10 Supplier-Induced Demand: Some Empirical Evidence and Implications

Robert G. Evans
UNIVERSITY OF BRITISH COLUMBIA

The professional relationship arises from the significant information differential between physician and patient, and permits the physician to exert direct, non-price influence on the demand for his own services. If the economic status of the physician affects the level and direction of such influence exerted, then models of the demand for care which do not include explicit consideration of supplier behavior are incompletely specified.

This paper outlines the effect on demand analyses of two alternative specifications of physician behavior, and notes that each can lead to 'perverse' response of price to increases in supply, or of quantity demanded to price. It then examines several pieces of empirical evidence from Canada and the United States which are consistent with substantial demand influence by physicians, with responses of generated output to physician stock around 80 per cent through increases in supply of physician-initiated services. The conclusion is that policy to limit price inflation, correct 'shortages' or restrain unnecessary utilization cannot be based on conventional supply and demand models.

Everyone knows that physicians exert a strong influence over the quantity and pattern of medical care demanded in a developed economy. The professional status of the physician and the peculiar 'doctor–patient relationship' are rooted in the dual roles which the physician must perform. He acts as the agent of the patient, providing expert direction or assistance in the interpretation of the patient's health status, the identification of the capacity of current medical technology to improve that status, and the skilled application of that technology. But at the same time he is a supplier of a particular class of services whose income and work satisfaction are related to the

¹ This is a revision of the paper given in Tokyo. Care has been exercised to retain the trenchancy of the commentators' observations.
volume of services he supplies and the price he receives for them. In this role the interests of the physician tend to conflict with those of the patient, particularly if medical practice is organized on an entrepreneurial fee-for-service basis. Such a setting creates strong economic incentives for the physician to overemphasize the supply of his own services to the exclusion of substitutes and to bias the patient’s ‘choice’ of services towards those which yield the highest net revenue per time unit for the physician. One purpose of the professionalization of the physician role is the formation of a set of attitudes which will counteract these incentives.

If physicians were simply entrepreneurs supplying a particular good, ‘medical care’, then one could analyze the behaviour of the medical market using the conventional theoretical tools of supply and demand analysis. The demand side of the market could be treated as a function of price and consumer characteristics such as age, sex, income, insurance status, and so on with efforts made to define and measure demand curves for aggregate or sub-markets. Shifts in demand due to such ‘policy instruments’ as types of insurance coverage could be studied without reference to supply-side behaviour. At the same time the supply side could be handled in a conventional manner, assuming perfect competition or monopoly and investigating the implications of each with or without recognition of the fact that the entrepreneur running the firm is also a principal factor supplier with his own income-leisure trade-off.¹

What is neglected, however, is the theoretical significance of the other aspect of the physician’s role, the patient’s agent. The physician can exert direct influence on the demand function of the consumer by altering the patient’s perceptions of his needs and of the capacity of medical technology to satisfy them. Thus, the medical service market cannot be simply dichotomized into demand side and supply side, with price serving as the only nexus between the two; rather we must allow for shifts in the demand curve itself in response to supplier

behaviour. Market clearing may take place directly through the
information which suppliers pass to consumers as well as by adjust­
ments in price.

In such a market, it may no longer be taken for granted that con­
tentional economic propositions hold. If, for example, physicians
respond to increases in price by exerting a stronger positive
influence on patient's perceptions of their 'need' for service, then it is
not impossible for the total response of quantity demanded to price
to be positive. The partial response may be negative as consumer
theory predicts, but the ceteris do not remain paribus. Instead they
vary in a systematic way with price.

Similarly an increase in the physician stock, if physicians have
discretionary power over demand, may lead to increases in both out­
put and prices. This is because the power to sell more at the current
price (by increasing each consumer's demand) is equivalent to the
power to sell the current output at a higher price (by increasing
demand and simultaneously raising price). An increase in number of
suppliers at a given volume of demand and price will lead to a fall in
each supplier's income; he can respond to this change by expanding
his effort on demand generation and then either working more or
raising prices, or both.¹

Of course, demand cannot be expanded indefinitely at any given
price level. There undoubtedly comes a point where continued in­
creases in supply will lead to the re-emergence of price-competitive
behaviour. The usefulness of discretionary models depends on an
assumption that such a point is significantly far from where we now
are, and that the social costs of reaching it would be relatively large in
terms of the negative health impact of 'over-doctoring' and quality
dilution as well as the economic costs of public investment and
subsidy.

A further implication of the same argument has to do with deter­
rent charges as a device for reducing utilization and moderating cost
increases under medical insurance plans. Insofar as such charges are
successful in lowering workloads at current prices, they will lower
physician incomes. If physicians have discretionary power over
demand, they will respond by shifting the demand curve rightward

¹ Such a reaction by physicians should not be interpreted as the deliberate
provision of unnecessary care. If physicians as a group believe that the public
'needs' more care than it now receives (and there is evidence that they do, e.g.
interviews by H. Schonfeld et al., 'Number of Physicians Required for Primary
Medical Care', New England Journal of Medicine, no. 286, 16 Mar 1972), then
they may well react to lowered workloads resulting from increased density of
physicians by attempting to take better care of their patients through more
frequent recalls and follow-ups, more extensive testing, consultations, and more
services generally.
and either raising price of increasing output back to previous levels.\(^1\)
The distribution of care may change, but overall expenditures are unlikely to fall.

Thus, the possibility that suppliers can individually or collectively exercise influence over the demand curve can have rather radical implications for the role of prices as market signalling devices on the effects of exogenous shifts in supply and demand on expenditure and price.

By itself, however, the notion that suppliers influence demand does not vitiate any of the conventional propositions of market analysis. If suppliers took no account of economic data in the information which they provided to consumers, e.g. if physicians disregarded prices, their own incomes and their workloads in making recommendations to patients, then the demand function as a relation between quantity and price could be treated as stable and negatively sloped because the other parameters of the utility function would still be exogenously determined.\(^2\) Unfortunately, the literature on physician behaviour abounds with references to the responsiveness of physician practice behaviour to economic factors.\(^3\) At the other

\(^1\) The emphasis on physician responses to overall workload changes indicates that the results of partial and general deterrent charges may be quite different. A small insured group which does not account for a large part of the income of its local medical community may indeed respond to deterrents by reducing utilization. On the other hand, a deterrent fee introduced in a national or provincial insurance plan is much less likely to affect overall utilization and more likely to exert upward pressure on medical prices. R. G. Beck, ‘An Analysis of the Demand for Physicians’ Services in Saskatchewan’, unpublished doctoral dissertation (Univ. of Alberta, Edmonton, Spring 1971) emphasizes the reactions of physicians as well as patients to the province-wide deterrent fee introduced in 1968, and points out that this fee happened to coincide with a significant upward revision in the provincial fee schedule, as well as apparently increasing utilization by certain classes of patients.

\(^2\) If the full information assumption for patients is relaxed, one can still insulate the non-price variables in the demand function from indirect supplier price and income effects by hypothesizing a purely professional physician who considers only technical factors in informing his patients, then wears his entrepreneurial hat when responding to the resulting demand. It is hard to see how an economist could regard such behaviour as either rational or optimizing.

\(^3\) For example, G. Monsma, ‘Marginal Revenue and the Demand for Physicians’ Services’, in H. Klarman (ed.), *Empirical Studies in Health Economics* (Baltimore: John Hopkins Press, 1970) surveys data on the response of physician practice patterns to rates of payment. U. Reinhardt, ‘An Analysis of Physicians’ Practices’, unpublished doctoral dissertation (Yale Univ., 1970) calculates rates of payment per minute for several types of common activities and notes how these bias activities. The *Task Force Reports on Costs of Health Services in Canada* (Ottawa: The Queen’s Printer, 1970) note instances of specific responses to fee schedule revisions – sharp increases in activities whose relative price had risen – as well as general ‘over-doctoring’ in areas with relatively large physician stocks as procedures were multiplied to maintain incomes. The list could go on indefinitely.
extreme, if physicians are pure income-maximizers it is clear that each individual physician will always exert as much influence as he can to increase demand for his own services. In this case, analysis of the medical services market becomes merely a version of the monopolistic competition model with advertising and a rather peculiar advertising cost function. The assumption of income maximization assures that unexerted discretionary influence over the demand curve does not exist, hence the demand curve can be treated as exogenous.

In a real world lying between these two extremes, however, discretionary power by physicians may persist. If it is a significant feature of the medical care supply process, then it must be incorporated into models of the medical market; and there appear to be two basic ways of doing this. One can retain the maximizing framework and postulate that physicians have some broader objective function including the exercise of discretion as well as income and/or workload. Alternatively, one can abandon the maximizing framework for some general ‘target’ model, in which physicians are assumed to have rough targets for both income and leisure (the origin of such targets is unclear, but no more so than the origin of the utility function!) and to adjust prices and/or discretionary behaviour so as to approach these moving targets.

The extended maximizing model can be applied to any market in which some price responsiveness by consumers exists due to full or partial self-payment; it cannot explain market behaviour in a full-insurance system such as Canada’s Medicare Plan. Such a model might specify each physician’s utility as a positive function of his income \( Y \) and a negative function of his workload \( W \) and the extent to which he exerts discretionary influence \( D \) to increase demand. \( D \) may also be interpreted as a positive preference for non-price rationing. If we regard each physician as a monopolistic competitor whose market share depends on the population/physician ratio \( R \), and each of whose patients demands an amount of care negatively related to the price charged and positively related to demand-expanding behaviour, then we may write:

\[
U = U(Y, W, D) \\
W = R \cdot f(P, D)
\]

where \( W \) is the physician’s workload, \( f(P, D) \) is the demand for care by his ‘representative’ patient, and \( R \) allows for shifts in workload

\(^1\) An advertising model with costless advertising is essentially rather uninteresting, while the usual explicit cost function for demand generation through advertising is implausible in this context.
arising from shifts in the exogenous population/physician ratio. The problem is to maximize $U$ subject to the workload constraint $W$, by setting price level, workload and discretionary behaviour. Income is equal to the product of price level and workload, assuming a constant proportion of overhead expense.

The comparative statics of such a model are relatively straightforward, if somewhat tedious, and it turns out that the conditions for a maximum are 'reasonable' a priori. But predicted responses to exogenous shifts are not determinate; in particular, an exogenous increase in supply (a fall in $R$) may well result in a rise in price. If the $D$ variable is suppressed, a fall in $R$ for $P$ given lowers both income and workload. The marginal utility of income rises and the marginal disutility of work falls, so the physician cuts price so as to increase workload and income. (Of course, the elasticity of demand for care faced by each physician must be greater than unity to make this model work.) But in the more general model, the physician may increase both $P$ and $D$ to get back to equilibrium. In general, the difficulty is that the $D$ variable introduces an extra off-diagonal term from the Hessian matrix of the constrained maximization problem into each shift-response equation, and nothing is determinate.¹

A maximizing model extended to include physician discretion is thus able to break the conventional association between rising prices and scarcity of excess demand, but at the cost of giving up all definite predictions about pricing behaviour. A further source of weakness in such models is the structure of the patient demand equation; conditions for a maximum require that the price elasticity of demand faced by each physician must be greater than unity. As noted below, this requires either that most measurements of the elasticity are erroneous, or that physicians are price-competitive with one another. A fortiori, such a model will not explain pricing behaviour in a fee-for-service market under comprehensive, universal medical insurance such as the Canadian Medicare system. The fact that such markets do indeed function, with price behaviour roughly similar to the United States self-paying market, suggests that the maximizing model may be inadequate.

A less formal, non-maximizing model may be constructed if we hypothesize that physicians have rough targets with respect to income and workload, based on their training, expectations and previous experience. Discrepancies between these targets and actual experience lead to adjustment behaviour; if income and workloads are below targets, demand generation may take place. If physicians

¹ This is a general problem with extended maximising models, analyzed by G. C. Archibald, 'The Qualitative Content of Maximizing Models', *Journal of Political Economy*, LXXIII, 1 (Feb 1965).
feel overworked and underpaid, upward pressure on prices may develop either collectively (revision of fee schedules) or individually (independent adjustments in billing behaviour).

In Canada, revision of fee schedules requires that a large enough number of physicians in any province find target incomes unattainable and seek to put through a uniform schedule revision; the physician whose targets run ahead of his colleagues' can only work harder. In the United States, where individual physicians have more power over their own prices, the factors limiting short-run adjustment are less clear. Nevertheless, there is considerable evidence that price adjustment is only one of several possible strategies a physician may employ in meeting his practice objectives.

The implications of such a model have been developed in more detail elsewhere, but in general they include upward pressure on prices when supply increases, as well as relatively little response of income per physician. Increases in regionwide fee schedules are likely to lower output per physician (although not unambiguously), and thus rising prices are ineffective as a rationing device for clearing markets. The primary function of price is (in conjunction with workloads) to achieve physician income targets. A model of this sort is distressingly fuzzy in its failure to explain the formation of target incomes or the distribution of target shortfalls between price increase and demand generation. It does, however, have the capacity to interpret a number of empirical observations which are anomalous in a conventional supply/demand framework.

If physician influence over demand is a significant feature of the medical service market, we would expect to find relatively little relationship between physician workload and physician density per capita. On the other hand, if demand is exogenous to the physician, workload should vary inversely with density. Within each of the provinces of Canada, physicians are now paid according to a standardized fee schedule; thus, one can use annual incomes of physicians within the same province as indicators of their relative workloads. In British Columbia, annual gross receipts of each physician from the provincial medical insurance plan are published by physician name, and it is thus possible to link up gross receipts with data on physician specialty, length of practice, location, form of practice (solo or group by size) and place of training. Furthermore, since each physician is located by school district and hospital district (the latter being aggregations of continuous school districts forming a single catchment area for purposes of regional hospital planning), it is possible to

---

measure demographic and other aggregate variables by region and to investigate their significance as predictors of physician gross income and workload.

This data base has been investigated for both 1969 and 1971 in order to determine *inter alia* the influence on physician workload of relative physician densities by region.¹ The procedure employed in both surveys has been to reduce the total physician census to a subset assumed to be in full-time medical practice, and then to relate the average incomes of these physicians to a set of independent dummy variables representing professional characteristics. This data base consists of 2,279 physicians in 1969 and 2,457 in 1971. Dummy variables are used to control for each of the characteristics listed above, thus enabling one to estimate for any particular set of characteristics the total amount by which the income of the physician which they represent would have exceeded that of the base reference physician (a general practitioner who graduated from the University of British Columbia within the last five years and is in solo practice in Vancouver). Details on data file are available from the author.

With physician workload (gross receipts) standardized for the characteristics of the physician, it is then possible to introduce measures of the relative availability of medical care in each region. These are unfortunately conceptually fuzzy, for several reasons. Availability is a relation between the exogenous 'demand' for medical care (in a fully insured population) and the exogenous 'stock' of care in each region. The crudest measures of each are total populations by region and total physician stock. But clearly, people are not homogeneous in their health characteristics; the exogenous demand of a population depends on age, sex, fertility, climate, occupation, socio-economic status, etc. Out of all these factors we carried out a rough standardization for age alone, using relative utilization weights for four different age groups to derive a standardized population for each region. Nor are physicians homogeneous; we therefore calculated physician/(standardized) population ratios for all physicians, general practitioners, and general practitioners and 'practicing specialists',² and

¹ R. Evans, E. Parish and F. Sully, 'Medical Productivity, Scale Effects and Demand "Generation"' (mimeo), Univ. of British Columbia, Department of Economics Discussion Paper No. 79 (June 1972) reports investigations with the 1969 data.

² 'Practicing specialists' is a concept borrowed from D. Anderson and A. Clough, 'The Location of British Columbia Physicians', *British Columbia Medical Journal*, II, 5 (Sep 1960). It excludes radiologists, pathologists, anesthesiologists, public health physicians and rehabilitation specialists, in an attempt to focus only on specialists who are or might be in direct primary contact with patients. The substitution between general practitioner and practicing specialist is obvious in the case of a paediatrician: it becomes more strained in the case of a neuro-surgeon.
fitted separate equations for each group. The problem of course is to define a subset of physicians who are substitutes for one another (more or less), and there exists no precisely correct procedure. Regional boundaries are a further problem, since one may choose small regions such as school districts, in which case referral across boundaries leads to mis-specification of available supply, or large regions such as hospital districts which may span several medical ‘markets’. We chose the latter to minimize the referral problem: tertiary referrals to the university-affiliated hospitals in Vancouver are missed, but secondary referrals to urban centres are in general within regions.

The coefficients on these physician/population density variables are as follows (t-statistics in brackets):

<table>
<thead>
<tr>
<th>Year</th>
<th>$AMD$</th>
<th>$GP$</th>
<th>$GPPS$</th>
<th>$PS$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1969</td>
<td>-0.158</td>
<td>-0.125</td>
<td>-0.179</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td>(2.68)</td>
<td>(1.46)</td>
<td>(2.75)</td>
<td>(0.27)</td>
</tr>
<tr>
<td>1971</td>
<td>-0.087</td>
<td>-0.274</td>
<td>-0.070</td>
<td>0.061</td>
</tr>
<tr>
<td></td>
<td>(1.27)</td>
<td>(2.28)</td>
<td>(0.88)</td>
<td>(1.06)</td>
</tr>
</tbody>
</table>

$AMD$ is the equation fitted for all physicians of whatever speciality, $GP$ are general practitioners only, $PS$ are selected specialties (see footnote 2, p. 169) and $GPPS$ is fitted on sums from $GP$ and $PS$. Equations are double-log, so coefficients represent elasticity. The $PS$ equation suggests that interaction among specialists has a positive effect on workload; all other equations indicate weak negative effects from density to workload. This is consistent with the hypothesis that physicians can to a large extent generate enough activity to insulate their workload against supply shifts. Adjustment is not complete, however, so a target income model would also suggest that supply increases generate upward price pressure. Moreover, the incomplete adjustment is consistent with increasing marginal disutility of demand generation.

Shifting ‘demand’ in response to changes in the effective supply of physicians is not of course the only interpretation of these results. One could argue that British Columbia has a physician shortage such that all physicians are working at full capacity regardless of the number available and that some patients are turned away. The plausibility of this argument is weakened by the fact that British Columbia (B.C.) has far and away the highest ratio of physicians to population in Canada, about 25–30 per cent above the national average (depending on definition). Many regions of B.C. have still higher ratios. Alternatively, one could argue that demand is a function of the shadow price of time and other access costs which are reduced when physician density increases; moreover, demand may vary from region
Evans – Supplier-Induced Demand

to region due to shifts in non-measured variables and physicians have simply distributed themselves in response to income differentials in such a way as to even out incomes. Neither of these arguments can be tested directly with aggregate data, since they appeal to unmeasured shadow prices and demand shifts, but they can be indirectly tested by evidence on the particular patterns of medical services supplied as physician density changes.

Data on the mix of services by physician and region in B.C. were not available for this study. However, a set of data on the operations of the Canadian non-profit private insurance plans which predated the universal federally initiated Medicare program is available from 1957 to 1967. These data indicate that across provinces the relative frequency of physician-initiated services such as consultations tends to be more closely correlated with physician availability than does the frequency of patient-initiated services such as first office or home visits. Moreover, the rapid increases in physician availability in Canada during the period 1957 to the present, due both to increases in manpower and to increases in services per physician, have likewise been associated with increases in the mix of total services accounted for by physician-initiated services. Consultations and diagnostic tests per capita have increased much faster over time than home or office calls; this is not consistent with either the argument that patients are responding to shadow prices or that medical services are merely reducing a backlog of unmet need. In the same context, it is notable that the shift from house to office calls in Canada over the past fifteen years has not been associated with any change in their relative prices. Thus, what evidence we have on inter-regional and intertemporal patterns of service mix is consistent with supplier generation of demand rather than consumer response to implicit prices or exogenous demand-side effects.

1 Evans, *Price Formation*, chap. 4. It must be confessed that data points are scanty and correlations are not perfect; data cover B.C., Alberta, Manitoba, two Ontario plans and the Maritime Provinces, with one observation for each.

2 Ibid., chap. 1. The problem is that price data are thoroughly unreliable, while data on increases in gross income per physician are drawn from taxation statistics and are quite good. The increases in gross receipts must be allocated between output per man and price, allowing for some quality change. Between 1957 and 1970 gross receipts per physician have risen 144.3 per cent, or 7.1 per cent per year. Regardless of how it is split up, this yields very high rates of increases in price or output or both. Over the same period, the Consumer Price Index rose about 2.5 per cent per year, and the stock of physicians per capita (in active fee practice) rose 1.5 per cent per year.

3 Of course, all such evidence can be rationalized in a conventional supply and demand framework. Inter-regional differences in physician availability may lead to differences in implicit costs of physician access, hence different demand patterns. Changes in service mix may be due to changes in patient tastes,
If, in fact, discretionary influence over demand is a significant feature of the market behaviour of physicians, then we should expect to find it reflected in price behaviour. And we do discover a general tendency for physician prices to be relatively higher in regions where physicians are plentiful than where they are scarce. While data are sketchy and relative levels are hard to establish owing to the discontinuous jumps which take place in listed fees, there is a clear tendency for medical care prices to be higher in those provinces of Canada with high physician/population ratios (Ontario, British Columbia) than in those with low (Saskatchewan, the Maritimes). The direction of causality in such an association is by no means obvious: higher prices per unit of output could as easily be the magnet which attracts more physicians. But the problem with this argument is that income relationships are much less clear: high prices are associated with both high (Ontario) and low (British Columbia) relative incomes. There is some evidence that higher than average increases in physician density are associated with lower than average increases in income; but the mechanism is clearly not through slower rates of price increase. Rather it appears to be incomplete ability to adjust quantity of workload per physician, perhaps operating with a lag.1

If the demand relationship is made subsidiary to supplier behaviour, then one ought to be able to use similar models to interpret United States and Canadian physician behaviour in spite of the differences in medical care financing. Patterns of organization on the supply side remain relatively similar. The existence of physician discretionary power is indicated by such United States results as Feldstein’s inability to derive any sort of ‘reasonable’ demand curve with a negative price term even in the context of a simultaneous market model under several different price adjustment specifications.2 His conclusion that physicians may hold prices below market clearing levels to enjoy the utility of non-price rationing is the inverse of the notion that demand generation at given prices is possible but involves disutility. Similarly, Newhouse3 finds systematic positive correlations between physician density and price in the United States, rejecting both ‘competitive’ and ‘monopolistic’ models of the market which use price as the only nexus between supply and demand. He notes operating through the physician as intermediary. Inter-regional variations in ‘health’ or tastes for medical care may lead to bidding up of prices in high-demand regions and drawing in of more physicians, hence a positive correlation between physician density and price. But a model rescued by a succession of ad hoc unobservables is an irrefutable and uninteresting model.

1 Evans, *Price Formation*, chap. 3.
2 Feldstein, *op. cit*.
3 Newhouse, *op. cit*. 
that rationing by physicians may explain his results. Reinhardt\(^1\) finds that the internal organization of physician practices is inconsistent with income maximization in that the shadow prices of physician time implicit in fee schedules are much higher than those implicit in physician/aide mixes. He suggests that a state of generalized excess demand creates a 'price umbrella' permitting this form of discretionary 'inefficiency by preference', but discretionary power over demand is an equally good umbrella. It permits the physician whose preference for income is strong to organize a fast throughput 'shop' without squeezing out his less aggressive competitors. Finally, the numerous studies which measure elasticities of demand for medical care at between \(-1\) and zero are rendered plausible: if physicians were income- or income/leisure-maximizers, they should always seek higher prices to raise income and lower workload, and the observed market result would depend on their having insufficient ability to collude so as to raise prices all round. But if in fact each physician generates his own demand, then no stable exogenous demand curve exists to be measured. Price being correlated with demand-generation effort will lead to a reduction in the measured response of demand to price, but the partial response of demand to price, \textit{everything} else held constant, is not being tested.

Thus, there exists a wide range of empirical evidence that the market for physicians' services is not self-equilibrating in the usual sense, that price does not serve primarily to balance supply and demand because there are important alternative channels of information which perform this function. The primary role of price is instead as an input to supplier incomes, which are not themselves the product of explicit maximizing behaviour but rather of target-seeking through the manipulation of several different control variables. In this context, as noted above, the explanation of supply, 'shortages', utilization and price and cost behaviour may be very different from the usual market models. To the extent that shortages, prices or costs are objects of public policy, policy prescriptions must take account of these differences.
