function (equation [9]) for different parametric levels of profit, $\pi = \bar{\pi}$. For any fixed profit level, this defines a family of closed iso-profit contours around point M. The slopes of these contours are given by $dq/dQ = -\pi_Q/\pi_q$. Next, we draw the demand curve in (Q,q) space: for each price, consumer demand for quality versus quantity is represented by an upward-sloping line. Lines representing lower prices lie below and to the right of those for higher prices. The monopolist's problem is to maximize profit, given a constraint determined by the regulator's chosen price, say, $\bar{P}$. Graphically, equilibrium occurs at the point of tangency between an iso-profit contour and the $\bar{P}$ line.[3]

Constrained equilibrium is shown by point $\bar{M}$ in Figure 4. $\bar{M}$ must lie between π_Q and π_q because the price constraint slopes upward, and it is only between π_Q and π_q that the iso-profit contours also slope upward (the contours are horizontal

3. This result can be derived by maximizing profit, subject to the constraint that $P = \bar{P}$. First-order conditions for this problem can be solved to yield $\pi_Q/\pi_q = P_Q/P_q$. The left-hand side of this equality is the slope of an iso-profit contour, and the right-hand side is the slope of the demand function at $P = \bar{P}$.

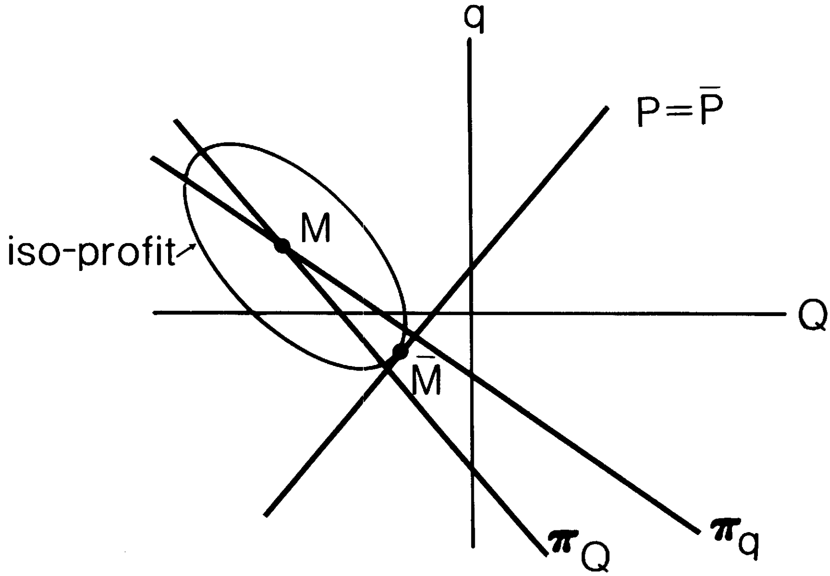

Figure 4. Equilibrium With Price Controls

where they cross π_Q and vertical where they cross π_q). We can generalize this finding to any controlled price: the monopolist's constrained equilibrium must lie within a "wedge" with vertex at M and arms formed by the π_Q and π_q lines. Since these lines are bounded away from the welfare-maximizing point at the origin of Figure 4, it follows that price controls cannot induce the monopolist to produce optimal levels of quality and quantity.

This conclusion is intuitive. The monopolist presently produces too much quality and too little quantity. The regulator forces the monopolist to cut its price in order to produce more quantity—but this causes quality to fall. For example, doctors may cut their visit lengths and boost the number of visits (or other billable services, such as laboratory tests) produced.[4] If price is cut to the level where quality is just right (along the Q-axis), quantity will be too low; if price is cut further to achieve the right level of quantity (along the q-axis), quality will be too low. Both targets cannot be hit at the same time.

As a technical exercise, we can also analyze the second-best properties of a price control system. This is done by totally differentiating the social welfare function (equation [11]) for different parametric levels of welfare to obtain a fam-

4. Medical care quality depends on much more than price levels—professional norms, community standards, defensive medicine, and peer reviews are also significant. Our prediction that quality will fall when prices are controlled below their equilibrium level represents a partial derivative, with the other factors held constant.

ily of iso-welfare contours centered about the socially optimal levels of quantity and quality. The slopes of these contours at $W = \bar{W}$ are given by $dq/dQ = -W_Q/W_q$. One of the contours will pass through point M (the monopolist's unconstrained equilibrium). It is not possible to prove that the slope of this iso-welfare contour is steeper than the slope of the π_Q line at point M, which is the sufficient condition for price controls to improve social welfare relative to M.

Again, this conclusion is intuitively appealing. Price controls may cause quality of service to decline rapidly. The offsetting increase in quantity induced by the controls may not justify the loss in quality. Thus, price controls may not contribute to a second-best welfare solution to the monopoly problem. Many critics may argue that *some* reduction in medical care quality (or the amenities and frills that pass as superficial indicators of quality) is socially justified. But this argument does not necessarily imply that price controls are an appropriate means to bring about the desired reduction in quality.

Conclusion

In this paper we have revisited the subject of competition in the physician services market, concentrating on physician-induced demand. Our visit has been deliberately narrow in scope. However, even a limited review of the recent literature on demand inducement has taken us through numerous studies and several econometric issues. This literature suggests that demand inducement may occur in the market for surgical services but that its extent is less than previously estimated. (Another interpretation is that the extent of demand inducement for surgery has declined over time.) Little evidence for demand inducement is found in the primary care physician services market.

We disagree with those who say that physicians generate demand to avoid price controls and that national health care spending is proportional to the number of physicians. The evidence does not support these arguments. But we do agree that substantial uncertainty may surround the physician's choice of diagnosis and treatment mode. However, this does not imply a breakdown of the agency relationship. We also suggest that small area studies can be improved by adding control variables that are required of any economic analysis of consumer demand.

We extended our earlier model of the physician's choice of quality to include quantity as a choice variable. We showed that a monopolist can do anything except overproduce both quality and quantity. Although some critics may argue that physicians' services are produced with too many amenities, it is not clear that price controls will bring this market to a second-best equilibrium. Price controls definitely cannot lead to socially optimal levels of both quality and quantity; only competition can.

If supplier-induced demand is quantitatively important, it has important implications for public policy. When doctors are paid by fee-for-service methods, they overproduce; under capitation, they underproduce. Thus, increasing the number of capitated plans would amount to trading one sin for another. At first glance, the only remedy would appear to be a highly regimented control system over price, quantity, *and* quality. Such a system would be administratively cumbersome, to say the least.

Suppose that the Bs are correct, and that asymmetric information between doctors and patients confers considerable market power on the former which is often exercised. Physicians and hospitals have often contended that consumer welfare is not served by providing consumers with information on the alternatives. In particular, they argue that consumers are easily confused by false and misleading claims and by comparisons of outcomes and cost that do not control for case selection. Ironically, at the same time providers frequently oppose rigid controls, which are a potential solution for combating the adverse consequences of supplier-induced demand. A more practical public policy would be one which deals directly with the root of the problem—namely, a set of strategies to improve truthful information available to consumers. Such strategies would include having the Federal Trade Commission and state government agencies vigorously enforce statutes that prohibit dissemination of misleading and false claims. If an adequate amount of information is not forthcoming from the private sector, such strategies should also include public support for information production and dissemination. For example, the Health Care Financing Administration recently disseminated information on hospital mortality rates. Wennberg et al. (1982) have also suggested providing consumers with better information on the value of health services.

At this time it is not known what types of information are most useful and to whom which types are useful. Since most people in the United States (including about two-thirds of those under age 65) are covered by private health insurance related to the employment of a family member, the most useful information might pertain to the quality of health plans rather than to the quality of individual providers. This information could be used by employers to select health plans to offer to their employees and by employees when they choose a health plan. Information on waiting times to obtain an appointment, consumer complaints, and the like might be disseminated to employees in much the same way as similar information on airlines is now distributed.

References

Auster, Richard D., and Ronald L. Oaxaca. 1981. Identification of Supplier Induced Demand in the Health Care Sector. *Journal of Human Resources* 16 (3): 327–42.

Brown, Malcolm. 1982. Problems and Prospects Concerning Canada's Medical Insurance System. In *Economics and Health: Proceedings of the Fourth Australian Conference of Health Economists*. Canberra: The Australian National University.

Bunker, John P., and Byron Brown. 1974. The Physician-Patient as an Informed Consumer of Surgical Services. *New England Journal of Medicine* 290: 1051–55.

Cromwell, Jerry, and Janet B. Mitchell. 1986. Physician-Induced Demand for Surgery. *Journal of Health Economics* 5 (4): 293–313.

Diehr, Paula. 1984. Small Area Statistics: Large Statistical Problems. *American Journal of Public Health* 74 (4): 313–14.

Dowd, Bryan E., Roger Feldman, Steven Cassou, and Jennifer Hunt. 1984. Summary Report: 1984 Employee Survey. Minneapolis, MN: Center for Health Services Research, University of Minnesota and Minnesota Coalition on Health Care Costs.

Feldman, Roger D., and David A. Ballard. 1981. The Role of Waiting Time in a Prepaid Health Care System: Evidence from the British National Health Service. *Eastern Economic Journal* 7 (34): 175–85.

Fuchs, Victor R. 1978. The Supply of Surgeons and the Demand for Operations. *Journal of Human Resources* 13 (Supplement): 35–56.

Fuchs, Victor R., and Marcia J. Kramer. 1972. Determinants of Expenditures for Physicians' Services in the United States 1948–68. Washington, DC: National Center for Health Services Research and Development.

Hay, Joel, and Michael J. Leahy. 1982. Physician-Induced Demand: An Empirical Analysis of the Consumer Information Gap. *Journal of Health Economics* 1 (3): 231–44.

Health and Welfare Canada, Planning, Policy and Information Branch. 1984a. *Canada Health Manpower Inventory 1984*. Ottawa, ON: HWC.

———. 1984b. *National Health Expenditures in Canada 1970–82*. Ottawa, ON: HWC.

Holahan, John, and William Scanlon. 1978. *Price Controls, Physician Fees, and Physician Incomes from Medicare and Medicaid*. Washington, DC: The Urban Institute.

McCarthy, Thomas R. 1985. The Competitive Nature of the Primary-Care Physician Services Market. *Journal of Health Economics* 4 (2): 93–117.

Newhouse, Joseph P., John E. Ware, Jr., and Cathy A. Donald. 1981. How Sophisticated Are Consumers About the Medical Care Delivery System? Santa Monica, CA: The RAND Corp.

Newhouse, Joseph P., Albert P. Williams, Bruce W. Bennett, and William B. Schwartz. 1982. Does the Geographic Distribution of Physicians Reflect Market Failure? *Bell Journal of Economics* 13: 493–505.

Office of the President. 1985. Health Status and Medical Care. In *Economic Report of the President*. Washington, DC: U.S. Government Printing Office.

Pauly, Mark V. 1978. Is Medical Care Different? In *Competition in the Health Care Sector: Past, Present, and Future*, ed. Warren Greenberg. Germantown, MD: Aspen Systems.

Pauly, Mark V., and Mark A. Satterthwaite. 1981. The Pricing of Primary Care Physicians' Services: A Test of the Role of Consumer Information. *Bell Journal of Economics* 12 (2): 488–506.

Phelps, Charles E. 1986. Induced Demand—Can We Ever Know Its Extent? *Journal of Health Economics* 5 (4): 355–65.

Reinhardt, Uwe E. 1985. The Theory of Physician-Induced Demand: Reflections After a Decade. *Journal of Health Economics* 4 (2): 187–93.

Rossiter, Louis F., and Gail R. Wilensky. 1983. A Reexamination of the Use of Physician Services: The Role of Physician-Induced Demand. *Inquiry* 20 (2): 162–72.

Sloan, Frank A., and Roger Feldman. 1978. Competition Among Physicians. In *Competition in the Health Care Sector: Past, Present, and Future*, ed. Warren Greenberg. Germantown, MD: Aspen Systems.

Sloan, Frank A., and William B. Schwartz. 1983. More Doctors: What Will They Cost? *Journal of the American Medical Association* 249: 766–69.

Sweeney, George H. 1982. The Market for Physicians' Services: Theoretical Implications and an Empirical Test of the Target Income Hypothesis. *Southern Economic Journal* 48 (3): 594–613.

Wennberg, John E., Benjamin A. Barnes, and Michael Zubkoff. 1982. Professional Uncertainty and the Problem of Supplier-Induced Demand. *Social Science and Medicine* 16: 811–24.

Wennberg, John E., and A. Gittelsohn. 1973. Small Area Variations in Health Care Delivery. *Science* 183: 1102–1108.