The Demand for Medical Care Services

A Retrospect and Prospect

Joseph P. Newhouse
The research reported herein was performed pursuant to Grant No. 016B-8001 from the U.S. Department of Health, and Human Services, Washington, D.C. The opinions and conclusions expressed herein are solely those of the author, and should not be construed as representing the opinions or policy of The Rand Corporation or any agency of the United States Government.

The Rand Publication Series: The Report is the principal publication documenting and transmitting Rand's major research findings and final research results. The Rand Note reports other outputs of sponsored research for general distribution. Publications of The Rand Corporation do not necessarily reflect the opinions or policies of the sponsors of Rand research.

Published by The Rand Corporation
The Demand for Medical Care Services

A Retrospect and Prospect

Joseph P. Newhouse

March 1981

Prepared for the
U.S. Department of Health and Human Services
PREFACE

This report was written as a survey paper for the World Congress on Health Economics, held at Leiden, the Netherlands, September 9-11, 1980. The paper surveyed the literature on the demand for medical care services and speculated about directions the literature may take in the next few years. Together with other papers from the conference, it will be published in the forthcoming proceedings of the conference by the North-Holland Publishing Company (editors Jacques van der Gaag and Mark Perlman).

The report should be of interest primarily to economists working in the field of medical economics. Those interested in the substantive results may wish to consult another paper by the author, Insurance Benefits, Out-of-Pocket Payments, and the Demand for Medical Care: A Review of the Literature, Rand Paper P-6134, also published in Health and Medical Care Services Review, July-August 1978. Some of the methodological issues discussed in the present piece are further elaborated in a paper by the author, Charles E. Phelps, and M. Susan Marquis, On Having Your Cake and Eating It Too: Econometric Problems in Estimating the Demand for Health Services, Rand Report R-1149-1-NC, also published in Journal of Econometrics, August 1980, pp. 365-390.

This study was supported by the U.S. Department of Health and Human Services under Grant No. 016B-8001.
SUMMARY

This report surveys the literature on the demand for medical care services, especially the literature of the 1970's, and conjectures where the literature may go next. It pays particular attention to the response of demand to changes in cost sharing arrangements; although other determinants of demand are treated as well.

The literature clearly establishes that demand responds to changes in cost sharing, although the magnitude of the response is not known with great precision. This review is organized by the sources of the data that various studies have used, because these data have largely dictated the study's methodology. Previous studies have used four sources of data: premium or claims information, natural experiments, comparison of demand of individuals with different insurance policies (usually utilizing household survey data), and designed experiments. The various advantages and drawbacks of methods using each source of data are discussed.

A challenge to the generalizability of these results can be raised if cost sharing is changed on a large scale. Some believe that physicians have an unexploited ability to affect demand sufficient to offset any autonomous or exogenous shift in demand. Although controversial in the literature, there are a number of reasons why physicians probably cannot fully offset exogenous changes in demand, so that the studies reviewed here remain of interest.

Several frontiers of demand analysis are discussed. These include specification of price, income, and health status variables; specification of the stochastic term and use of appropriate estimators; specification of the dependent variable; the "norms" argument (the argument that physicians gear their treatment advice to the average or modal patient) and technological change. How price and utilization respond to changes in demand is a particularly important area to explore because there is a reason to think that the standard competitive paradigm may not apply in a heavily insured market. The fully insured consumer has no incentive to seek an efficient supplier; hence, the standard competitive supply curve may not exist. The role of insurance companies or third parties in influencing price is also not well understood. Thus, the theory of both price and utilization at the market level is in need of attention. But the substantial progress of the past decade bodes well for the next.
## CONTENTS

PREFACE ........................................ iii  
SUMMARY ........................................ v  

Section  
I. INTRODUCTION .............................. 1  
II. THE PRICE ELASTICITY OF DEMAND .......... 3  
   Studies Based on Insurance Claims or Premium  
   Data ........................................ 3  
   Studies Based on Natural Experiments .......... 5  
   Studies Based on Comparison of Individuals with  
   Different Insurance Policies .................. 7  
   Designed Experiments ......................... 10  
III. THE EFFECTS OF OTHER VARIABLES ON  
     DEMAND ................................... 12  
IV. DOES SUPPLIER-INDUCED DEMAND INVALIDATE  
    THE DEMAND FUNCTIONS IN THE  
    LITERATURE? .................................. 14  
V. FRONTIERS OF DEMAND ANALYSIS ............ 16  
   Specification of the Price, Income, and Health  
   Status Variables ............................... 16  
   Specification of the Stochastic Term and  
   Appropriate Estimators ....................... 17  
   Specification of the Dependent Variable ........ 18  
   The Norms Argument and Technological Change ... 20  
   How Do Price and Utilization Respond to Changes  
   in Demand? ................................... 21  
VI. CONCLUSION .................................. 24  
REFERENCES ..................................... 25
I. INTRODUCTION

During the past decade, empirical knowledge about the demand for medical care, especially its price or insurance elasticity, has increased enormously.¹ Most of the relevant research has taken place in North America, especially in the United States, and has used American data. This survey concentrates on those studies.²

The United States' debate about the appropriate role, if any, of patient cost sharing in a national health insurance plan has helped stimulate American research into the magnitude of insurance elasticities. The elasticity plays an important role at two levels. The United States Department of Health and Human Services has developed a model to estimate the monetary costs of alternative insurance plans; the elasticity of demand is a necessary input. At the same time, economists have developed normative models of insurance in which the optimum depends on the elasticity of demand (Zeckhauser, 1970; Feldstein, 1973; Arrow, 1973).

Analyses of price and insurance elasticities tend to be of less significance outside the United States because other countries have historically had little or no interest in patient cost sharing. Nonetheless, quantifying nonprice determinants of demand appears to be of universal value (although the determinants may differ in their importance from one country to the next). For example, how travel and queueing times affect demand is useful to know when making location decisions about facilities or medical manpower. Similarly, how trends in demographic variables affect demand is important to know when planning the appropriate number of physicians to train.

Before reviewing empirical estimates of demand functions, some problems in quantifying demand require comment. Broadly speaking, the literature has measured demand in two ways: expenditures and physical units of utilization. Although expenditures provide a ready metric for aggregating disparate services, they suffer from two potential disadvantages. First, insofar as prices do not reflect marginal cost, variation in expenditures has ambiguous welfare implications. Second, ³

¹Phelps and Newhouse (1974a) show that price and insurance elasticities differ only by an empirically negligible income effect, provided the gross price is invariant to variation in insurance.
²For reasons of space I do not take up analysis of demand for insurance or demand for prepaid plans such as Health Maintenance Organizations. See Phelps (1973), Keeler, Morrow, and Newhouse (1977), Roemer and Shonick (1973), and Luft (in press). Nor do I treat the demand for health. See Grossman (1972).
if the gross price for a given medical service changes when insurance changes, induced variation in expenditure need not closely correspond to induced variation in real resource use.

Alternatively, one can measure demand in physical units, such as hospital days or physician visits. Although such variables appear to measure real resource use, they suffer in practice from incompleteness. In the case of inpatient services, the variables typically employed are admission rates and length of stay (or their product, hospital days), but neither measures the intensity of service—the most rapidly rising component of hospital expenditures in developed countries. The same criticism applies to the usual measure of demand for outpatient services, namely, physician visits. Visits can differ markedly in their intensity (e.g., in the number of diagnostic tests performed), and a simple count of visits clearly does not capture such variation. These imperfections should be borne in mind when assessing the literature: I return to the issue of measuring demand below.
II. THE PRICE ELASTICITY OF DEMAND

In 1971, Rashi Fein testified before the United States Congress that he knew of no evidence indicating that deductibles and coinsurance "lowered utilization" (Fein, 1971). Although one might cavil at this claim, the studies of this matter, cited below, came almost entirely at or after this time. Elsewhere (Newhouse, 1978a) I have reviewed many of these studies and have described their estimates of price elasticity, service by service. However, rather than repeat that material here, I focus instead on the methodologies various analysts have employed to derive estimates. The nature of the data used in each study has, to a large degree, dictated the methodology used. I have therefore classified studies by the type of the data used to estimate elasticities. Each type of data possesses certain strengths and weaknesses.

STUDIES BASED ON INSURANCE CLAIMS OR PREMIUM DATA

These studies use quite simple methods. Expenditure data on insurance claims, when available, are analyzed as a function of insurance plan. When individual claims are not available, premiums for different insurance policies contain implied elasticities of demand with respect to insurance. For example, suppose premiums are quoted for policies with 10-percent and 25-percent coinsurance. If demand elasticity were zero (and if loadings were proportional to actuarial value), the premium for the 10-percent coinsurance policy would be 20 percent \((100-10)/(100-25)\) higher than the premium for the 25-percent coinsurance policy. If the difference in premiums between the two policies is greater than 20 percent, one can infer the elasticity from a formula found in Phelps and Newhouse (1974a).

This method has been applied to estimate the response of demand to variation in coinsurance and deductible rates. The medical services analyzed include hospital and physician services (combined), hospital services alone, and dental services alone. Phelps and Newhouse (1974b) found demand for hospital and physician services combined to be 6

---

percent higher at 10-percent coinsurance than at 25-percent coinsurance.\textsuperscript{1} Newhouse, Rolph, Mori, et al. (1980) estimated that demand per person at a $1000 per person per year deductible was about two-thirds of demand at a $50 per person per year deductible (1975 dollars). Most of the response occurred in the range near $50, as theory would suggest (Keeler, Newhouse, and Phelps, 1977). Freiberg and Scutchfield (1976) found that a decline in out-of-pocket cost per day of hospitalization from $44 to $2.50 (which approximates a change in the coinsurance rate from 50 percent to zero) was associated with a 50-percent increase in hospital admissions. Phelps and Newhouse (1974b) found that demand for dental services was 30 percent higher at zero coinsurance than at 20-percent coinsurance.

This method has the following strengths:

1. It is usually straightforward to specify the variation in price because one is examining a well-defined change in coinsurance or deductible rates. (Premiums for policies with varying scopes of services, however, must be adjusted so that they are comparable, and this is not always straightforward.)
2. Because the data usually pertain to large employer groups, one can reasonably presume that insurance is exogenous or that any self-selection is minimal.
3. The variation in utilization occurs for a small group in the market; thus, these estimates can be fairly taken to represent demand elasticities.

On the other hand, this method has the following drawbacks:

1. Very little information is available about the characteristics of the insured, making it difficult to appraise the distributional effects of cost sharing. One could, in principle, examine how premiums vary for groups of individuals with differing characteristics, but in practice it is unlikely that the variation in group characteristics will suffice to permit many inferences.
2. One will not collect any information about services that are not covered. If, for example, outpatient mental health services are not covered, one cannot learn how demand for them responds to variation in insurance.
3. Some services may be covered in certain policies but not in others (e.g., outpatient mental health). If such services have non-zero cross-price elasticities with other covered services, one must sometimes make untestable assumptions about the

\textsuperscript{1}Where possible I have reported percentage differences rather than elasticity estimates because one can clearly keep in mind the range of coinsurance under discussion.
joint probability density function of demand for covered and noncovered services to obtain unbiased estimates of own-price elasticities.

4. To use this method to estimate the effects of a deductible (in contrast to coinsurance), one must know the probability density function of demand, something that may be difficult to obtain in practice. For example, policies that have a deductible will typically not generate information on demand by those who spend less than the deductible. At levels of deductibles that most individuals exceed, this is not a serious problem; but for larger deductibles, it forces one to make unverifiable assumptions about the shape of the probability density function of demand in the region below the deductible.

5. Individuals (or providers) may not claim reimbursement for all services actually used because of transaction costs incurred in filing a claim. As a result, one must typically make unverifiable assumptions about the rate of underfiling. The seriousness of this problem clearly depends on the magnitude of underfiling, but is of some importance when one is estimating the effects of variation in a deductible, because underfiling is concentrated in the region of total expenditure near the deductible.

6. Results cannot necessarily be generalized to those not covered by group insurance plans.

STUDIES BASED ON NATURAL EXPERIMENTS

These studies typically compare demand by a group of individuals before and after their insurance policy changed. For example, Scitovsky and Snyder (1972) and Scitovsky and McCall (1977) measured how frequently Stanford University employees made physician visits before and after the employees' coinsurance rate changed from zero to 25 percent. Visits fell approximately 25 percent one year after the change, a decline that maintained itself four years later. The method has also been applied to hospital services by Heaney and Riedel (1970) and Williams (1966), to dental services by Morehead, Donaldson, and Zanes (1971) and Grubb (1964), and to drugs by Smith and Garner (1974). In the case of dental services and drugs, full insurance caused demand to approximately double relative to situations in which services were not insured. In the case of hospital services, the response was, speaking very roughly, about half as great.

Roemer, Hopkins, Carr, et al. (1975) estimated that a $1 copayment
for office visits in the California Medicaid program decreased office visits\(^3\) but increased hospitalization to such a degree that total expenditure increased. This finding caused them to characterize the copayment as "penny-wise and pound-foolish."

For estimating price elasticities, natural experiments share many of the advantages of using premium or claims data:

1. The price variation can usually be specified in a straightforward manner.
2. Because the same group of individuals is compared on a before and after basis, self-selection problems do not arise.\(^4\)
3. Usually the change occurs for a relatively small group in the overall market, but not always. For example, Beck (1974) analyzed the natural experiment that occurred when Saskatchewan implemented a province-wide copayment. The resulting decrease in utilization could have reflected elements of both supply and demand. A similar remark applies to comparisons of utilization before and after the United States Medicare program (Bombardier, Fuchs, Lillard, et al., 1977).
4. Natural experiments may permit examination of whose demands are more affected by the change (i.e., estimation of interactions between price elasticities and demographic variables). For example, in the data collected by Scitovsky and Snyder, demands of female dependents were more responsive than those of any other group (Phelps and Newhouse, 1972).

Natural experiments have the following drawbacks:

1. The before and after method confounds the effects of the change in price with changes in any other variable affecting demand (e.g., an influenza epidemic in one period and not another). The more preferable design—a comparison with either an equivalent control group or an equivalent alternative treatment group—is not available.
2. The group among whom the natural experiment occurred may not be representative; e.g., Stanford University employees may differ from the United States population. Insofar as the group is large enough to study variation in response within

\(^3\) See also Brian and Gibbens (1974).

\(^4\) The Roemer et al. study (1975) may be an exception; in this study there was a nonequivalent control group that was exempted from the copayment in the before period. Unfortunately, those who were hospitalized in the before period may have been disproportionately in the no control (copayment) group. If so, hospitalization in the control group would have fallen spuriously (see Heims, Newhouse, and Phelps, 1978).
the group, this disadvantage may be overcome by appropriate weighting.

STUDIES BASED ON COMPARISON OF INDIVIDUALS WITH DIFFERENT INSURANCE POLICIES

The economics literature has most commonly used this method. When the unit of observation is the individual (or household), the method utilizes data from surveys that ask the household about its insurance and use of services. These data, together with the covariates available in the survey, are then used to estimate a demand curve (Colle and Grossman, 1978; Davis and Reynolds, 1976; Holtmann and Olsen, 1976; Manning and Phelps, 1979; Newhouse and Marquis, 1978; Newhouse and Phelps, 1976; Phelps, 1975; Rosenthal, 1970; Rosett and Huang, 1973). Alternative data on area averages have been used in place of data on individual households (Davis and Russell, 1972; M. Feldstein, 1971, 1977; P. Feldstein, 1973; Fuchs and Kramer, 1972; Rosenthal, 1964).

These cross-sectional studies exhibit a wide variety of price elasticity estimates. Some of the variation in the estimates may reflect true differences in the populations from which the observations were drawn; some reflects differences in specification (Newhouse, Phelps, and Marquis, 1980). Other variations may stem from a (true) nonconstant demand elasticity, with observations coming from varying ranges of price or insurance. The estimates of Newhouse and Phelps (1976) are among the lowest values; they estimate elasticities around -0.1; studies of Martin Feldstein (1971, 1977) yield somewhat higher estimates (around -0.5), as do those of Fuchs and Kramer (1972) and Davis and Russell (1972). Perhaps the highest elasticities reported in the literature are those of Rosett and Huang (1973), who find an elasticity of around -0.35 at a coinsurance rate of 0.2, but an elasticity of -1.5 at a coinsurance rate of 0.8. The estimates around -0.1 more closely resemble estimates derived by using other methods, but comparisons are difficult for reasons discussed below.

These studies have the following advantages:

1. Data can be collected from a representative sample (e.g., a national probability sample).
2. Data collected at the household level usually contain enough information about other characteristics of the household to allow one to estimate interactions of price elasticities with demographic variables.
3. One can obtain information on the consumption of services that are not insured either because they fall outside the scope of coverage or below a deductible.

4. Data from geographic aggregates allow for the possibility that physicians (or other health care providers) gear their treatment advice to some average or modal insurance coverage in the local market area, instead of, or in addition to, the coverage of the individual being treated (the so-called "norms hypothesis").

But there are the following drawbacks:

1. Price or insurance is difficult to specify. Actual insurance policies can vary along so many dimensions that simple parameterizations, such as the frequently used average coinsurance rate, can cause inconsistent estimates (Newhouse, Phelps, and Marquis, 1980). For example, if one computes the average or marginal coinsurance rate for those facing a deductible, and uses that rate to estimate a price elasticity, the resulting estimate will be biased away from zero because the error term will be negatively correlated with the average coinsurance rate.° Such problems are necessarily present when the data are averages from a geographic entity such as a state, and the direction of the inconsistency cannot always be signed a priori.

Another misspecification of the price variable occurs if no direct information on price is available, and price is estimated by dividing total expenditures by quantity consumed. In that case measurement error in the quantity consumed variable can result in a well-known bias away from zero in the estimate of the price elasticity. Fuchs and Kramer (1972) attempt to avoid this problem by using an instrumental variable estimator, but asymptotic properties of consistency may be of little solace with a small number of observations.

2. Use of aggregate data raises the possibility of aggregation bias or instability. For example, suppose one wishes to predict what will happen if a uniform national plan is enacted. In such a case the change in insurance across individuals will vary depending upon their prior coverage. Unless prior coverage is

°Use of the average coinsurance rate if a deductible is present introduces error into the measurement of the own-price variable as well as correlation with the error term; this errors-in-the-variables problem causes an inconsistent estimate of the price elasticity, and the direction of this inconsistency cannot be signed a priori. (See Newhouse, Phelps, and Marquis, 1980). In practice, the negative correlation of the price variable with the error term seems likely to dominate the direction of the overall inconsistency.
independent of the individuals' insurance elasticities, or varies in a known way, aggregating across individuals will introduce bias.

3. Aggregation across services can present another kind of aggregation problem, if services vary in the extent of coverage, because theory permits such aggregation only when relative prices are constant. This problem is particularly dramatic in the Rosett and Huang (1973) study; they examined expenditures at various average coinsurance rates, but the rate is computed as the average rate for all observed expenditures. Those entering the hospital typically had their hospital bill covered; those who did not go to the hospital typically had a much smaller expenditure that was not covered by insurance. Hence, insurance appeared to induce a large increase in total expenditure. If one knew the coinsurance rates for hospital and outpatient services, of course, one could enter such rates in the demand equation.

4. The theory of demand for insurance suggests that individuals with high demands for medical care will seek relatively greater coverage (Phelps, 1976). If so, insurance is endogenous. Unfortunately, no instruments are known that strongly identify the effects of insurance. For individual data, the best conceptual variable appears to be work group size, because the size of the loading falls systematically as work group size increases (Phelps, 1973), and work group size arguably does not affect the demand for medical services directly. Nonetheless, a two-stage least squares estimator, when applied to a sample of several thousand observations, did not yield statistically significant estimated price elasticities (Newhouse and Phelps, 1976). In light of the evidence of non-zero price elasticities that other methods show, it seems unlikely that this imprecision occurred because the true price elasticity is zero: rather, the difficulty seems to be weak identification. Whether household surveys will be sufficiently large to achieve reliable results is thus problematical.*

5. Both price and utilization variables may be subject to considerable measurement error. Households are asked to recall utilization over some past period; studies show that the proportion of total variance that is error variance in household

---

*The distribution of insurance in Newhouse and Phelps' data was also not favorable; few individuals had insurance for ambulatory cases. A population with greater variance in the insurance variable would yield more price estimates, although significant estimates were obtained by using ordinary least squares (OLS).
reported annual hospital and dental expenditure could be as high as 40 percent (Marquis, Marquis, and Newhouse, 1976; Marquis, 1980). Random error in measuring utilization, of course, only degrades precision. But random error in the price variable will bias the coefficient toward zero, and the household cannot necessarily be relied upon to furnish error-free information about insurance. The large American surveys (the Medical Care Utilization and Expenditure Survey; the Center for Health Administration Studies surveys) attempt to verify the household-supplied insurance information with employers or insurance companies, but verification efforts do not always succeed and can miss policies that the household respondent simply does not mention.

6. The estimates from the studies cited in this section come from a wide range of insurance; some are true price elasticity estimates with no insurance present; others have observations from a wide range of coinsurance rates. Thus, the interval within which insurance or price varies may be quite different across studies. This difference can make it difficult to compare these studies, one with another, and also with studies based on other methodologies.

DESIGNED EXPERIMENTS

Two experiments (at least) have been designed to gather information about the price elasticity of medical care. One took place in Kansas during an eight-month period in the 1960's (Lewis and Keairnes, 1970; Hill and Veney, 1971). The other is the Health Insurance Study, a multiyear project that is still in progress. Some initial results from the Health Insurance Study are reported in another paper prepared for The World Congress on Health Economics (Manning, Morris, Newhouse, et al., 1980).

Designed experiments randomize (or use related techniques) individuals to an experimental treatment(s) and to a comparison group. The treatments vary the price of care. The Kansas study tested the hypothesis that reducing the coinsurance rate for outpatient care would reduce total expenditure by reducing demand for hospital services; in contrast to the results of Roemer, Hopkins, Carr, et al. (1975) described above, the hypothesis was not supported.

The Health Insurance Study (Newhouse, 1974) was designed to estimate the own-price elasticity for most medical services, the interactions, if any, of that elasticity with various demographic character-
istics, and the cross-price elasticity between inpatient and outpatient care. Preliminary results suggest that demand for care is about 50-percent higher when care is free than when it is subject to an income-related catastrophic insurance plan. Results for the cross-price elasticity between inpatient and outpatient care are consistent with those from the Kansas study.

Designed experiments, when appropriately executed, have the following advantages:

1. Insurance is exogenous and readily parameterized.
2. Insurance is approximately orthogonal with other covariates.
3. The range of variation in price is high relative to most natural experiments, thereby yielding information on elasticities over a wider range of coinsurance as well as increasing the precision with which one can test the hypothesis that price affects demand.

Designed experiments have the following disadvantages:

1. Artifacts of the experiment, such as its temporary length, may interfere with reliable estimation. For example, if the study in Kansas had run for more than eight months, results may have differed (although in this particular instance the results from the Health Insurance Study suggest not). More generally, individuals may behave differently in an experiment than they would in an actual program (a Hawthorne Effect), and methods to test whether behavior is different do not always exist.

2. Designed experiments are relatively expensive; if not properly executed, they risk substantial wasted resources.
III. THE EFFECTS OF OTHER VARIABLES ON DEMAND

Empirical studies have found that a number of nonprice variables importantly affect demand. These include health status, age (which may reflect imperfect measurement of health status), sex, income, time price, and race; the effect of education has been more difficult to isolate.¹

Health status, when included, usually explains more variation than any other single variable (Newhouse and Phelps, 1976). The observed relationship between health status and utilization might, however, represent reverse causality (Manning, Newhouse, and Ware, in press). The mechanism is as follows: Household surveys typically measure health status at the time of the interview and utilization over the past twelve months. But individuals who visit the physician at higher rates are more likely to have health problems discovered and may perceive their health to be less good than those who do not visit a physician very frequently. Thus, self-rated health status may be endogenous, and the coefficients in the literature inconsistent. Still, there can be little doubt that, other things equal, those in poorer health will have a greater demand for medical services.

Estimated income elasticities have varied from near zero to approximately 1.0. Grossman (1972) has suggested that wage and nonwage income will affect demand differently because an increased wage raises both the value of time used in consuming medical services and the value of any reduction in sick days. To date, the evidence is consistent with this view, although the effect of nonwage income upon demand has not been estimated with any precision.

The greater the distance one must travel to receive medical care, the less the demand for medical care (Acton, 1975, 1976; Phelps and Newhouse, 1972; Simon and Smith, 1973). Distance is one of several variables that affect the time price of care; using Becker’s general approach, Phelps and Newhouse (1974a) and Acton (1975, 1976) have developed the theory of how time price affects the demand for medical care. If elasticity with respect to total price (time price plus money price) is constant, observed time price elasticities should rise as money prices fall, and observed money price elasticities should rise as time prices fall. The latter prediction is borne out by the greater (in absolute

¹On occasion, the estimated effect of education has been negative (Colle and Grossman, 1978); at other times it has not explained much variation (Manning, Morris, Newhouse, et al., 1980).
value) price elasticities for home visits than office visits, and the former prediction is consistent with the relatively high values for time price elasticities at zero or near-zero money prices (Phelps and Newhouse, 1974b).

Of considerable interest is the interaction, if any, between income and price elasticities; in particular, are the poor more sensitive to price than the more well to do? Evidence from Canada (Beck, 1974) suggests that this is the case; United States data are inconclusive or show the opposite (Phelps and Newhouse, 1972; Newhouse and Phelps, 1976; Manning and Phelps, 1979; Manning, Morris, Newhouse, et al., 1980). The Canadian data have the considerable advantage of a very large sample; on the other hand, they measure the effect of a copayment that was implemented at a province-wide level; if queues had decreased as a result of the copayment, the resulting change in the time price could have led to less of a decline in demand for those with high prices of time (who would be disproportionately nonpoor). But failure to find an effect in the smaller American data sets (where time price presumptively did not change) could merely reflect a lack of power.  

---

Manning and Phelps found that the absolute value of price elasticities for child dental services increases significantly with income, but this finding, of course, may be specific to dental services.
IV. DOES SUPPLIER-INDUCED DEMAND INVALIDATE THE DEMAND FUNCTIONS IN THE LITERATURE?

A serious challenge can be raised about the estimated demand functions just described. Many argue that the important determinant of demand is supply, especially the supply of physicians (Detsky, 1978; Evans, 1974; M. Feldstein, 1967; Roemer and Shain, 1959). In particular, the physician is said to have sufficient discretion in advising the patient (i.e., unexploited demand-creation ability) to offset exogenous changes in demand. But if the demand of only a small number of individuals changes, the physician has little incentive to act to offset demand changes; hence, the studies of demand cited above could greatly overestimate the response of demand to changes in insurance or to changes in demographic variables such as income and age.

Recent evidence, however, suggests that the physician’s unexploited demand-creation ability is limited, and therefore the physician’s ability to offset exogenous changes is limited.¹

1. Schwartz, Newhouse, Bennett, et al. (1980) have shown that, as their numbers have grown, specialities of a given type have systematically entered smaller and smaller towns in the United States. These data directly refute the hypothesis that specialists can gratify their supposed preference for living in larger cities by creating ever more demand. Rather, the data are quite consistent with standard economic location theory, which implies that as the number of physicians has grown, the market area surrounding each has shrunk. As a result, small towns that were previously part of larger market areas have now become the focus of new market areas. Although some amount of unexploited demand-creation ability could exist, physicians appear to locate as if they are constrained by standard market forces. An alternative explanation of the observed diffusion is that physicians’ location preferences have changed in favor of smaller towns. Direct evidence on location preferences is difficult to obtain, but such evidence as does

¹Sloan and Feldman (1978) reviewed the findings of earlier studies and found them consistent with the standard, neoclassical view of negligible or no demand-creation ability as well as with the alternative view that the physician has important unexploited demand-creation ability.
exist does not suggest any change of this nature (United States General Accounting Office, 1978).

2. During the period 1969 to 1977, controlling for specialty and location, physicians' real incomes in the United States fell approximately 1.75 percent per year (Newhouse, Williams, Bennett, et al., 1979). Data on physicians' incomes in Hadley, Holahan, and Scanlon (1979) are consistent with these figures; these same authors also show that Canadian physicians' real incomes fell after the implementation of Medicare and the subsequent influx of physicians. If physicians have sufficient demand-creation ability to fully offset adverse changes in the external environment, one presumes that they would not tolerate a decrease in real income.

3. Victor Fuchs (1978) has estimated the magnitude of surgeons' ability to induce demand. He estimates that a 10-percent increase in surgeons will induce a 3-percent increase in operations. The true elasticity figure may, however, be less if an increased number of surgeons is correlated with an increase in quality that in turn increases demand.2 Even accepting Fuchs' estimate, the extent of unexploited demand-creation ability is clearly limited.

Although one therefore cannot reject the hypothesis that physicians have some unexploited demand-creation ability, these observations cast serious doubt on the extreme version of the supply-creates-its-own-demand argument; namely, that physicians have sufficient unexploited demand-creation ability that they can or will offset any changes in market demand that might be induced by changes in insurance or other variables. Thus, demand analysis remains of interest, but how changes in demand at the market level affect utilization is an important frontier, as discussed below.

---

2Fuchs' estimates rely on a variable measuring per capita hotel and motel receipts in an area to identify the effect of supply on demand. The premise is that such receipts are higher in attractive areas and so will cause an increase in supply that is independent of demand. If, however, areas with many tourists also have more border-crossing by patients (e.g., both surgeons and hotels in very large metropolitan areas draw from larger market areas), then Fuchs' estimates are biased away from zero. A more cautious objection is that areas with many tourists may have more accidents requiring surgery; a ski resort would be an unfair example.
V. FRON TIERS OF DEMAND
ANALYSIS

In this section I speculate on the direction of demand research in the next decade. Such speculation is hazardous; almost certainly I have omitted important directions. But the risk of appearing foolish—either ex ante or ex post—will have been worth taking if the discussion serves to stimulate fruitful research.

SPECIFICATION OF THE PRICE, INCOME, AND HEALTH STATUS VARIABLES

Most empirical work has assumed that the insurance policy could be treated by one parameter—a nonvarying coinsurance rate—and has analyzed demand per unit of time, typically one year. If the coinsurance rate varies with total expenditure (e.g., a deductible or upper limit), the appropriate theoretical price variable is unobservable; moreover, it varies with the time left in the accounting period and the amount of expenditure necessary to change the coinsurance rate when decisions about treatment of a given illness are made (Keeler, Newhouse, and Phelps, 1977). A more appropriate unit of observation when insurance is of this form is the illness episode rather than total demand per unit of time. During the 1980’s, data from the Health Insurance Study should facilitate analysis in which the illness episode is used as a unit of observation, because sufficient information will be available to enable expenditures to be grouped into illness episodes and dated within the accounting period.

Specification of relevant cross-price variables will probably improve. For example, estimates will probably be made of the cross-price elasticity between outpatient mental health services and general medical services. Little is known about this elasticity, but it is clearly an important topic when one is assessing the merits of insurance coverage for outpatient mental health services. A similar comment applies to the cross-price elasticity between prescription and nonprescription drugs when one is contemplating insurance coverage of drugs.

Most empirical studies have measured income as current income (Andersen and Benham, 1970, are an exception). Yet current income suffers from two well-known defects: It contains transitory income, which will bias the estimated income elasticity toward zero if demand in fact relates to permanent income. It is also presumptively endoge-
nous because of the effect of sickness on both income and demand for medical services. Both of these problems could be avoided by using a measure of permanent income, yet economists have not agreed upon a standard method for measuring permanent income. The fact that various instrumental variables all provide consistent estimators offers rather little comfort, given the usual sample sizes. Measures that better approximate permanent income (e.g., the use of several years of income) should improve the estimation of both income elasticities and the interaction of price elasticities with income. They may also make possible a better test of Grossman's hypothesis that nonwage-related income should affect demand differently than wage-related income.

Measurement of health status variables seems likely to improve in several ways. The comprehensiveness with which health is measured will increase; e.g., physical, mental, social, and physiological dimensions rather than just self-assessed general health perceptions may be included. Greater attention will probably be paid to possible biases from random measurement error in the health status variable, either through the employment of formal reliability measures to adjust least squares estimates or the use of structural equation techniques (van de Ven and van der Gaag, 1979). Account will be taken of endogeneity of current health (Manning, Newhouse, and Ware, in press).

SPECIFICATION OF THE STOCHASTIC TERM AND APPROPRIATE ESTIMATORS

Most demand modelling has used very simple stochastic assumptions, typically assuming that the distribution of the error term is normal or lognormal and that observations are independent. Because of the Central Limit Theorem, the assumption of normality is appropriate when using aggregate (area-wide) data. But for individual data, the assumption of normality is not a good approximation. Lognormality is a better approximation, but preliminary analysis of Health Insurance Study data suggests that one can reject at the 1-percent level the hypothesis that medical expenditures come from a lognormal distribution. The actual distribution departs from lognormality in part because of a group that spends nothing during a particular time period (typically about 20 percent of the people in any one year). Additionally, the conditional distribution of annual expenditures (conditional upon spending some positive amount) has systematic departures from lognormality in the right-hand tail.

Typically, researchers using family-level data have not accounted for intrafamily correlation, yet such correlations appear pronounced
(Manning, Morris, Newhouse, et al., 1980; van der Gaag and van de Ven, 1978). In general, one can anticipate the use of more appropriate stochastic specification, which should lead to more efficient estimators.

SPECIFICATION OF THE DEPENDENT VARIABLE

Most analysts have measured demand either as expenditures per unit of time or in physical units consumed per time period (e.g., visits or hospital days per year). The relationship between physical units and expenditures will probably be explored much more fully in the next decade.

As pointed out above, measures such as visits and hospital days are seriously incomplete, because they fail to analyze the intensity of service. Ancillary services, for example, everywhere represent a substantial component of expenditure on both outpatient and inpatient services, but few analyses of demand for ancillary services have been undertaken. (Scitovsky and Snyder, 1972, estimated that the price elasticity for outpatient ancillary services was approximately half that of the elasticity with respect to visits.)

How one analyzes intensity of services depends in part on the meaning one attaches to prices. A quite conservative approach keeps the estimation as much as possible in physical units. Such an approach could extend beyond the demand for various types of medical procedures to encompass choice of physician. It is important to know the determinants of choice between, say, a specialist and a general practitioner in those countries where choice is permitted (e.g., United States, Canada), because governments in those countries tend to see the specialty distribution as a policy instrument, and their actions may affect both the supply of and demand for different types of manpower. To study how individuals choose providers, one could place providers into a discrete category (e.g., general practitioner, board-certified pediatrician) and analyze how variables such as insurance, income, and education affect choice.

Unfortunately the conservative approach just outlined tends to break down when large numbers of categories must be analyzed, and large numbers of categories will be the rule, not the exception. For example, providers differ in many ways other than specialty, and medical care consists of hundreds—if not thousands—of different types of services or procedures.

The alternative approach uses price (or an analogue) to aggregate across types of providers or services, but this approach is hardly problem-free. First, different prices appear relevant to different aspects of
the expenditure decision, implying that expenditures need to be disaggregated. Goldman and Grossman (1978), for example, point out that fixed costs such as travel and waiting times apply to units of service such as visits but do not necessarily apply to the intensity of service, or to what they define as quality, namely the type of provider used. They have built upon the theoretical work of Rosen (1974), who suggested that the analyst approach the demand for different products by first estimating an hedonic price equation and then using the implied prices of product characteristics to estimate the demand for various types of product characteristics.

Because of the Rosen method's current popularity with economists generally, and its promise for addressing the issue of quality variation, it is worthwhile to emphasize two practical difficulties with the method (both noted by Rosen). In the usual case, the hedonic price equation (the first stage) cannot be linear, or there will be no variance in the price of quality in the sample. But to justify one nonlinear form rather than another is typically a difficult task, and the estimated marginal prices could be quite sensitive to the choice of functional form.

Even assuming, however, that the nonlinear functional form of the hedonic equation is correct, a serious identification problem can arise. To identify a demand curve, one must have at least as many variables that affect the supply price of quality and do not affect demand as one has dimensions of quality. If variation in quality takes place in several dimensions, the requirement for several excluded exogenous variables can be onerous.

One clearly exogenous variable (or set of variables) is that of possible differences among local areas in factor prices or regulations that affect the cost of production. Whether the variation in factor prices that exists will suffice to yield reliable estimates is, of course, an empirical issue, but the matter seems problematical. At a minimum, one would like data from several local markets.

Goldman and Grossman (1978), who were working with data from one local market, proposed a different solution to the identification problem and the problem of variation in the price of product characteristics across consumers. They postulated that consumers faced different prices because of assumed differences in the cost of search. The variables they used to measure the cost of search, however, arguably affect demand directly, in which case identification has still not been achieved.1 Identification may be possible, however, by analyzing

---

1The variables used were race, length of residence, and location. Race may affect demand because of differences in culture (e.g., tolerance for pain; see Zola, 1956). Those living longer in an area or living in a particular subarea may systematically differ in respects other than their knowledge of the local medical care system, and such unmeasured differences may affect demand.
choices of consumers who face varying prices because of varying insurance coverage; such an approach also permits one to use a linear hedonic price equation. Rosen's methods, or extensions of them, may prove quite useful to the degree that meaningful prices can be used; as discussed below, however, observed prices in heavily insured markets may have little or no significance.

THE NORMS ARGUMENT AND TECHNOLOGICAL CHANGE

Further tests of the norms hypothesis seem likely and are important to conduct because they will shed light on the degree to which demand studies based on individual-level data yield meaningful elasticity estimates. The tests carried out to date have examined whether variation in area-wide insurance affects utilization of medical care by individuals, holding constant individual insurance coverage. The tests have not supported the hypothesis, but precision has been poor (Newhouse and Marquis, 1978). Similar tests may well be repeated on other data sets.

Another variant of the norms hypothesis deserves testing, one that focuses attention on variation across time rather than variation across space. Theory suggests that the rate of technological change in medical care reflects the market willingness to pay for new products. Suppose ideas for new products or services are generated at a constant rate, but that those actually meeting the market test reflect income and insurance coverage generally. Specifically, more meet the test as insurance and income grow. Expenditures on medical services over time could then rise more rapidly than would be predicted from cross-sectional variation in income and insurance (which holds technology constant). The markedly higher income elasticities in time series than in cross-sectional data are suggestive of such a phenomenon, although other explanations for this difference can be found. The straightforward approach to this issue requires measuring technological change and analyzing its determinants, a very difficult problem. Some progress on this issue may be possible in the next decade, but it will not be easy.

---

2Willingness to pay in most instances should be measured at the level of the world market. Temporal variation in insurance and income changes, therefore, must be measured for the entire world, not just a single country.

3Charles Phelps has pointed out to me that greater insurance can also reduce the risk to the investor, thereby raising the flow of new products.
HOW DO PRICE AND UTILIZATION RESPOND TO CHANGES IN DEMAND?

One of the major applications of demand analysis in most industries is that of predicting how price and utilization (or quantity bought and sold) will change if determinants of demand or supply change. The use of estimated demand curves in forecasting future beef prices is a well-known example. But the applicability of the beef market paradigm to medical care is in dispute; both the stability of the supply curve and the mechanisms used to clear the market are not settled issues.

As insurance tends toward completeness, the consumer loses the usual incentive to purchase from less costly, more efficient providers (Evans, 1971; Newhouse, in press). If insurance simply reimburses costs or charges, the meaning and stability of the supply curve are called into question. Inefficient producers are not necessarily weeded out, and one cannot assume that an estimated supply curve represents an aggregation of marginal cost functions, all representing efficient production.

How price is determined in markets with substantial insurance is not a settled issue either. In many, if not most, countries, price is nominally determined or negotiated by a governmental or centralized body; considerations of public finance and political economy will clearly play a role in determining price.

In other instances, such as in some types of private insurance in the United States (as well as the American Medicare program), price determination is decentralized. The literature typically assumes that the extent of insurance makes little difference to the process by which price is determined; i.e., insurance simply shifts the demand curve in some fashion, and a standard supply curve exists and is stable. Prices then adjust to equate supply and demand. This usual tale appears applicable if insurance covers only a minor portion of the market: Insurers could reimburse at a market price, determined in the usual fashion; but when most consumers are insured, the behavior of the insurer, as well as that of the consumer, becomes presumptively relevant to the reimbursement that the provider receives. Although the standard partial equilibrium model could still apply, its usefulness is open to question. How prices and utilization behave in heavily insured markets looms as one of the most important frontiers to explore.

Three examples may clarify matters. Many believe that the United States has too many surgeons, and too few primary care physicians, such as pediatricians. The study of Fuchs, Hughes, Jacoby, et al. (1972),

---

4One exception is Holahan, Hadley, Scanlon, et al., 1979; see also, Steenwald and Sloan, 1974.
for example, suggests underemployment among surgeons. Suppose these beliefs are true: what might explain them? The pattern is suggestive of markets with prices fixed at levels other than the price in competitive equilibrium. Could prices be fixed? Insurance coverage for care supplied by physicians whose services are mostly rendered in the hospital, such as surgeons, has historically been quite extensive, whereas insurance for physicians whose services are primarily delivered on an outpatient basis, such as pediatricians, has been more scanty.\footnote{Might extensive insurance coverage have fixed fees in a way that has induced excessive entry into specialties such as surgery? The standard partial equilibrium model suggests not; if prices are flexible, surgeons should not be underemployed. Alas, we seem to have no theory other than the standard model to explain how prices are set in heavily insured markets and whether they might be set so as to induce an inappropriate distribution of physicians across specialties.} A second example pertains to hospital financing. Although the private hospital in the United States is doubtless subject to some ultimate demand or product market constraint, the extent of reimbursement insurance makes it appear that the constraint may not now be binding.\footnote{Over 90 percent of hospital expenditure is reimbursed by third parties, and for many hospitalized individuals, the marginal dollar is fully reimbursed. Presumptively, there is little or no basis for price competition. This reasoning led the Carter administration to propose public regulation of hospital budgets while it was simultaneously proposing the deregulation of several other industries. Yet if the hospital is not subject to constraints imposed by the product market, what determines its production choices? Why are not prices higher even than they are? And why do certain hospitals become bankrupt, while at others labor disputes and strikes occur? If the hospital does operate under some product market constraints, what exactly is the nature of those constraints? Until we have answers to such questions, the theory of how suppliers respond to changes in demand will be incomplete.} The third example concerns equilibrating or market clearing mechanisms. When complete insurance for outpatient services was introduced in Montreal in 1970, waiting times for an appointment increased from six to eleven days (Enterline, Salter, McDonald, et al., 1973). This increase presumptively cleared the market. But why were

\footnote{There is a good economic explanation for this pattern; loading fees are much higher for small outpatient expenditures than for large inpatient expenditures.}

\footnote{Indeed, rents could still exist. One can show a significant correlation between the percentage of revenue derived from third parties and average income in a specialty.}

\footnote{Insurance in the United States typically reimburses the hospital its costs or a "price" quoted by the hospital ("charges").}
waiting times for an appointment used as an equilibration mechanism? What does theory predict about the dimensions along which equilibration will occur when price is controlled? (One can imagine other dimensions, such as longer waits in the office, shorter visits, fewer revisits, or more telephone visits.) How does the use of alternative equilibration mechanisms affect demands of various individuals? For example, do longer delays for appointments affect demand by children more than by adults because of a higher incidence of self-limiting diseases among children? The decline in the rate of visits by children in Montreal is consistent with such an hypothesis. Some beginnings on finding answers to this question have been made for inpatient services (Feldstein, 1967; Rafferty, 1971; van der Gaag, Rutten, and van Praag, 1975); almost nothing is known about outpatient services. With answers to such questions, we will better understand the effect of a change in demand upon who obtains what kind of medical service for what type of medical problem.
VI. CONCLUSION

To seriously entertain the notion that we do not have a convincing explanation for the determination of price and utilization when insurance is complete or nearly so seems heretical; if true, it would tempt an observer to characterize health economics as among the most backward of economic subfields. But however dismal the current situation in health economics, it seems vastly better than it was fifteen years ago, certainly with respect to demand analysis.

Fifteen years ago almost no theoretical work on insurance had been done; reflecting this state of affairs, empirical analyses of demand specified insurance in an ad hoc fashion. For example, they sometimes employed a dummy variable indicating that the person was insured; at other times, they used its aggregate analogue, the percentage of an area's residents with any kind of insurance. In general, empirical demand studies were more concerned with testing hypotheses than with seriously estimating elasticities.

Both conceptual and empirical knowledge are now much improved. The availability of experimental data in the 1980's should ensure further progress on many fronts. Indeed, we should have reasonably definitive answers to the important question that small-scale demand studies can answer: When one individual's circumstances change, how does that individual's demand change, assuming that the behavior of all other consumers is unchanged? Our progress will probably be slower in understanding the effects of changes in demand at the market level, but even for that prickly set of issues, the progress of the past decade bodes well for the next.
REFERENCES


Freiberg, Lewis, and F. Douglas Scutchfield, "Insurance and the Demand for Hospital Care: An Examination of the Moral Hazard," Inquiry, 13:1, March 1976, pp. 54-60.


Keeler, Emmett, Joseph P. Newhouse, and Charles E. Phelps, "Deducti-


Newhouse, Joseph P., "The Erosion of the Medical Marketplace," The Rand Corporation, R-2141-1-HEW, December 1979; also pub-


Phelps, Charles E., and Joseph P. Newhouse, "Coinsurance, the Price


van der Gaag, Jacques, and Wynand P.M.M. van de Ven, "The Demand for Primary Health Care," *Medical Care*, 16:4, April 1978, pp. 299-312.


