Chapter 9

PHYSICIAN AGENCY*

THOMAS G. MCGUIRE

Boston University, Economics Department

Contents

Abstract
Keywords
1. Introduction
2. Demand and supply for physician services
   2.1. Prices and quantities
   2.2. Entry conditions and monopoly profits
   2.3. Competition among physicians
3. Physician behavior with complete information
   3.1. A monopolistically competitive firm selling a service
   3.2. A third-party payer and administered demand and supply prices
   3.3. Noncontractible "quality," supply-side cost sharing in managed care contracts, and competition for patients
   3.4. Network incentives in managed care
   3.5. Efficient production of physicians' services
   3.6. Summary
4. Uncertainty about treatment effects and asymmetric information
   4.1. Irreducible uncertainty
   4.2. Unobservable physician actions
   4.3. Unobservable physician characteristics
   4.4. Summary
5. Physician-induced demand
   5.1. Theory of demand induction
   5.2. Physician-to-population ratios, income effects, and induction
   5.3. Fees and induction
   5.4. Other evidence bearing on PID
       5.4.1. Defensive medicine
       5.4.2. Self referrals

462
462
463
468
469
469
473
475
475
481
487
492
494
496
497
498
502
503
503
507
509
513
516
516
517

*Financial support from the National Institute of Mental Health (K05-MH01263) is gratefully acknowledged. I am grateful to Margarita Alegría, David Cutler, Randy Ellis, Richard Frank, Jacob Glazer, Miriam Hatoun, Hsien-ming Lien, Albert Ma, Joseph Newhouse, and Mark Pauly for their comments.

Handbook of Health Economics, Volume 1, Edited by A.J. Culyer and J.P. Newhouse
© 2000 Elsevier Science B.V. All rights reserved
Abstract

This chapter reviews the theory and empirical literature on physician market power, behavior, and motives, referred to collectively as the issue of “physician agency.” The chapter is organized around an increasingly complex view of the demand conditions facing a physician, beginning with the most simple conception associated with demand and supply, and building through monopolistic competition models with complete information, and finally models with asymmetric information. Institutional features such as insurance, price regulation, managed care networks and noncontractible elements of quality of care are incorporated in turn. The review reveals three mechanisms physicians may use to influence quantity of care provided to patients: quantity setting of a noncontractable service, influencing demand by setting the level of a noncontractible input (“quality”), and, in an asymmetric-information context, taking an action to influence patient preferences. The third mechanism is known as “physician-induced demand.” The empirical literature on this topic is reviewed. Theories based on alternatives to profit-maximization as objectives of physicians are also reviewed, including ethics and concern for patients, and the “target-income” hypothesis. The target-income hypothesis can be rejected, although there is empirical support for non-profit maximizing behavior.

Keywords

agency, monopolistic competition, quantity-setting, physician-induced demand, target income, insurance, price regulation, networks, managed care

JEL classification: I11, L84
1. Introduction

Physician behavior is the central issue in health economics, as many writers have recognized. Arrow (1963) challenged economists to deal with asymmetric information, absence of markets for risk bearing, and the privileged social role of physicians. Making the analogy between physician payments and return to capital in the rest of the economy, Paul (1980) observed that although profits are a small part of value added (less than the 20% of health costs represented by physician payments), return to capital directs capitalists' decisions about investment and production. Similarly, payments to physicians are laced with incentives, and these incentives, to hospitalize, to treat, to take time to diagnose carefully, and over a longer term, to select a geographic location, mode of practice or specialty of practice, direct resources in health, thus determining costs and outcomes. Fuchs (1974) aptly called the physician the "captain of the team." Managed care notwithstanding, drugs, surgery, and other health care inputs cannot be had without physician initiative and concurrence. Physicians are trained to exercise this authority.

This chapter reviews the theory and empirical research on how physicians influence the medical services used by patients. In so doing, it confronts fundamental questions about physician motives and market power, referred to together here as the issue of "physician agency." Special attention will be devoted to a new literature dealing with physician behavior in the context of incentives created by managed care. The goal of this chapter is to draw on the contributions of many writers to develop a working model of physicians that can handle the key elements of physician and patient interaction and the associated institutions. Literature reviewed will be drawn from the field of health economics.

In neoclassical theory, the firm sets price and quantity in order to maximize profit subject to the constraint of market demand. Every phrase in this paradigm has been questioned when it is applied to physician-firms. Do physicians maximize profit? Many have argued that physicians are motivated differently than other business people, that they are, for example, concerned for their patient's health, and make different tradeoffs when it comes to their own gain or the utility of their patient-customers. Are physicians constrained by market demand? Physicians are said to work at an advantage in relation to their customers because of their superior knowledge of the patient's medical condition and of what treatments are likely to be most helpful. According to this argument, physicians behave differently because they may exercise market power in

---

1 Feldman and Sloan (1988) and Gynor (1994) contain useful reviews of the market for physicians' services.

2 Medical sociology also contains a very large literature on physician and patient behavior. For one overview, see Mechanic (1990). Some of the literature in medical sociology takes a different approach to modeling choice and action, rejecting the "rational choice-based approach" employed in economics and emphasizing group structure and social networks in determining actions. See Pesosaido (1992).
ways inaccessible to other sellers, controlling, not being constrained by, patients’ demand. On the other hand, it is possible to pose the question in current medical markets, do physicians even set price and quantity? Ubiquitous third party payers set prices to patients through terms of coverage and to physicians through terms of reimbursement. Managed care plans seek to impose their will on physicians’ decisions about treatment. In light of these complications, it is not surprising that doctors and patients are not a ready application of neoclassical economic theory.

At the same time, using currently available textbooks as a point of reference, there is no consensus about an alternative approach to physician–patient interaction. Some texts present no model at all. Among the texts committing to a model, no model or approach garners more than one vote from the electorate of authors. And no text uses a model as a general way of organizing the discussion of physician behavior.

Textbook authors avoid commitment perhaps because the paradigm of the profit maximizing seller subject to a market demand constraint is not well-accepted in health, while at the same time there is no agreeable alternative. Authors of theoretical papers in journals question the neoclassical paradigm and create models to isolate the implications of some element of the physician–patient interaction (regulated prices when quality is noncontractible, for example), and while this work adds insight, the balance of the model may ignore other elements that matter for other purposes. Furthermore, fundamentals of the problem, motives, power, and imperfect information, are very complex. The social institutions overlaying doctor and patient decisions, insurance, reimbursement, and recently, managed care, themselves add to the challenge.

Empirical studies in health have a disconcerting tendency to turn up results running counter to simple neoclassical models. Writing in the inaugural issue of the *Journal of Law and Economics*, Kessel (1958) confronted the paradox of how hundreds of thousands of physicians could price discriminate, when competition among profit-maximizing firms must eliminate the practice. Kessel rejected a “charity hypothesis” as an explanation (which altered physician motives, replacing profit-maximization by

---

3 Eastaugh (1992) reviews a classification of possible models based on the two dimensions of number of sellers in the market and whether the physician can induce demand, without presenting an explicit model. Feldstein (1979) uses a simple monopoly model. Phelps (1997) presents a model of monopolistic competition, but without insurance, price regulation, or incorporation of information issues. Santerre and Neun (1996) review elementary market models and relate them to physician markets, without proposing any particular conception of the working of the market. Getzen (1997) alludes to asymmetric information and agency, and imperfect competition but presents no model of the physician-firm. Polland, Goodman and Stano (1997) refer to the principal/agent problem and the concept of perfect agent (defined as what the patient would do with the information) but also have no model. The physician faces a downward demand curve and may induce demand by moving out the demand curve. There is no integration of an agency perspective. Zweifel and Breyer (1997) is the only text I am aware of that proposes a specific model with an information/agency component. Their model is very special however, since it regards outcome as contractible (that is, a basis for payment). Most approaches in the journals regard output (quantities of care), not outcome (health), as a potential basis of payment in contracts. Rice (1998) contains a critique of the welfare economics of health care, but no explicit positive model of physician price and quantity setting.
a shared ethic to provide health care even if patients lacked an ability to pay) in favor of the hypothesis that physicians set prices as a collusive oligopoly. He accused the American Medical Association (AMA) of coordinating physician pricing, threatening sanctions in the form of denial of membership in the local AMA branches and hospital privileges if patients with lower demand elasticity (the rich) were not charged a higher price than those with higher elasticity (the poor). That an outside organization (even the 1950’s AMA) could coordinate the intimate economic exchange between patients and doctors seems, in hindsight, an unlikely explanation for price discrimination.

Perverse empirical results, with the signs of fundamental economic relations reversed, continued to emerge from more formal econometric research. Fuchs (1978) found that an increased supply of surgeons, controlling for demand factors, increased market price. Rice (1983) found that a decreased price of physician services caused an increase in supply of services. As a possible explanation, Fuchs, Rice, and others proposed that physicians sacrifice profit to pursue a “target income.” The target income hypothesis does the job in a mathematical sense, replacing the normal positive $p$ and $q$ relation governed by an upward-sloping supply curve by the negative relation given by the rectangular hyperbola $p \times q = T$, where $T$ is the target. The resolution comes at a cost, however, of introducing an objective for physicians that many regard as implausible.

Another set of mundane facts motivates the modeling efforts in health care. In a patient’s contact with the doctor, the doctor’s position is not, “here is the price of my services, how many do you want?” It is more like, “here is what you should do.” Physicians set quantity, and more generally, the treatment patients should use, and may make this decision in light of the factors affecting them. Empirical studies [e.g., Gaynor and Gertler (1995)] find that when normal demand-side variables, such as demand-price, income, and clinical need are controlled for, variables affecting the supply of care, such as supply price, physician attitudes, or partnership incentives, influence what happens to the patient. How is this to be understood?

Physicians “induce demand” is one answer. The physician-induced demand (PID) hypothesis, associated with Evans (1974), is essentially that physicians engage in some

---

4 Most of Kessel’s fascinating article is devoted to a social history of the AMA’s opposition to prepaid group practices. Kessel interprets this opposition as occurring because prepaid groups, by charging a uniform premium for membership, prevent physicians from being able to price discriminate. One can accept that the AMA’s resistance to prepaid groups was economically motivated without needing to appeal to the price discrimination hypothesis. Prepaid groups reduced demand for physician services by imposing non-price rationing, thereby reducing physician income.

5 In the first econometric test of the functioning of markets in health care, Martin Feldstein (1970) found (using a time series of national data) an upward-sloping demand and a downward-sloping supply curve, spawning another special theory of physician behavior. He proposed that physicians set prices below market clearing in order to ration demand to retain an ability to pick and choose among their patients to treat the “interesting cases.” See also, Steinwald and Sloan (1974), Sloan (1976) and Dyckman (1978).

6 In this, consider $p$ to be the margin above cost. Speculation about a target income held by doctors can be traced back to Feldstein (1970).
persuasive activity to shift the patient's demand curve in or out according to the physician's self-interest. Patients have incomplete information about their condition, and may be vulnerable to this advertising-like activity.

In recent years, the economic theory of physician behavior has emphasized contracting and information issues. Economists have recognized for a long time that a significant market is missing in the health sector: payments or insurance based on health outcomes. Arrow (1963) observed that an efficient (first-best) health insurance policy would specify payment contingent on the individual's state of health. The moral hazard problem would be resolved if an individual who suffered a sudden health problem were paid a specified amount by the insurance company; afterwards, the individual could make his own decision to purchase health care. A state-contingent payment scheme protects the individual from the financial risk of illness ex ante and retains incentives for the patient to consume health care efficiently ex post. Nevertheless, insurance policies or physician payments contingent on health status are nonexistent because health status is too costly to verify. A second contractibility problem is also important and fundamental. Some elements of treatment are not reported. Insurance coverage and provider payment are based on reports of measures such as number of "visits" or "days" in the hospital, or accounting "costs," which only partially reveal the resources devoted to treatment. A physician or other health care provider must be relied upon to prescribe the clinical content of the services connected with a "visit" or a "day," and to invest (costly) effort into making these services productive in terms of the patient's health. Thus, the physician almost always supplies her own input into the production of health care for the patient. This input, which is often referred to as "effort," but also could be understood as "quality," is simply not contractible; the market for insurance and payment policies based on the physician's effort is also missing.  

The economic problem in health care transactions can be deeper than issues of contractibility. An output (change in health status) or an input (physician effort) may not even be observable, let alone contractible. If output or some inputs are known to the physician but unobservable to the patient (and a third-party), the problem of asymmetric information is introduced. Information asymmetry is used to motivate many papers in health economics, even if the information issues are not modelled explicitly.

Following the introductory material in Section 1, Section 2 begins with the simplest model of the physician market, demand and supply. Historical and current information about prices, income, supply, and specialization are presented and interpreted within a demand and supply framework. The possible collusive role of monopolistic practices, such as restrictions on entry into medical school are considered. This first economic perspective on the market remains useful for certain purposes, particularly for understanding trends in physician income and supply. As is well-known, demand and supply

---

7 Ma and McGuire (1997) point out that even "visits" or "days" are not contractible, only the reports of visits or days are contractible. The requirement that patients and physicians report usage truthfully puts limits on the ranges of feasible insurance-payment systems.
are generated by price-taking buyers and sellers. And while the demand and supply framework has some utility, it does not come to grips with the special decisionmaking processes that characterize patients and physicians. To do so requires an explicit model of the behavior of the physician firm (and its customers). For the remainder of the chapter, the focus will be on behavior at the firm level.

Section 3 places physician-firms in a monopolistically competitive market, the versatile market structure favored by health economists. It addresses models of complete information—that is to say, models with no uncertainty on the part of either the physician or the patient about the benefits of medical care (and thus, no asymmetric information). Patients in this section have a stable set of preferences. For institutional reasons, we regard prices as being set by payers. Physician quantity-setting, however, emerges naturally within this model. The nonretradability feature of physician services implies that profit-maximizing physicians do not allow patients to choose the utility-maximizing quantity given the price the patient pays. This is the first of three ways a physician can influence quantity identified in this chapter.

Within a model of complete information, physicians can decide about quality or effort—a key input into health, which, even though observed by patients, may be impossible to verify. This is the second mechanism for physicians to influence quantity: through choice of a noncontractible input (quality) that influences patient demand. The input is observable by the patient (hence its effect on choice) but cannot be paid upon by a payer. Insurance, price regulation, physician contracting, and managed care are all introduced in Section 3.

Section 4 addresses information issues, resting, as Arrow (1963) and others have argued, at the heart of the physician-patient relationship. This section first considers the effect of shared uncertainty between patients and doctors, and then introduces the important element of asymmetry of information between patient and physician. When patients believe doctors know more than they do, doctors may be able to persuade patients to demand more or less care. For this to work, information must be asymmetric, and the physician must be taking an unobservable action to influence demand. This third mechanism for quantity determination captures the meaning of "physician-induced demand" as it is used in the literature.

As a preview to the material covered in Sections 3 and 4, Table 1 contains a summary of the three mechanisms—direct quantity setting stemming from nonretradability, altering demand by setting an observable but noncontractible input (quality), and persuasion based on asymmetric information. Some of the empirical evidence that supports the existence of these mechanisms is discussed in Sections 3 and 4, though these sections are primarily theoretical. Sections 3 and 4 study the implications of physician profit maximization within increasingly complicated demand environments.

The large body of empirical research in health economics concerned with the existence of induced demand is reviewed in Section 5. Although there has been a clear and widely accepted definition of physician induced demand (PID) for more than 25 years, the measurement and meaning of PID has been one of the most contentious issues in health economics.
Section 6 is concerned with physician motivation, and gives consideration to alternatives to profits as objectives guiding physician behavior. Standards of practice and ethics, concern for patient welfare, and pursuit of a target income are all considered theoretically and empirically.

In undertaking a review of physician behavior, a decision has to be made about how much respect should be paid to the conventional model of a profit maximizing firm constrained by market demand. As an expedition in "normal science," a review regards observations and empirical findings in health economics as puzzles to be solved, if possible, within the broad paradigm of neoclassical economics. On this front, Pauly (1980, p. 177) advised some time ago:

We should not be too quick to depart from standard maximizing economic models in order to explain behavior in the medical care industry. Supposedly anomalous features of that industry sometimes vanish when more appropriate sets of data are used, while other apparent institutional differences require only redefinitions of price, ownership-entrepreneurship, and markets to make models analogous to the traditional ones applicable.

Kuhn’s theory or paradigm is “an object for further articulation and specification under new or more stringent conditions” (1970, p. 23). At some point, of course, the articulations become so unwieldy that the standard paradigm should be rejected altogether. But this judgment can only be made after a systematic effort to integrate the ideas and findings within the dominant paradigm, and in the presence of an alternative theory.

2. Demand and supply for physician services

The demand and supply model of physician markets has appeal when the object of study is the industry, not the firm, and the focus is on aggregate supply, as measured by
number of physicians, or on prices as measured by average wage or annual income. Acknowledging that a more complex model is necessary to explain the details of price and quantity setting at the patient level, competition and entry conditions nonetheless govern income and average compensation. Historically, the demand and supply perspective has been used to study the effects of restrictions on entry into medical school, for example. As physicians begin to sell their services to organized buyers in managed care, the forces of demand and supply can be expected to determine the terms doctors can expect in those contracts.

2.1. Prices and quantities

Physician incomes throughout most developed countries are very high, among the highest for any occupational group. In the 1970’s and 1980’s, physician earnings in the US grew relative to college graduates and lawyers [Gaynor (1994)]. In 1994, average physician net income, according to the AMA, was $182,400.\(^8\) Some statistics about the distribution of physicians by specialty and their incomes are contained in Table 2. The number of physicians supplying health care is regulated by licensing laws, including provisions for students trained outside the US, and the capacity of the 126 US medical schools. About 16,000 students graduate from U.S. medical schools each year, and these are joined by 5,000 immigrant physicians. The accumulation of these flows has built the current stock of 550,000 physicians in active practice, a stock that is growing at about 1.5% per year. More than half of all physicians are part of a group practice.

In terms of physician-to-population ratios, for various years in 1990’s, the US at 254 physicians per 100,000 population is higher than the UK (164), Japan (177), and Canada (221), but lower than many other high-income countries, such as France (280), Sweden (299), Germany (319), or Belgium (365).\(^9\)

Physicians average about 55 hours of work per week. Figuring 48 weeks per year, the average physician net wage was $65 per hour in 1994. Net is about half of gross (Table 2 contains details and sources) so the average price charged by physicians in 1994 could be figured at approximately $130 for each hour of their time. This of course varied according to the specialty of the physician, and the work activity involved. The geographical distribution of physicians at the state level has been successfully studied using a basic demand and supply model [Benham et al. (1968), Fuchs (1978), Frank (1985)].

2.2. Entry conditions and monopoly profits

At one time, there were few institutional interferences in the operation of market forces in physicians’ services. In the middle of the 19th century, skilled laborers earned more

---

\(^8\) 1994 was the first year in which physician earnings fell in nominal terms, according to data collected by the AMA [Simon and Born (1996)].

\(^9\) Information on physician population ratios is from various years in the 1990s, available from the World Health Organization at www.who.int/whosis.
Table 2a
Physician income 1994

<table>
<thead>
<tr>
<th>Specialty</th>
<th>% of total</th>
<th>Average income</th>
</tr>
</thead>
<tbody>
<tr>
<td>All physicians</td>
<td></td>
<td>$182,400</td>
</tr>
<tr>
<td>Specialty</td>
<td></td>
<td></td>
</tr>
<tr>
<td>General/family practice</td>
<td>15%</td>
<td>$121,200</td>
</tr>
<tr>
<td>Pediatrics</td>
<td>8%</td>
<td>$126,200</td>
</tr>
<tr>
<td>Psychiatry</td>
<td>7%</td>
<td>$128,500</td>
</tr>
<tr>
<td>Internal medicine</td>
<td>28%</td>
<td>$174,900</td>
</tr>
<tr>
<td>Pathology</td>
<td>3%</td>
<td>$182,500</td>
</tr>
<tr>
<td>Obstetrics/gynecology</td>
<td>6%</td>
<td>$200,400</td>
</tr>
<tr>
<td>Anesthesiology</td>
<td>5%</td>
<td>$218,100</td>
</tr>
<tr>
<td>Radiology</td>
<td>5%</td>
<td>$237,400</td>
</tr>
<tr>
<td>Surgery</td>
<td>14%</td>
<td>$250,200</td>
</tr>
<tr>
<td>Type of practice</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employee or contractor</td>
<td>33%</td>
<td>$158,400</td>
</tr>
<tr>
<td>Self-employed</td>
<td>67%</td>
<td>$210,200</td>
</tr>
<tr>
<td>Solo - 49%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 + MDs - 51%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Income in net of practice expenses.

Table 2b
Physician expenses 1994 (office practice costs)

<table>
<thead>
<tr>
<th>Average expenses of self employed</th>
<th>$183,100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distribution</td>
<td></td>
</tr>
<tr>
<td>Non-physician employee wages</td>
<td>35%</td>
</tr>
<tr>
<td>Office rent</td>
<td>26%</td>
</tr>
<tr>
<td>Medical supplies</td>
<td>8%</td>
</tr>
<tr>
<td>Malpractice liability insurance</td>
<td>12%</td>
</tr>
<tr>
<td>Equipment</td>
<td>3%</td>
</tr>
<tr>
<td>Other expenses</td>
<td>16%</td>
</tr>
<tr>
<td>Total</td>
<td>100%</td>
</tr>
</tbody>
</table>

Source: Getzen (1997), from AMA publications.

than the average physician [Starr (1982, p. 84)]. Licensure was uneven among the states. Physicians came in many different flavors, with homeopath, who believed like cured like, competing with, among other groups, allopats, who believed in cures by opposites [Frehch (1996, pp. 52–53)].

The free market in curing was eliminated by the AMA's lobbying for uniform national licensing. Between 1880 and 1890, every state licensed medicine. With the reorganization and growth of the AMA between 1900 and 1910, and a rise in membership from 8,000 to 70,000 [Starr (1982, pp. 110, 112)], membership began to be associated with hospital privileges and control of expert witnesses in malpractice cases. Reform of
medical education was also a goal of the AMA. State medical boards were beginning to regulate medical schools. Flexner accelerated the process and helped transform medical training into its present form. In his 1910 report, Medical Education in the US and Canada, Flexner recommended an increase in the quality of medical education, uniform training (based on the Johns Hopkins model), and a decrease in the number of schools and students per school.

Since the Flexner Report, the number of medical schools and their capacity has been subject to control by the AMA. Between the period of 1910 and 1965, restrictions on the number of physicians trained decreased the supply of physicians in relation to population, from about 1.6 physicians per thousand in 1910 to less than 1.3 in 1965, a remarkable economic achievement during a period of rapid growth in income and insurance coverage driving up per capita demand for medical care.\(^\text{10}\)

The federal government intervened with the Health Professions Educational Assistance Act of 1963, increasing the capacity of US medical schools and making immigration easier for foreign-trained physicians. Output of medical schools doubled to 15,000 between 1965 and 1975. Since 1965, the physician-to-population ratio has crept steadily upward.\(^\text{11}\)

A demand and supply framework can be used to address whether the limited supply of physicians has protected physician incomes and generated rents. There are more than two applicants for every spot in medical school. If medical school tuition were set at cost, this in itself would be prima facie evidence for a restriction on entry leading to rents. However, since medical school tuitions only cover a fraction of the cost of education [Ganem et al. (1995)], excess demand for spaces at the subsidized price does not necessarily imply the number of doctors is below the quantity at which demand would equal supply at a market-clearing price for training. The number of physicians could be at the competitive level, but the cost of training is subsidized, perhaps to enable medical schools to collectively choose a pool of new physicians with desirable social characteristics (“high quality,” ethnic diversity, social consciousness, geographic distribution).

Figure 1 contains a simple demand and supply diagram for medical school places. The number of places at medical schools is set by administrative policy. Tuition is subsidized. We observe then an excess demand at the subsidized price. If there were no

---

\(^\text{10}\) One of the side effects of central control of medical school admissions was the spread of discriminatory admission practices. Medical schools increased discrimination against women, blacks, and Jews. Some medical schools stopped admitting women, or limited enrollment to five percent of the students [Starr (1982, p. 124)]. Black medical schools fell from seven pre-Flexner to two post-Flexner. As a result, the percentage of physicians who were women or who were black fell for fifty years after Flexner [Frech (1996, p. 54)]. Jews suffered quotas. For example, more than half of the applicants to Cornell Medical College in New York City in 1940 were Jewish, but the school limited the number of admits to 10-15 percent, making it ten times harder to get in for a Jew than for a non-Jew [Frech (1974, p. 125)].

\(^\text{11}\) The Health Professions Educational Assistance Act of 1976 restricted the inflow of immigrant physicians. Hoetner (1986) has emphasized the increasing role of competition from foreign medical graduates (FMGs) in American medicine.
subsidy, the number of places demanded would be less than are demanded at the subsidized price. It is impossible to tell if the quantity demanded at the full price is more or less than the regulated quantity. Figure 1 illustrates one possibility, that they are equal.

A large number of empirical studies have looked at the question of entry restriction, attempting to measure any "excess return" on the student’s investment in medical education. In an era before federal subsidy of medical education, Friedman and Kuznets (1954) found evidence of monopoly returns that they attributed to AMA restrictions on entry. The presence of a tuition subsidy for medical schools (which does not generally exist for law or business schools) implies that, if supply were restricted to the competitive level, even the marginal physician would enjoy rents (equal to the subsidy – see Figure 1). More recent studies which evaluate the return to tuition investment contend with evaluating students' opportunity costs and correcting for physicians' long hours, without a clear finding of excess return due to entry restrictions in medicine [see, e.g., Sloan (1970), Leffler (1978), Marder et al. (1988)]. Weeks et al. (1994), for example, studied the return to education for physicians, lawyers, dentists, and MBAs, and found post-high school rates of return of 20–25% for all groups.

It seems very plausible that during the first half of this century, restrictions on the output of medical schools elevated physicians' economic position. When supply is constant and demand increases a lot, sellers benefit. As we have noted above, during the 50 years between 1900 and 1950, the number of new medical school graduates held
constant. Meanwhile, using figures from the Statistical Abstract of the United States, the population of the US doubled (from 76 million to 151 million), the average age of the population increased from 23 to 30, and real per capita Gross Domestic Product increased by a factor of six. Insurance coverage was also beginning to be prevalent by the end of this period, and doctors were becoming more competent in contending with disease. All of these factors, population, age, income, insurance, and technologically driven increments in value, increased demand.

More recent studies, using data on incomes, correcting for opportunity cost and hours worked, do not support a clear conclusion that at the current rate of production of medical schools, the output of doctors keeps physicians' incomes above a competitive level. In the last 10–20 years, outputs of medical schools have expanded considerably. In addition, market forces may be responsible for the dissipation of rents created by the original restrictions on entry. Buyers of health care – hospitals, insurers, prepaid plans, as well as individual patients – may have substituted against high-wage physicians, making more use of nurses and other "physician extenders." If physician net incomes include a rent, this will tend to attract applicants, allowing medical schools to choose students who are "high quality," exactly those who are likely to have a high opportunity cost. Some rents may have been transferred back to buyers by price restrictions. Finally, as Gaynor (1994) has argued, increasingly aggressive antitrust activity directed at physicians may have inhibited collective physician exercise of market power.

2.3. Competition among physicians

For a given number of physicians, it is in the interest of the group to minimize competition among themselves. Kessel (1958) proposed that organized medicine operated a pricing conspiracy. An alternative to price collusion is an agreement to divide the market, a generally more effective way to inhibit competition because it prevents unwanted competition on non-price as well as price dimensions. Market division maintains the inelasticity of each practitioner's demand.

The counter strategy by buyers is to increase the sellers' demand elasticity. Buyers' interest in increasing the elasticity of demand facing sellers has been appreciated in health care markets. As Dranove, Shanley and White (1993) have argued in the case of hospitals, and Scherer (2000) in the case of pharmaceuticals, when patients make choices of medical supplier, demand for suppliers is likely to be inelastic. When organized buyers (insurers, HMOs) make choices, however, demand can be more elastic, driving prices downward. The same is likely to hold true in physician markets. Organized medicine worked for many years to protect the "sacred" physician–patient relationship, ensuring that patients, not third-parties, made choices of doctors. As Reinhardt (1996, p. 9) recently put it, "Until about the mid-1980's the patient's freedom

12 The number of other health personnel, such as nurses, did increase substantially over this period. While this had a mitigating effect on the restriction of supply of doctors, it is further evidence that there was a large increase in demand for health care services.
to choose among the doctors and hospitals at the time of illness was sacrosanct in the United States” [his emphasis]. This institution of patient choice was maintained by political activity by organized medicine. In a losing battle, the AMA declared “contract medicine,” whereby a doctor might be hired by an employer to provide health care to its workers, unethical in 1969 [Freh (1996, p. 69)]. Kessel (1958) and Havighurst (1978) document organized medicine’s longstanding opposition to prepaid group practices. Scholars new to health economics may find this surprising since today, physician contracting takes so many forms, and prepaid group practices are a well-established set of institutions.

Notwithstanding attempts to prevent intraprofessional competition, there is ample evidence that competition among physicians, and between physicians and substitutable professionals, is in force. Frank (1985) estimated a demand and supply model of physician pricing and location, finding a standard relationship between price and quantity. Frank et al. (1987) studied earnings of podiatrists, foot specialists who compete primarily with orthopedic physicians. More competition from orthopedists as well as within their own specialty reduced earnings, even in a model in which any simultaneity between practitioner per population and earnings was not considered. Newhouse et al. (1982) found that the geographical distribution of physicians was consistent with competitive location theory. Benham et al. (1968) and Escarce et al. (1998) provide evidence that locational choices respond to demand conditions.

The demand and supply framework is useful for discussing aggregate trends in physician earnings, earnings by specialty, and earnings by region. [See, e.g., Simon and Born (1996).] Physicians are relying on contracts with third parties for a rapidly increasing share of their earnings [Emons and Wozniak (1997)]. In a recent survey of health plans, Rosenthal et al. (1999) reports that in 1998, 63% of HMOs in California pay most of their primary care practitioners by some form of capitation. The terms of these contracts, e.g., the level of capitation, are likely to be governed by demand and supply forces, and represent an important new area for research. As we noted in the beginning of this section, demand and supply analysis will be useful in study of the overall level of physician compensation, particularly when the buyer is a managed care plan, in a position to substitute one physician for an other, or to substitute other personnel for physicians.

Competitive analysis relies heavily on the use of a “zero-profit” condition. Entry drives profits at the margin to be equal to zero, and this analysis can be used to understand specialty and location, capitation payments, or other issues in compensation. At the level of the individual patient, however, entry conditions are not sufficient to understand the question of quantity setting. To attend to treatment determination requires us

---

13 For example, the Group Health Association (GHA) in Washington, DC, an early prepaid group was deemed to violate the AMA’s code of ethics. Physicians worked for a salary in an organization not controlled by doctors. Furthermore, patients’ choice of doctor was restricted to a physician member of the plan. The local society expelled or disciplined physicians affiliated with GHA, and prohibited members from associating with GHA physicians. The local society was successfully prosecuted by the Justice Department for antitrust violations. See Havighurst (1978).
to move beyond demand, supply and entry conditions, and to study the implications of profit maximization.

3. Physician behavior with complete information

In this section we assume that demand conditions are given to the physician. Patients have fixed preferences for health care, and physicians can take no action to change those preferences. Furthermore, we assume in this section that information is complete. The physician recognizes patient preferences, and patients can observe and evaluate the characteristics of care supplied to them by doctors. These assumptions can be defended for at least some areas of medical care. Pauly (1988) has argued that patients can be regarded as being reasonably well-informed for about one quarter of the care they consume, such as for routine care and for care of chronic illnesses. The assumption that patients can accurately judge the value of the health care they receive is implicit in the large literature on demand which employs evidence about demand response to infer implications for optimal insurance. [See, e.g., Manning et al. (1987), Feldman and Dowd (1991); see Rice (1998) for a critique.]

We also assume in this section that physicians maximize profit. Eisenberg (1986), a physician, contends that physicians are motivated by financial self-interest, concern for their patients, and concern for the social good, devoting a chapter to each in his thoughtful book on physician decision making. His first chapter is about self-interest, and that will be our starting point too.

As we will see, in a complete information model, physicians possess two ways to influence quantity. First, with market power, the physician can set the quantity of a non-retradable service (Sections 3.1, 3.2). Second, if there is an element of noncontractible quality in medical care supply, the physician can influence quantity by choice of this input (Section 3.3).

3.1. A monopolistically competitive firm selling a service

In virtually all characterizations of physicians in economics journals and textbooks, the physician is portrayed as having some market power. Monopolistic competition is a versatile structure for representing market power, and is the expressed favorite of many writers [Frech and Ginsburg (1975), Pauly and Satterthwaite (1981), McGuire (1983), Klevorick and McGuire (1987), Dranove (1988), Dranove and Satterthwaite (1991, 1992, 1999), Getzen (1984), Zuckerman and Holahan (1991), Pauly (1979, 1991), Phelps (1997), Frech (1996), Newhouse (1978), Folland et al. (1997), Gaynor (1994)]. Monopolistic competition includes an element of monopoly (downward-sloping demand) and an element of competition (large number of competitors – each firm ignoring strategic interactions). Because of location, specialty, quality, or some other element of taste, patients do not regard physicians as perfect substitutes. Information imperfections could also generate a monopolistically competitive structure. There may in fact be good
substitutes for a given physician, but if the patient doesn’t know who these are, the patient may be willing to pay more for the services of the familiar doctor, an idea proposed by Pauly and Satterthwaite (1981). For now, however, we will keep informational issues in the background and simply assume that the structure is monopolistically competitive because of recognizable differences among physicians.

A point in favor of monopolistic competition is that it is an appealing alternative to models that rely on collusion among physicians to explain some observed patterns of price and quantity. Consider Fuchs’ (1974, p. 71) story of the “surgeon surplus:” “A comprehensive, detailed study of general surgeons in one suburban community in the New York metropolitan area revealed that the surgical workload of the typical surgeon was only about one-third of what experts deemed a reasonably full schedule.” Fuchs further characterized the market as follows: “For most types of surgery, the quantity physicians would like to supply at the going price is far greater than the quantity demanded.” How can this excess supply persist? Why does competition fail to reduce price and increase demand and workload? A conspiracy is the explanation that Kessel might have proposed. Surgeons might be colluding to keep prices high, to maximize their joint profits. But more plausibly, as Fuchs notes (p. 73), “Many surgeons believe, perhaps rightly, that demand would not increase appreciably in response to a price cut . . . ,” in other words, each physician faces a downward-sloping demand. Collusion is not required to observe price above marginal cost and for there to be “excess supply.”

The classic evidence for a monopolistically competitive structure is a demand curve with a negative slope [Haas-Wilson (1990), Kleavorick and McGuire (1987), McCarthy (1985), McLean (1980)]. If the instrument for competing for patients is quality [Gaynor and Gertler (1995)] or even the aggressiveness in “inducing demand” [Dranove (1988)], evidence that such a decision variable influences demand supports the monopolistically competitive assumption. Many other empirical features of the market are also consistent with monopolistic competition. Studies of physician practice costs conclude that physicians operate on a downward-sloping portion of their average cost curve [PPRC (1992), Escarce and Pauly (1998)]. Firm-level advertising only pays if there is imperfect competition [Feldman and Begun (1978), Haas-Wilson (1986), Rizzo and Zeckhauser (1990)]. Wong (1996) found that physicians respond to factor price changes in a manner consistent with monopolistic competition.

The basic conception of the physician-patient relationship embodied by monopolistic competition is that physicians are imperfect substitutes in the eyes of patients. A patient has a demand for the services of a particular physician, as opposed to demand for “physicians’ services” in general. Although some observers have written about the “demise” of the physician-patient relationship [see Sloan et al. (1993, p. 51) for one discussion], surveys continue to show that a clear majority of patients have what they

\[14\] Much, but not all of surgical care was insured during the 1960’s. Many patients, including the elderly, faced some price and if they were price responsive in choice of physician, would have given surgeons an incentive to reduce price to raise volume.
regard as a "regular source of care" [Moy et al. (1998)]. Even without a primary care
provider, patients may rely on physicians to supervise their care during an episode of
illness. The demand curve a patient has for a physician is not the same as the demand
the patient has for physicians' services, a distinction, despite the general enthusiasm for
the monopolistically competitive structure that has not been given much attention in the
literature.\(^{15}\) Many papers motivate their model with words associated with monopolistic
competition, and then analyze a single firm.

Although interaction is not strategic in monopolistic competition, actions of one
physician, such as a price change, affect demand for other physicians. In what follows,
we present a model of monopolistic competition in which the physician has some mar-
ket power but the patient has some alternatives. We will model the patient's alternatives
as simply as possible in order to enable us to focus on the behavior of a representative
physician.

Another important feature of the market will also be taken into account. Physicians
sell a service; a diagnosis or treatment provided to one patient cannot be resold by
that patient to some other customer. As Gaynor (1994, p. 224) observes in his review,
"services are by their nature inherently heterogeneous and nonretradable." The nonre-
tradability of physician services has important implications for price discrimination and
more generally for price and quantity setting. Farley (1986) called attention to the con-
nection between non-retradability and price discrimination in physician markets, but
the implications of nonretradability have not been fully appreciated in the context of
physician markets.\(^{16}\)

We can now proceed to set up a model of patients and physicians that we will build
on throughout this chapter. The quantity of physician services is \(x\). The patient benefits
from services according to \(B(x)\), denoted in dollars. The marginal benefit function is
\(b(x) = B'(x)\). \(b(0) > 0; b'(x) < 0\). We employ a benefit function rather than a demand
curve since profit maximization implies price-quantity pairs that may not be "on" the
demand curve. The \(B(x)\) function captures any health shocks implicitly, so that \(B(0)\)
may be negative. Time costs, inconvenience, and other costs and benefits of using med-
cal care experienced by the individual are incorporated in \(B(x)\).\(^{17}\) By assuming that the

\(^{15}\) The hospital market is also generally regarded as monopolistically competitive. See Pauly and Redisch
(1973) or Draiove, Shansky and White (1993). Frank et al. (1987) estimate an earnings model for podiatrists
based upon the idea that medical practitioners appropriate hospital rents. Draiove and White (1996) have
another view about how doctors can secure rents from monopolistically competitive hospitals. Since each
specialty group's fees (e.g., cardiologists) fees are a small part of each hospital's costs, and patients (or
payers) decide on the hospital on the basis of total cost, each specialty will try to raise its fees as much as
possible to enlarge its share of the rents.

\(^{16}\) Polland, Goodman and Stano (1997, p. 377) recognize that physician services are nonretradable and sup-
sport price discrimination. Draiove's (1988) model of demand inducement implicitly recognizes this prop-
erty by including the assumption that the patient's choice is to "consent" to treatment or not. In the general lit-
erature, it is well understood that nonretradability is behind models of price discrimination [Varian (1989)].

\(^{17}\) The utility function generating this benefit can be expressed as \(U(y + B(x))\). Demand for \(x\) is independent
of income in this formulation. See Ma and Riordan (1998) for discussion of the implications of alternative
forms of utility and benefits.
benefit function depends only on the quantity of \( x \), we abstract from the role of other goods, including income, influencing the valuation of services. Physician services are produced at constant cost per unit \( c \).\(^{18}\) If \( p \) is the price of physician services (insurance will be introduced shortly), physician profit is \( \pi = px - cx \), and patient net benefit can be written \( NB(x) = B(x) - px \). Define \( x^* \) as the solution to \( b(x) = c \), the efficient level of \( x \). Let \( NB^* = B(x^*) - cx^* \), the maximum possible patient net benefit. Also, for purposes of reference define \( x^m \), the level of \( x \) that maximizes \( B(x) \), or, the solution to \( b(x^m) = 0 \). See Figure 2.\(^{19}\)

In monopolistic competition, the patient has substitutes. In general, a patient could consume services of many physicians at the same time, and benefits from physicians' services would be a function of the set of services consumed. We simplify this by forcing

---

\(^{18}\) Empirical research indicates that AC is falling. This is consistent with fixed costs and a constant marginal cost. See Escarce and Pauly (1998), and Physician Payment Review Commission (1992).

\(^{19}\) It is worth calling attention to an assumption that the decision to be made by the doctor and the patient is a decision about the “quantity” of one variable, \( x \). Often in papers on health care, this quantity is denominated in dollars. While this follows convention in economic models, the literature in medical sociology or medical decision analysis views the matter differently, with the doctor and the patient having to decide about a “treatment” which can consist of a mix of different services. Typically, there are a discrete number of treatment alternatives, described, for example in a decision tree. See, e.g., Weinstein et al. (1980).
\[
B N - (x)B = d
\]

\[
(x) B - d = (x)B
\]

The above equations can be solved numerically to find the roots of \( x \) and \( d \). The key-order conditions (Einstein) in this context are:

\[
\{(B N - x d) + c N - x d = 0 \}
\]
total surplus available. Note that the patient is not a price taker. At the price of $p$, the patient would prefer to consume fewer services than $x^*$ but nonretradability lets the doctor set quantity. One can think of the consumer surplus gained above $NB^0$ by consuming up to the point where $b(x) = p$ (the moderately shaded region in Figure 3) as just being offset by the consumer surplus lost from consuming beyond this point to $x^*$ (the dark region). In effect, the physician makes an all-or-nothing offer to the patient, extracting all available consumer surplus. This is not surprising, since with market power and the nonretradability feature, the physician possesses the prerequisites for the exercise of first-degree (or perfect) price discrimination [Varian (1989)].

In Kessel’s world of the 1950s, physicians could set prices (and quantities) without contending with third-party regulations. Price discrimination across patients emerges naturally from the model in Program I. Consider different patients with different benefit functions. Suppose one patient has a higher willingness to pay indicated by a higher $B(x)$. Equation (3.4') tells us immediately that the higher willingness-to-pay patient will pay more for the same services. Nonretradability shelters the price discrim-

---

20 It is evident from Figure 3 that if $NB^0 = NB^*$, the physician is forced to give the consumer $x^*$, at the competitive price, $c$. 
\[ 0 > \frac{d}{dx} \text{ where } x \neq c \]
instituted in 1966, adopted a similarly permissive UCR system. The UCR system could limit fees paid in any one year, but gave an incentive to physicians to increase charges, eventually forcing payers to set dollar amounts to be paid by procedure for various specialists.\footnote{Data produced as part of Hsiao's fee reform research show how out of alignment Medicare fees had become by the patchwork regulatory system in force through the 1980's. For the single highest dollar volume procedure, a cataract removal, Medicare paid more than $1500 per hour of work in 1986. For the second highest volume procedure, an office visit, Medicare paid less than $100 per hour. See Glazer and McGuire (1993) for more information on high-volume fees prior to reform.}

In the US, Medicare, state Medicaid plans, Blue Cross plans, and within recent years even commercial insurance, set maximum prices they will pay physicians for each procedure\cite{Eisenberg1994}. In a comprehensive study of physicians' practices in the late 1980's, Hsiao and colleagues\cite{Hsiao1988, Hsiao1988b} devised a relative weighting scale for physician services that forms the basis of Medicare's physician fee schedule. Some other payers pay based on Medicare's schedule, while others use their own fee schedule, based on historical payment or on demand and supply conditions. Outside the US, when governments pay physicians by fees, these fees are regulated, as in Canada.

Reinhardt\cite{Reinhardt1975} classified physician markets in two dimensions, price setting vs. price taking for physicians, and on the dimension of inducing demand or not. The price setting/price taking distinction is the traditional difference between a firm with market power facing a downward-sloping demand (conferring some price-setting power) and a firm with no market power facing a horizontal demand at the market price. Payers have stripped physicians of price-setting power. Does this mean that physicians have no "market power?" Equation (3.4) tells us the answer to this question is no. When payers set price, market power continues to convey an advantage to physicians, as we see in this section, because physicians retain the ability to set the quantity of their nonretradable service.

Third-party payers may also regulate the prices paid by patients. Insurance or other form of third-party payment reduces the financial price to the patient at the time services are used. The price paid by the patient is less than the price received by the physician, the difference being the amount contributed by the third-party payer. Prices patients pay can be complex, involving deductibles, copayments, and limits; here we will assume the price a patient pays is a constant share $\theta$ of the price paid to the physician, $\theta$ is the coinsurance rate. Thus, $0 < \theta < 1$. Third-party payers also set the price doctors receive, a practice that can be understood as an attempt to prevent the increase in willingness to pay of patients created by insurance from being shifted to physicians in the form of a higher price. The price paid to the physician, $p$, is now set by the insurer, with $p > c$, to ensure physician participation.

Physician profit depends on $p$, and patient net benefit depends on $\theta p$. As before, the patient has a benefit function $B(x)$ and has an alternative offering net benefit of $NB^0$. Physician profit maximization with a regulated price is described in Program II. Quantity $x$ is the physician's only decision variable.
Program II:

\[ L = px - cx + \lambda (B(x) - \theta px - NB^0), \]  \hfill (3.6)

with the first-order conditions:

\begin{align*}
L_x & : \quad p - c + \lambda (b(x) - \theta p) = 0, \\
L_{\lambda} & : \quad B(x) - \theta px - NB^0 = 0. \hfill (3.7)
\end{align*}

So long as \( p > c \), the physician profits from more \( x \). Rewriting (3.8) as (3.8') we see that quantity is set so as to just satisfy the net benefit constraint. We label the solution to (3.8') as \( x' \), and illustrate it in Figure 4:

\[ x' = \frac{B(x') - NB^0}{\theta p}. \hfill (3.8') \]

Equations (3.8') and (3.7), and Figure 4 show that so long as \( p > c \), \( \theta p > b(x') \), since \( \lambda \) is positive.\(^{22}\) When price is constrained, the doctor exercises market power by

\(^{22}\) \( \lambda \) can be greater than or less than one, depending on how much \( p \) exceeds \( c \). If profits per unit sold are very small, the value of a relaxation in the constraint to the physician will be less than a dollar. The opposite case is also possible.
setting quantity beyond the point the patient would choose given the price he faces. The physician wants to do so because she makes a profit on every unit sold, and services are nonretratable. The patient’s alternative of leaving to receive $NB^0$ limits how far $x'$ can be pushed. Another constraint, that $b(x) \geq 0$, could reasonably be added to the problem, limiting the quantity setting to quantities which convey some non-negative benefit, but this would not change the essential character of the result.

It is plausible that physicians can require patients to use more than they would like given the prices they face. A physician can put pressure on a patient to agree by conveying that if the patient does not accept the treatment, his alternative is to seek care from another practitioner. Patients in some cases may be able to avoid some overtreatment (from the patient’s economic point of view) by failing to comply with prescribed treatment after some point. This may be done with visits that must occur over time, for example, and after some point the patient can simply stop going to that physician. Even in the case of visits, however, the physician may be able to exert some influence. Hickson et al. (1987) found, when paid a fee with a solid margin over cost, pediatricians scheduled well-baby care in excess of that recommended by the American Academy of Pediatrics, but did not do so when they were paid a salary. Physicians recommend more treatment for insured patients, even in artificially constructed clinical scenarios (Mort et al. (1996)). Chassin et al. (1987) found in a fee system that a sixth to a third of commonly performed procedures provided no (or negative) marginal benefit.

For many treatments, an all-or-nothing quantity setting strategy may be very effective for the physician. Suppose there are few or many tests that could be run. The patient would like “few” given the price, but wants his doctor, who insists on “many,” to administer and interpret the tests. Treatment might be simple or complex, but the patient’s physician might insist on complex. The many medical situations in which treatments are provided if and only if both physician and patient agree fit squarely within the model of physician quantity setting.

The welfare economics of health insurance have been based on the assumption that the patient is a price taker in medical markets. If this assumption is not correct, and as we have seen, it is contradicted by the assumption of a profit-maximizing seller of a service, the analysis of optimal insurance, and optimal health payment more generally, would require reworking. This is currently an open area for research.

A limitation of the model in Program II is the assumption that $NB^0$ is fixed. In a monopolistically competitive model, if an insurer reduces price, it reduces price for all physicians in the market, not just the physician described in Program II. The value of the patient’s alternative, represented by $NB^0$, must therefore change as well with a change in the administered price. The simplest complete model of imperfect competition is

---

23 Patients fail to keep about 20% of scheduled visits in some studies. (See Oppenheim et al. (1979) and Smith and Yawn (1994).) Giuffrida and Gravelle (1998) contains an economic model of patient compliance. Compliance there is regarded as desirable, though coming at a cost.

24 See Dranove (1988) who confronted a similar issue in a monopolistically competitive model. The effect of more competition on patient quantity is ambiguous in his analysis.
two physicians who compete for patients distributed according to some dimension (e.g., distance), as has been analyzed by Glazer and McGuire (1993) and Ma and McGuire (1998). In a generalization of Program II, Ma and McGuire (1998) show the case of two physicians, competing in a Hotelling-type model, when an insurer sets price market-wide, $dx/dp < 0$ for both physicians; in other words, both physicians increase quantity in response to a regulated price fall. The reason for the positive quantity response to a fall in the administered price is the same as in the simple single physician model of Program I: lowering a regulated price channels a physician’s market power to quantity.

Intuitively, the explanation is as follows: In the presence of an administered price, the physician exercises monopoly power by setting quantity. Quantity $x$ is above the level the patient would demand at the price he pays, so increasing $x$ represents exercise of more monopoly power. Competition with other physicians limits how much the physician wants to increase $x$, because an increase in $x$, with other physicians’ behavior constant, leads some patients to leave this physician’s practice. Now, reducing the price paid per unit of $x$ reduces the penalty the physician pays for losing a patient. The loss associated with an increase in $x$ is reduced, and the physician (and all physicians) are led to increase $x$ in response to a regulated price fall. This result is unlikely to be fully general to models of monopolistic competition, which can have complex patterns of substitution across sellers. Furthermore, at some point, lower price must induce exit from the industry, tending to decrease quantity, at least in aggregate. It does establish, however, that a negative $dx/dp$ is consistent with a complete information, monopolistic competition model.

A model akin to the one just discussed was used to study the effect of administered prices in Medicare. Mitchell and Cromwell (1982), and Zuckerman and Holahan (1991) assumed the demand represented demand of the group of Medicare beneficiaries, and the cost curve of the physician sloped upward. Quantity provided might be limited by demand (they assumed price-taking demand behavior), or, if the marginal cost curve cut the $p$ line from below before the quantity consumers wanted to buy, quantity would be supply-determined.

The main application of this model of either demand or supply-limited quantity was to consider the effect of changing levels of Medicare fees and the effect of allowing physicians to “balance bill” Medicare patients. Medicare set fee allowances that determined what it paid to doctors, and determined the beneficiary’s coinsurance payment. Initially, physicians did not need to limit themselves to this price. At the beginning of the Medicare program, physicians could charge any price they pleased, requiring the patient to pay the “balance bill” equal to the difference between Medicare’s allowed charge and the physician’s price. Medicare has steadily restricted physicians’ authority to balance bill (today it is constrained to about a 10% window).

The question arises: what effect does a restriction on balance billing have on physician markets? In the Mitchell and Cromwell (1982) and Zuckerman and Holahan (1991) analysis, the physician could price discriminate, charging patients with a high willingness to pay a balance bill, but no balance bill for patients with a low willingness to
pay. The conclusion about balance billing from this analysis was the following: since only inframarginal patients were balance billed, balance billing functioned simply as a transfer from patients to physicians. These papers concluded that balance billing could be eliminated with no effect on quantity supplied.

This is an uncomfortable conclusion. Is nothing lost if prices paid to doctors are reduced by eliminating balance billing? The obvious concern is about the “quality” of services, in addition to quantity. If physicians have a choice about the quality of their services, we can see that the supply-constrained equilibrium is not likely to hold up. In the supply-constrained case, patients demand more (are willing to accept more) services at the (regulated) price they face, but the physician stops providing them because the marginal cost has risen to the supply price. Now, let the physician choose quality. By “quality” we mean some aspect of services that increases the value of services to consumer, is costly to the physician, and is not reimbursed by the payer. In a supply-constrained case, the physician can reduce quality, reduce marginal cost, increase profits, and then supply more services in response to the lower marginal cost. The reduction in willingness to pay stemming from the quality fall is not a problem for the doctor since the demand constraint is not binding. This process of reducing quality can continue, in fact, until demand does bind. The supply-determined case in these models will not be an equilibrium if quality is variable. When quality is variable, then, balance billing does have efficiency effects. Glazer and McGuire (1993) show that if Medicare sets fees correctly (an important proviso), all patients, those that pay as well as those that do not pay a balance bill, are better off if balance billing is permitted.

In this section we have shown that physician quantity setting and an increase in quantity associated with a decrease in regulated prices both emerge from this simple model, features normally associated with special market power (inducement) and motives (target incomes) of physicians. Second, efficiency problems in the market for physicians’ services take the form of too much quantity and an inefficient level of “quality.” The quantity problem arises from two sources, from patient insurance and from the quantity-setting power of physicians with administered prices. The inefficient level of quality may result from quality being unreimbursed within a regulated health payment system. In the next two sections, we show how managed care contains additional instruments for dealing with quantity and quality setting in physician and patient interaction.

---

25 Price discrimination was between two groups. The physician set one level of balance bill applying to part of the patients, and accepted the Medicare fee, with coinsurance, for the others. Accepting the Medicare fee as full payment is referred to as “accepting assignment.”

26 In Glazer and McGuire (1993), physicians can discriminate on quality as well as quantity, between patients charged the balance bill and those who are not.

27 It seems likely, as administered prices are pushed low to control costs, that quality will be set too low. Experience from the Medicaid program feeds this suspicion.
3.3. Noncontractible "quality," supply-side cost sharing in managed care contracts, and competition for patients

We define "quality" as a noncontractible input into the production of health for the patient. Noncontractible means that it cannot be used as a basis for payment. The care or effort that a doctor puts into a decision or treatment matters to the patient, but is difficult to incorporate into a payment system. More concretely, one could simply think of the "time" doctors spend in conducting a procedure [Glazer and McGuire (1993)]. Some physicians are paid per unit of time (e.g., psychiatrists, and anesthesiologists in Medicare). Yet actual time spent is very difficult to verify, and payments for most activities are not based on an explicit report of time. As McCall (1996, p. 51), an MD, notes, "several time-consuming activities, vital to providing good medical care, pay doctors nothing or next to nothing [including] conducting careful medical interviews, educating patients, staying up-to-date with medical advances." In recommending how patients should judge their doctors, he says, "the amount of time a doctor spends interviewing you, examining you, and explaining things reflects how genuinely concerned that doctor is for your welfare" (p. 52). "Time" is one good candidate for an observable but noncontractible input into the patient's health. Others—diligence, care, attentiveness—synonyms in this circumstance for "effort"—can be thought of as well.

Here we will use the term "quantity" to designate those physician inputs which are contractible, and "quality" to denote those which are not. From the patient's point of view, both types of inputs, quantity and quality, matter for the benefits of health care. We retain the assumption that the patient has full information about the physician, even though quality is not contractible.\textsuperscript{28} If the level of quality ("effort") supplied by the physician is $q$, then the benefits to the patient of treatment by the physician can be rewritten $B(x, q)$, with derivatives $B_x > 0$, $B_{xx} < 0$, $B_q > 0$, $B_{qq} < 0$. Effort is costly to the physician so now $c(q)$, and $c_q > 0$, $c_{qq} > 0$.

In the last several years, managed care payers have become more creative in writing contracts with physicians, incorporating various incentives for physicians to be careful about quantity. These forms of contracts appeared earlier in the HMO and hospital sector, where they were called prospective payment or "supply-side cost sharing" [Ellis and McGuire (1986), Newhouse (1996)]. They also go by the name risk sharing, and take the particular form of capitation and "withholds." In a capitation contract, a physician or group of physicians is responsible for a defined set of the health care costs for a patient over a period, typically a year. Under a "withhold," a physician or group is paid according to negotiated fees, but a bonus or a withhold is paid at the end of the contract period if certain cost or other targets are met.

\textsuperscript{28} Gaynor and Gertler (1995) suppose that a physician choses the "effort" put into work, and this will affect the quantity demanded of the physician. Medical group practices, studied by Gaynor and Gertler, reward individual work differently, according to whether payment policies are based on averaging or based on individual productivity. Gaynor and Gertler find that physicians in group practices which pool income make less effort, and this can lead to large reductions (up to 50%) in the volume of work physicians do.
Recent surveys have documented the prevalence of the changing contractual relationships between physicians and managed care payers. Emions and Wozniak (1997) use data from the 1996 AMA Socioeconomic Monitoring System to monitor contracts. Risk sharing contracts are more common with primary care physicians than with specialists, based presumably on the rationale that the primary care doctor has more control over the aggregate use of patients. In 1996, 36 percent of all physicians had at least one capitation contract, and for these physicians capitation revenues averaged 25% of the total. For some specialties, such as pediatrics, capitation revenues were 30 percent of practice totals. Withholds were about as common as capitation, with 36 percent of physicians having contracts with withholds. For these physicians, 19 percent of their revenues came from such contracts. Similar figures are reported by Remler et al. (1997) from a national survey, and from Hellinger (1996) for a slightly earlier period.

Capitation, withholds, and bonuses all fit within a framework of supply-side cost sharing. Supply-side cost sharing is present if when a service is provided (a cost incurred), some of this cost is borne by the provider because of the payment contract. Supply-side cost sharing can be low-powered as in a withhold system with a mild penalty for exceeding a cost target, or high-powered, as in a capitation system. In a fully-capitated system, the provider bears all of costs at the margin. We can write the general form of a per-patient contract with supply-side cost sharing as:

\[ R + p_s x, \quad \text{with } R > 0, \quad c > p_s \geq 0. \]

\( R \) is the portion of the payment made independent of the services provided – the capitation amount, the partial capitation amount, the bonus. \( p_s \) is the payment per unit of service. The contract features supply-side cost sharing if \( c > p_s \). In a capitation contract, \( p_s = 0 \); in low-powered contracts, \( p_s \) is closer to \( c \). A payer can calibrate incentives to a provider by alternating the composition of a contract between payment by \( R \) and payment by \( p_s \).\(^{29}\)

To appreciate the incentive properties of these contracts we must enrich the model we have been using so far to include competition for patients. Up to now, the physician faced the potential loss of a patient if the patient were given a package of price and quantity that failed to satisfy a net benefit constraint. More generally, a physician could be considered to have a probability of keeping a patient, with the probability increasing as the physician gives the patient more net benefit. Another interpretation of such a formulation is that the physician might attract more patients of a certain type if the net

\(^{29}\) One form of withhold contract would be if the physician receives a bonus \( B \) if costs do not exceed a target \( T \). For each dollar costs are above the target, the physician loses a portion \( r \) of the shortfall. This contract can be written as \( p_s x + B - r(p_s x - T) \), which fits within the \( R + p_s x \) form with supply-side cost sharing. The \( R + p_s x \) contract is linear, however, and cannot exactly capture nonlinear features of physician contracts, for example, if the physician "bonus" is constrained to be positive.
benefits she provides in her practice were higher. With either interpretation, we can express the number of patients the physician serves as a positive function of net benefit offered: \( n(NB) \), with \( n' > 0 \). Since \( p_s \) may be set below cost, we can no longer write demand price as a share of supply price. Instead, we will let the payer set demand price at \( p_d \).

We now have a richer profit maximization problem for the physician, that can be summarized in Program III.

**Program III:**

\[
\pi = n(NB)[R + (p_s - c(e))x], \quad \text{where } NB = B(x, e) - p_dx. \tag{3.9}
\]

Profit is a product of the number of patients and the profit per patient. The physician’s contract may have a prospective component per patient \( (R) \) and a fee \( (p_s) \) for each unit of service. The number of patients depends on the net benefit the physician supplies. Net benefit depends on both the quantity the physician supplies and the effort (quality) she puts in. Quantity is contractible, effort is not. The patient’s insurance is represented by \( p_d \), the price the patient pays for each unit of service. The physician chooses \( x \) and \( e \) to maximize profit.

The first-order conditions (3.10) and (3.11) describe the physician’s maximization:

\[
\begin{align*}
\pi_x: \quad n'(B_x - p_d)[R + (p_s - c)x] + n(p_s - c) &= 0, \tag{3.10} \\
\pi_e: \quad n'B_x[R + (p_s - c)x] - nc_ex &= 0. \tag{3.11}
\end{align*}
\]

These can be rewritten as

\[
\begin{align*}
\frac{B_x - p_d}{NB/x} \left[ \frac{R/x + p_s - c}{p_s - c} \right] &= -\frac{1}{\varepsilon_{n,NB}} \quad \text{where } \varepsilon_{n,NB} = n' \cdot \frac{NB}{n}, \tag{3.10'}
\\
\frac{R/x + p_s - c}{c} &= \frac{\varepsilon_{c,e}}{\varepsilon_{n,e}} \quad \text{where } \varepsilon_{c,e} = \frac{\partial B_x}{\partial e}, \quad \varepsilon_{n,e} = n' \cdot \frac{\partial NB}{\partial e} \cdot \frac{e}{n}. \tag{3.11'}
\end{align*}
\]

---

30 Formally, potential patients of this physician could have alternative net benefits distributed according to \( NB + d \), where \( d \) ("distance") takes some distribution. A patient goes to this physician if the net benefit he receives is greater than \( NB + d \).

31 We ignore any effect of income on premiums since we assume the benefit of health care is independent of income. If we were to use this model for a formal analysis of optimal insurance and payment, risk and the insurance contract would need to be considered.

32 The model here is closely related to Ma and McGuire (1997). In that paper, the physician chooses quality or effort and the patient chooses quantity of treatment. Price paid by the patient and paid to the physician are set by a payer. This model is also related to one studied by Feldman and Sloan (1988) where a monopoly physician sets quantity and quality to patients in the presence of price controls. Patients are price takers without insurance. The firm in Feldman and Sloan (1988) is a monopolist, and so faces no competition for patients.
We can use (3.10') and (3.11') to relate the policy instruments of the payer to the efficiency problems in treatment determination. There are two efficiency targets, one for \( x \) and one for \( e \). Quantity \( x \) tends to be overused because of moral hazard and physician market power. Quality or effort \( e \) may not be set efficiently in general because it is noncontractible. The payer, equipped with a payment system that includes supply-side cost sharing, has two instruments: the overall level of payment, and the prospectiveness of the payment, corresponding roughly to \( R \) and \( p_s \). As papers by Ma (1994) and Roberston (1994) first made clear, prospective payment can induce a provider to undertake noncontractible effort if this effort leads to more business. Increasing the profitability per patient (say by increasing \( R \)), leads to more of the noncontractible quality.

In general, (3.10') and (3.11') would need to be solved simultaneously to find the physician's profit-maximizing quantity and quality decision as a function of the payment system parameters. We can see from (3.10') and (3.11'), however, how the payment system can help solve the efficiency problems, and therefore why a managed care plan would be interested in writing a risk-sharing contract.

The first-order condition, (3.10'), describes the physician choice of quantity. \( B_x - p_d \) is the marginal net benefit a patient receives from more \( x \) and \( NB/x \) is the average net benefit the patient receives. Up to this point, the marginal net benefit has been negative: the physician has pushed quantity beyond the point where \( B_x = p_d \). This benefit elasticity is multiplied by a payment-system term and is equated to the negative inverse of the elasticity of the number of patients with respect to net benefit. The payment system term is the ratio of the average net revenue per unit of \( x \), \( R/x + p_s - c \), to the marginal net revenue, \( p_s - c \).

Consider first the case when there is no prospective payment and \( R = 0 \). The payment system term becomes one and the benefit elasticity is equated to the negative inverse of a market "demand response," the change in the number of patients with respect to a change in the net benefit provided. Consider the effect of an increase in the demand response elasticity \( \varepsilon_{n,NB} \). To bring the left-hand side of (3.10') into equality with a smaller (in absolute value) negative number, \( x \) must fall to bring \( B_x \) and \( p_d \) closer together. (The marginal changes faster than the average.) In words, (3.10') shows that the physician is restrained in pushing \( x \) too far because of the prospect of losing business. A similar idea, that physician quantity setting is restrained by market demand, was proposed by Dranove (1988) in a model of "physician-induced demand." There, patients were not sure of what they needed, but became increasingly suspicious of their physician's recommendations as the doctor's style of practice became more and more aggressive. As we show here, the same results hold in a model of complete information, recognizing that physicians have some quantity-setting ability. Market demand response may thus modify the physician's tendency to push visits beyond the point the patient would demand given his insurance. In a perfectly competitive model, this demand response is very high, and as (3.10') indicates, the discrepancy between marginal benefit and price paid by the patient disappears altogether. In general, with monopolistic competition, and without supply-side cost sharing, there will still be some quantity setting beyond the point a price-taking patient would prefer.
Now reconsider the payment term in (3.10'), allowing \( R > 0 \). With supply-side cost sharing, \( p_x < c \), and the payment term is negative. This has the very important effect of reversing the sign of \( B_x - P_d \). Thus, with supply-side cost sharing, the physician no longer has an interest in pushing quantity beyond the point the patient may demand, and indeed, will tend to limit the quantity to less than what the patient would demand (since the sign of the marginal benefit/price difference is reversed). By making the payment system more prospective, increasing the weight on \( R \), and decreasing it on \( p_x \), the payer can give the doctor incentives to cut back on quantity, perhaps even hitting the first best, where marginal benefit equals cost. The promise of supply-side cost sharing is that it can compensate in this way for the insured patient's moral hazard, and lead to the efficient quantity of health care, where \( B_x = c \). A payer could increase the degree of supply-side cost sharing, decreasing \( p_x \), while increasing \( R \) to maintain the overall average profitability of services. In that way, the quantity of \( x \) could be reduced towards the efficient level.

The payment system also affects the physician's choice of effort or quality, reflecting the physician's tradeoff between higher cost with higher quality, but more patients with a higher willingness to pay. Equation (3.11') shows that in profit maximization, the physician equates the percentage markup of average fee over cost to the ratio of two elasticities: the cost elasticity of effort over the demand response elasticity of effort. For inducing effort, it is only the average profitability that matters. This makes sense since effort is not explicitly reimbursed. If profitability goes up, attracting new business has more value, and effort rises. An increase in effort will lead to a rise in the right-hand side of (3.11'), bringing it into equality with the left-hand side increased by the rise in profits. The basic reason is that while both \( \varepsilon_{c,e} \) and \( \varepsilon_{n,e} \) are positive, marginal costs are increasing in \( e \), but marginal benefits are falling. An equation like (3.11') has been the basis of studies of optimal provider reimbursement by Ma (1994), Rogerson (1994) and others. Since more quality leads to more customers, paying more for each customer can induce higher quality services.

Many empirical studies have confirmed the effect of form of payment on doctor behavior. In a study in a position to distinguish physicians' desired supply from actual use, Hickson et al. (1987) found that pediatricians aggressively scheduled visits when they were paid generously by fee-for-service, in comparison to both standards of care promulgated by the American Academy of Pediatrics, and in comparison what their colleagues with comparable patients were doing when they were paid by salary: 4.9 visits per year vs. 3.8 visits per year. In both cases patients kept only some of the visits:

\[ \text{33 Ellis and McGuire (1986, 1993), Newhouse (1996), Ma and McGuire (1997).} \]

\[ \text{34 The literature on quality and effort has generally regarded this to be a patient-specific variable. Papers study the problem in a single-payer, uniform patient context. If some quality dimensions are practice-wide rather than patient-specific, the analysis of the determinants of quality would need to be broadened. The classic analysis of quality determination in a firm with a single set of customers (but without a regulated price) is Spence (1975).} \]
3.6 vs. 2.9, but the percentage kept was about 75% in both cases, suggesting that physicians could influence patient utilization. Jennison and Ellis (1987) found the same set of physicians provided more visits when paid by fees than under a capitation contract. Compared to doctors in HMOs, who do not make profits by ordering tests, Epstein et al. (1986) found that doctors in fee-for-service practice ordered 50 percent more EKGs and 40 percent more chest X-rays, the tests they perceived to be highly profitable. Rates of low-profitability tests were not elevated. At a walk-in care clinic, once salaried doctors were afforded bonuses to increase volume, lab tests went up 23 percent and X-rays, 16 percent [Hemenway et al. (1990)]. Stearns et al. (1992) found large changes in utilization in response to a shift from fee-for-service to capitated payments to a group of physicians. In a large study with extensive controls for patient characteristics, Greenfield et al. (1992) found that patients paid by fee-for-service in a large group practice were 27% more likely to be hospitalized than patients of the same group paid by capitation.35

The analysis in this section shows the power of payment system design for dealing with the two basic efficiency problems in physician treatment determination. As Ma and McGuire (1997) point out, however, even when the number of instruments equals the number of targets, the payment system may not be able to achieve both of these efficiency targets. For one thing, truthful reporting in payment systems constrains the choice of the payment system. In practical terms, $p_t$ can only be decreased to 0. If it were to go below 0, both the physician (penalized for each unit consumed) and the patient (facing a positive copayment $p_d$) would have an interest in misreporting quantity to the payer.36 Although the payer has two instruments, the level of payment $R$ and the degree of prospectiveness, $p_x$, the second instrument is limited. The payer, because of constraints on reporting, may not be able to attain the control on moral hazard desired by use of supply-side cost sharing.

Beyond this, in reality there are more than two targets. Physicians may be risk averse, for example, and unwilling to bear the degree of supply-side incentives the payer would otherwise want to impose. Furthermore, quantity of one contractible element, $x$, may not be the only contractible input into treatment. Multiple noncontractible inputs may exist. Payment systems are crude in the sense that the reimbursement and insurance contract creates incentives that are uniform across a range of services. Fortunately, other powerful instruments are also available to managed care plans.

3.4. Network incentives in managed care

Managed care plans have at least two other sets of instruments in addition to supply-side cost sharing to contend with moral hazard. Utilization review allows the third party

35 See Hellinger (1996) for a review of some of these studies.
36 See Brundin and Ma (1998) for more extended discussion of the implications of relying on reports from patients and providers for design of payment systems.
payer to interject judgment about what services should be provided to patients. The most obvious mechanism used by managed care in this respect is denial of care. Denial countermands the decisions of patients and doctors under the sway of moral hazard. Outright denial of care, however, appears not to be very common. Rermel et al. (1997) found that while managed care plans initially denied 3.4 percent of physicians' requests to hospitalize patients, 2/3 of these denials were reversed on appeal, leading to an ultimate denial rate of only one percent.

Managed care plans also assemble a "network" of providers. Patients in a managed care plan typically must obtain covered care from a provider in a network. When an out-of-network provider is used, coverage may be significantly less, or absent. Managed care plans do not accept all doctors, and occasionally drop doctors from their networks. Emons and Wozniak (1997), using AMA data, report that 13 percent of all physicians applied for and were denied at least one managed care contract in 1996, and 6 percent were dropped involuntarily from a network during that year. Limiting patients' choice of doctors decreases a consumer's valuation of a health plan, all else equal. What does a plan gain in exchange for this restriction on patient choice?

A managed care plan seeks lower prices from physicians who are granted network privileges. If restricting the number of participating physicians is going to lead to a better price for the managed care plan, it must be that competition is imperfect; otherwise, for any quantity target, the plan would minimize price by admitting all available suppliers. In pharmaceuticals (where there can be no doubt competition is imperfect), networks are called "formularies" and by restricting choice of drugs, managed care plans and others can enhance competition within a class of drugs and bargain for a better price with manufacturers [Scherer (2000)]. Dranove et al. (1993) labeled a similar phenomenon "payer-driven competition" in the case of hospital services. When a managed care plan can direct patients to lower-priced hospitals, it increases hospitals' price elasticity of demand, eliciting lower prices.

A health plan may use a network to pursue other objectives as well. In the same way as in price competition, if the managed care plan can see and evaluate quantity per patient, membership in a network can be used as an incentive to control moral hazard. At the level of an individual patient, a managed care plan may have a limited ability to judge the appropriateness of utilization. But at the level of a physician's practice, on which network decisions are based, patient severity will tend to average out (if even imperfectly) and more conclusions can be drawn. A managed care plan may not know if Mrs Smith needed a Caesarean section, but it may be able to say that Dr. Jones' practice, with a 50 percent rate, is not the one the plan wants in its network. Similar remarks apply to observable elements of "quality."

Networks can also be formed by physicians in order to market themselves to managed care plans and other buyers. Provider-formed networks can interfere with managed care plans' contracts, and may pose an antitrust threat in markets where the network has market power (Haas-Wilson and Gaynor (1998), Greenberg (1998)). "Any-Willing-Provider" laws at the state level inhibit managed care plans from establishing networks [Ohsfeldt et al. (1998)].
In terms of the model set out above, supply side cost sharing creates incentives by altering the payment system component of the physician's revenue function. A network alters the demand-response portion of the revenue function, the \( a(NB) \). With a network, patients are not free to go to any doctor according to how they view the net benefits they will receive. Patient flow is at least partially controlled by the plan. In this way, the plan can feed a doctor patients if the patients are getting the care the plan wants them to get. In a regime with moral hazard, physician referral rates can go up (not down) for compliant physicians as quantity is cut back and patient net benefit falls.

Network effects are only beginning to receive explicit attention in the literature. Within a model of monopolistic competition, Ma and McGuire's (1998) managed care plan penalizes a doctor by denying some patients who would otherwise seek out the doctor if the doctor deviates from the plan’s target utilization. They show that the plan may need only a small penalty to enforce the behavior it wants. They measure the network effect associated with a mental health managed care plan and find it to be associated with a large (roughly 40 percent) decrease in the quantity of visits per episode of care.

3.5. Efficient production of physicians' services

The literature on physician behavior has often examined whether the production of physician services, or the production of health services, takes place efficiently. The relation of physicians to one another (solo versus group practice) to other personnel, and to hospitals, has all been studied theoretically and empirically. Some of the motivation for this work stems from information and contracting issues.

Solo and group practice forms each have advantages and disadvantages, known in the partnership literature (Farrell and Scotchmer (1988)). Solo practice internalizes incentives, but group practice allows more flexible allocation of “lumpy” inputs. DeFelice and Bradford (1997) test for productivity of physicians in these two modes and find them to be about equal. Gaynor and Gertler (1995) regard the question of the optimum size of the group as involving a tradeoff between risk and incentives to elicit noncontractible effort.

The economic relation between physicians and hospitals has been questioned in terms of whether it would lead to an efficient use of the two main inputs into the production of hospital care. Traditionally, physicians and hospitals have functioned as independent economic units, with the seemingly odd arrangement that a physician could admit a patient to a hospital free of charge (to the physician) and order hospital staff to incur costs as the physician saw fit. The zero price hospitals charge for “privileges” at the hospital appears not to have originated as an equilibrium, but through custom. If one regards the hospital as “capital” and the physician as “labor,” the normal economic organization in a capitalist economy would be for the hospital to hire the physician, pay the physician a wage, while charging the patient a unified bill for hospital care. Pauly and Redisch (1973) suggest an opposite interpretation, that in health care, groups of physicians (the medical staff) act as a worker cooperative, hire capital, pay capital its opportunity cost.
only (hospitals are non-profit), and garner any surplus in the form of higher reimbursement for themselves. Physicians can exercise market power by restraining the number of physicians who are members of the "cooperative." This relationship may, however, not disturb incentives for production efficiency. If physicians have a claim on any "residuals," they can be presumed to have incentives to minimize cost of treatment to patients [Pauly and Redisch (1973), Pauly (1980)].

Within the conception of the labor-dominated firm, price setting and other contracting practices of third-party payers have altered the division of surplus between labor and capital. Physicians are combining with hospitals in the form of physician–hospital organizations (PHO's) or engaging in contracts, such as exclusive contracts, which, in effect, circumvent the custom of the zero price for privileges. A physician might be paid some amount, for example, in exchange for sending all her patients to a certain hospital. Under pressure from organized payers, physician–hospital relationships are changing rapidly, involving more vertical integration and closer contracting [Robinson (1997)]. The number of physician practices owned or managed by hospital-based systems increased 60% in one year between 1994 and 1995 to 11, 234 (Modern Healthcare, June 3, 1996.) As always, anticompetitive effects are possible in tying contracts when at least one of the parties has market power [Frech and Danger (1998)]. Formal analysis of the effect of tying contracts between hospitals and physicians has only begun to contend with an active, contract-writing, third-party payer. [See Ma (1997).]

The physician running a practice is an owner-manager, and as such must combine other inputs with her own time to produce physician services. Since physicians are scarce and highly paid, a certain amount of substitution of other personnel is required for efficient production. Reinhardt (1972) and Brown (1988) are concerned whether a physician hires the efficient number of partially substitutable personnel. In managed care plans, when physicians lose authority about hiring aides, the ratio of physician extenders to physicians goes up a great deal. This could be a move to a more efficient mode of production, or a move to produce a different form of service. Econometric studies of physician cost functions have been undertaken for the purposes of evaluating efficiency [Pope and Burge (1992)] and calibrating "cost-based" payments [Escarce and Pauly (1998)]. The difficulty of measuring all inputs and outputs precludes drawing a clear conclusion about efficiency from these studies.

37 Harris (1977), an economist and practicing physician, regarded the operation of a hospital as subject to two lines of authority: the medical staff that run patient care and the management which made long-run resource and marketing decisions.

38 The incentive to combine inputs efficiently brought about by this form of internalization is disturbed by differential insurance coverage for the inputs. For example, if hospital services are fully insured and physician services are not, hospital services will be overused in Pauly's (1980) framework.

39 Jerry Cromwell (1996) considers nurse anesthetists to be "nearly perfect substitutes for anesthesiologists" but the nurses are paid at half the rate of the doctors. Substitution has been thwarted because payers, particularly Medicare, pay for inputs rather than outputs. In an integrated delivery system hiring workers according to salaries, the substitution is possible.
3.6. Summary

In a setting of complete information, recognizing that physician services are nonretradable yields the first of the three ways that a physician can influence quantity of health care used by patients. Physicians can set quantity apart from what a price taking patient would choose, subject to a constraint on patient exit. When physicians face regulated prices, market power is channeled into quantity setting, and a decrease in regulated price may lead to an increase in quantity, without a need to appeal to inducement based on asymmetric information or target income motivation. Expanding the purview to a market with many patients, the physician’s quantity setting can affect the total demand at the practice level.

A complete information model can also be used to explicate the second mechanism a physician has for influencing quantity. Physician care contains observable but noncontractible elements, referred to in the literature as “effort” or “quality.” These inputs may complement or substitute for the contractible inputs. By setting the level of the noncontractible input, the physician can influence demand for the contractible ones. Though noncontractible, effort or quality is not outside of the power of health plans. Partial control over the number of new patients coming to a physician provides a mechanism a health plan can use to manipulate doctors’ decisions. Supply-side cost sharing can introduce a penalty for providing “too much” care. This can be balanced with a per patient payment that makes noncontractible quality worthwhile. Managed care networks allow plans to manipulate demand response and give the plan another instrument to influence doctor behavior.

The power of demand response depends, of course, on the patient’s being able to observe and evaluate elements of a physician’s practice. This ability is called into question in the next section.

4. Uncertainty about treatment effects and asymmetric information

Arrow (1963) titled his paper, “Uncertainty and the Welfare Economics of Medical Care” [italics added]. More recently, Wennberg (1985) has argued that uncertainty is the most important factor influencing physician behavior. Many writers on health care have deep misgivings about the application of simple economic models in health care, based largely on how these models ignore the uncertainty and informational asymmetries surrounding health care. In a recent paper, medical sociologist Donald Light (1997, p. 299) writes,

"Health care is often emergent as diagnosis and treatment unfold. Clinical decisions are contingent on what is found and how the patient reacts. Cases are highly variable, and the course of treatment is uncertain." Furthermore, he stresses that, "There is great information asymmetry, because the clinician knows so much more than anyone else" [his emphasis].

According to Wennberg, the sources of uncertainty include the following: first, there is classification of the patient in terms of disease condition, or initial health status; sec-
ond, there is uncertainty about the effects of treatment for a given condition, even in controlled conditions; and third, patient preferences may not be known (at least to the physicians). The presence of uncertainty is sometimes argued to lie behind the observed variations at the population level in the rates of treatment (Phelps, this Handbook).

In health economics, analysis has emphasized situations in which there is uncertainty, and in addition, where information about effects is not shared equally, that is, situations of asymmetric information. Before discussing asymmetric information between the physician and the patient, it is useful to consider first the implications of what Pauly (1978) referred to as "irreducible uncertainty:" the absence of information about the consequences of health care treatment that is shared equally by the doctor and the patient.

4.1. Irreducible uncertainty

In Section 3, the benefits a patient receives from health care are \( B(x) \). The negative second derivative of \( B(x) \), \( b'(x) < 0 \), can be interpreted as stemming from two factors. Patients may have a declining marginal valuation of health care because (1) as more health care is consumed, the marginal impact on health status is lower, or (2) as more health status is gained, the marginal utility of health status itself falls. To capture both of these effects, we could write \( B(x) = V(H(x)) \) where \( H(x) \) is the relationship between \( x \) and health status, and \( V(H) \) is the utility of health status function. Both \( V'' \) and \( H'' \) could be less than zero. Suppose \( V'' < 0 \), as seems likely. We can then consider there to be risk aversion with respect to health status.

A simple way of introducing uncertainty is to suppose that the patient (and the doctor) are uncertain about initial health status, and this uncertainty is additive in relation to the improvement in health rendered by consumption of \( x \). Then, the patient’s benefits from treatment must be regarded as expected benefits:

\[
E[B(x, u)] = E[V(H(x) + u)],
\]

where \( u \) is a random variable with mean zero and variance \( \sigma_u \). If the degree of absolute risk aversion is decreasing with income, it can be shown that the expected benefits from medical services \( x \) are increasing in the level of uncertainty as represented by the variance of \( u \).

Irreducible uncertainty about initial health status or the effects of treatment imposes risk on patients. A patient may want to offset this risk by consuming more health care. Judged in terms of its effect on health status, some demand may appear as not worthwhile, in the sense that the cost is high in relation to the expected increase in health status. However, from the point of view of ex ante utility, the utilization may be worthwhile. Patients may be rationally seeking to insulate themselves against health status risk.

\footnote{This assumption implies that the third derivative of \( V() \) is positive. Intuitively, as people have more income they are less risk averse. See Arrow (1971) or Pratt (1964).}
by consumption of medical care. Woodward et al. (1998) show how this can work in a decision-analysis framework concerning diagnostic tests for liver cancer. Expected utility is modeled explicitly. The authors conclude that "risk aversion can increase the perceived value of diagnostic procedures and thus raise optimal diagnostic expenditures" (p. 149). (See also the discussion in Grossman's (2000) chapter of this Handbook.) We now proceed to consider cases in which there is uncertainty, and in which information is distributed asymmetrically.

4.2. Unobservable physician actions

Some of what the patient does not know may be known by the doctor. Economic situations involving asymmetric information are referred to as agency problems, wherein the principal (the patient) is affected by an action taken by the agent (the doctor).41 In some cases, the asymmetry of information between doctors and patients, or indeed between doctors and payers, might be so severe that there is no way for any outside party to know what a doctor did or knew. Some patients (e.g., the very young, the very old) may not be able to report, and some aspects of treatment (e.g., pain relief) may leave no trace. Economic incentives will be little help in eliciting effort in such activities.

There will be many other cases, however, where there is asymmetric information, but mechanisms by which evidence can be collected regarding the doctor's behavior. Better health care can be expected in general to lead to better outcomes, at least probabilistically. If good outcomes can be paid upon, this can motivate doctors, a principle on which some of the burgeoning literature on "performance contracting" is based.42 Even if outcomes cannot be paid upon, that is, are "noncontractible," outcomes may be observable to clients. It may be infeasible to pay doctors on whether they are able to cure back pain because it is too costly to validate a patient's report. Nonetheless, the patient knows if his back still hurts. If the doctor is rewarded for doing a better job, because the patient is more likely to return or to recommend this doctor to friends, the doctor is encouraged to take unobserved actions to improve quality. Note that this mechanism is similar to that studied above in the complete information case. Instead of observing effort directly, here the patient instead observes an imperfect indicator of effort, outcome.

The patient may see the outcome of the doctor's action, but because outcome is also affected by other unobservable factors (e.g., the uncertainty just discussed above), the patient does not know for sure whether the doctor's action was appropriate. In his general review of agency theory, Arrow (1986) returned to the case of doctor and patient:

---

41 There is a presumption in the economics literature that more information to the patient is always good. This may not be correct. Physicians may fail to disclose some information to patients for valid paternalistic reasons. If a child might have cerebral palsy, it could be argued, why tell the parents until the situation clarifies? On the other hand, patients complain that doctors tell them too little, too late. See Sloan et al. (1993, pp. 56–67).

42 See Lu (1999) and Shen (unpublished) for papers that include models of how providers respond to performance incentives, and measures of the consequences, intended and unintended. See Zweifel and Breyer (1997) for an analysis in which outcomes are contractible.
The physician–patient relation is a notorious case... The very basis of the relation is the superior knowledge of the physician. Hence, the patient cannot check to see if the actions of the physician are as diligent as they could be (p. 1184).

Physician diligence, referred to in Section 3 as "effort," could take many forms. If physicians know something that could benefit the patient, e.g., initial health status, then revealing this accurately and completely to the patient would help the patient decide how much health care to seek. If physicians must find something out, diligence could consist of taking actions necessary to identify the patient's condition.43

Several authors have discussed the applicability of agency theory to physicians and patients [Dranove and White (1987), Mooney and Ryan (1993), Gaynor (1994)], and some papers have developed particular agency models [Blomqvist (1991), Lundback (1998), Zweifel and Breyer (1997)]. Arrow (1963, pp. 964–965) anticipated one important result in the principal-agent literature when he recognized that a potential solution to the informational asymmetry is to transfer all risk to the physician. He called this an "ideal insurance" plan that would protect the patient against the portion of health status risk that could be ameliorated by the doctor's actions. "Under ideal insurance, the patient would actually have no concern with the informational inequality between himself and the physician, since he would be paying by results anyway, and his utility position would in fact be thoroughly guaranteed."44 In their discussion, Dranove and White (1987) note that transferring risk to the agent is infeasible because health status is in general noncontractible. In common language, patients could exploit the situation by claiming "it still hurts." Mooney and Ryan (1993) later make the same point by stressing how health markets deviate from a standard principal-agent model because of the inability to contract on outcome.45

Information asymmetry between buyer and seller is of course not unique to health care. The broader literature contains papers addressing the issue of "experts" consumers must rely on for advice and subsequent services. In a fashion similar to health care, experts in car repair or legal services may have an incentive to call the consumer's problem serious when it is really not, in order to increase demand for their services.46 One theme in this literature is how competition among experts (who may or may not vary in

43 These two aspects of diligence are referred to in the principal-agent literature as "hidden information" and "hidden action" problems. See Arrow (1986).
44 A qualification on this is that there may be some irreducible uncertainty that cannot be ameliorated by physician actions, as discussed above.
45 In the standard principal-agent model where outcome is observable and contractible, risk can be transferred to the agent. However, if the agent is risk averse, this would require the principal to compensate the agent for accepting the risk. Typically, then, the optimal contract is a contract in which risk is shared between both parties, the principal balancing the costs of paying a risk premium against the inefficiency created by the lower-powered incentives to the agent. See Laffont and Tirole (1993). Zweifel and Breyer (1997) apply such a model to the patient-provider case. Newhouse (1996) discusses the applicability of this type of model to hospital contracting. Paying on "outcomes" may encourage providers to select patients who are healthy and avoid the sick. See Shen (1999) for a model of this behavior and an empirical study.
46 Gaynor’s (1994) review contains an extended discussion of this literature.
skills) limits the potential exploitation of consumers [Wolinsky (1993), Emo (1997)].
Some price competition is part of these models. Market processes can be expected to
work differently in health care, however, where experts cannot set prices because of the
prevalence of fixed demand and supply prices.

We can add asymmetric information into the model developed so far. Assume there
is an input controlled by the physician which influences health outcomes, but that is
not observed by the patient. A literal "input" is one interpretation of this new element.
A physician could be diligent or not in terms of putting in effort of some form not
observed by the patient. Another interpretation, however, is in terms of information.
A physician might reveal or not reveal what she knows, or only partly reveal information.
More accurate information presumably benefits the patient, by for example, revealing
to the patient ahead of time the benefits and risks of some procedure that the patient
might use. To ensure that the unobserved action cannot be inferred, there must also be
another input, also unobservable; otherwise, the patient could draw accurate conclu-
sions about effort from observing outcome. Thus, write patient benefits as \( B(x, e, u) \),
where, as before, \( x \) represents contractible inputs. In contrast to Section 3, we will now
regard \( e \), effort, as unobservable, implying, of course, that it is also noncontractible.
Furthermore, \( u \) is a random variable, also unobserved, that influences outcomes to the
patient. The patient is assumed to know the functional relationship among the variables
included in \( B(\cdot) \), to observe \( x \), and to see the outcome, the realized or observed value of
health care benefits, which we can call \( B_o \).

A simple illustration of such a model is when effort is \((0, 1)\): the physician consults
her expert colleagues or she does not. In this case, the Bayesian patient can infer a
likelihood that physician took effort \( e = 1 \), based on his observation of \( B_o \) and \( x \). This
could be called \( L_1(B_o, x) \). The likelihood the physician took effort \( e = 0 \), is simply then
\( L_0(B_o, x) = 1 - L_1(B_o, x) \).\(^{47}\) In a modification of Program III in Section 3.3 above,
we can write the physician's profit as:

\[
\pi = n(L_1[R + (p_s - c)x],
\]

where \( n \) is the number of patients the doctor sees, now a function of the doctor's like-
lihood of providing high effort. \( R \) is the per patient payment, \( p_s \) the payment per unit
of the contractible input \( x \), and \( c \) is cost per unit. If the patient can presume that \( e = 1 \)
characterizes a physician in subsequent encounters, the patient's observation of \( B_o \) can
be informative. A physician may regularly consult colleagues, or regularly reveal infor-
mentation to patients about their health status. In these cases, patients learn about what
they might expect in future encounters. Demand response to this information is impor-
tant, because it is the mechanism by which the physician is rewarded for high effort.
The combination of prospective and per unit payment is important because a payer is
interested in affecting the contractible as well as the noncontractible input.

\(^{47}\) It is possible to generalize this in the direction of repeated observations in which \( L_o \) is a function of a first
and subsequent encounters, and in the direction of a continuous effort decision by the doctor.
Pauly (1980), Dranove (1988), and Rochaix (1989) have analyzed models similar to (4.2) when all payment is in the \( p_x \) component (which then must be greater than or equal to \( c \)), and there is no prospective payment. In these papers, there is a demand response influencing the physician’s choice of \( e \). Pauly’s fee-paid doctors can take an action to increase the volume of contractible services used by patients, but the doctor must tradeoff the benefits of this action against the cost of discouraging patients from coming to a doctor they regard as “inaccurate.” Dranove’s (1988) physician makes a treatment recommendation. The patient is more likely to reject the recommendation if the physician has a “bad” reputation as being an “overprescriber.” In Rochaix (1989), the threat of patient exit constrains doctors to be more conservative than they otherwise would be. These papers were oriented to the issue of “demand inducement,” conceived as an activity a doctor could undertake that would not involve resource costs, but would increase quantity demanded. Said in terms of our model, the physician might observe something about the patient, and then decide whether to reveal this to the patient. Demand inducement as portrayed by Pauly and Dranove could be thought of as the doctor telling the patient they are the “type” that would benefit a lot from health care, even if they are not. This would induce the patient to be willing to consume more of the contractible input forming the basis of physician payment. This constitutes the third mechanism for quantity setting contained in Table 1. The issue of demand inducement will get more attention in the next section. We simply note here that this model covers this case, but is more general. The action could also require the physician to incur some cost. The demand-response mechanism would still reward the action and be a way to elicit the unobserved action.\(^{48}\)

Generalizing the payment system to be \( R + p_x x \) adds an important perspective. As Newhouse (1996) notes, if payment is pure capitation, a provider may have an incentive to use her informational advantage to discourage treatment. And as in Section 3, the physician’s decision about effort will be governed by both the level of the prospective payment \( R \), and the payment per unit of the contractible input, \( p_x \).

As just discussed, the physician’s choice of unobserved effort is put forward as a characteristic of the physician that the patient (and payer/regulator) could expect to hold true in repeated encounters. In other cases, a physician might make a choice in treatment of a particular patient that might not be a reliable indicator of a pattern of behavior. If the level of effort cannot be anticipated, demand response in terms of number of patients might not materialize. All may not, however, be lost in terms of eliciting effort. Another form of demand response is the patient’s behavior in the course of treatment. Treatment is actually a sequence of actions by the patient and the doctor, as Light emphasized above. Patients may learn something over the course of treatment about their outcome, and have a chance to react to a doctor’s unobserved effort. A simple example would be if the sequence were as follows: the patient seeks

\(^{48}\) Chalkley and Malcolmson (1996) analyze the difficulties in writing provider contracts when the demand-response mechanism is not operative.
treatment; the doctor takes more or less care; the patient responds favorably or not to treatment. Then, if the patient’s response is poor, the patient must seek more treatment. The “seeking more treatment” is a contractible action, that depends on the non-contractible outcome, that itself depends on the unobservable physician action. “Internalizing” the costs of poor outcomes is a rationale behind prospective payment and capitation, but there is a more general problem lurking here in which the payment to the provider depends on subsequent actions taken by the patient. The power of any such mechanism to redress inefficiencies in physician decisions is yet to be explored.

4.3. Unobservable physician characteristics

We can interpret $e$ in $B(x, e, u)$ in a related way, as an unalterable characteristic of a physician. It could be “quality,” but quality understood as a fixed characteristic, such as acumen in diagnosis. The economic issues involved in an unalterable characteristic are somewhat different than a behavior to be induced. If patients value this characteristic differently, efficiency requires matching patients to the right doctor. If the mix of quality can be altered in the long run by changing the composition of doctors, efficiency requires quality to be rewarded.

The physician is an “experience good,” as has been noted by Gaynor (1994) among others. A patient literally has to try a doctor, and then make an inference about the doctor’s quality, including any issues about a match with the patient’s preferences. Because learning is imperfect and slow, the market’s reward for higher quality is likely to be inadequate. Hoerger and Howard (1995) review the literature on consumer search for physicians. In their own study of search behavior, they found that only about 1/4 of women seriously considered an alternative to their prenatal care provider. Satterthwaite (1979), Pauly and Satterthwaite (1981), and more recently Dracone and Satterthwaite (this Handbook) emphasize that asymmetric information about physician type is a basis for monopolistic competition. Pauly and Satterthwaite (1981) observe that if there are more doctors in a market, a patient, through personal contact or by information provided by friends and relatives, will tend to know less about any given doctor because the information is spread more thinly. More doctors may therefore not increase competitiveness of market, but by diluting the quality of information a patient has in some sense, and increase patient allegiance to a known doctor.49

A simple learning model has implications about markets and doctor quality [McGuire (1983)]. First, since patients must infer quality by good outcomes, and it is always possible that a “bad” doctor is lucky, the market’s valuation of the quality of a doctor will suffer from regression towards the mean. Poor quality doctors will be valued too

49 Wong (1996) found little support for the information-demand elasticity connection.
highly, and high quality doctors will be valued too low. Second, given uniform prices, no patient believes his doctor is less than average quality.50

4.4. Summary

Simple (or irreducible) uncertainty and asymmetric information both have implications for decisionmaking by doctors and patients. A model incorporating both can be built on the structure introduced in the previous section. There are many areas for potential research, along the lines of making more explicit the nature and consequences of asymmetric information. Patient observation and learning and their implications for physician behavior and market equilibrium have barely been explored in the literature so far. Viewing treatment as a sequence of patient and physician actions, for example, seems a potentially rewarding area.

In terms of physician influence over quantity, this section has added “persuasion” to the two already identified in Section 3. As we have said here, in the presence of asymmetric information, the physician can take an unobservable action (with or without incurring resource costs), that will influence patient valuation of care. This activity, denoted “demand inducement” in most of the literature on physician behavior, is the subject of many empirical papers in health economics.

5. Physician-induced demand

Evans (1974) opened his influential paper on “supplier-induced demand” with the statement (ironic in hindsight), “Everyone knows that physicians exert a strong influence over the quantity and pattern of medical care demanded in a developed economy.” The meaning and measurement of supplier or physician-induced demand (PID) has in fact been one of the most contentious topics in the economics of health care. The title of Phelps’ (1986) paper, “Induced Demand: Can We Ever Know Its Extent?” conveying more dismay than conviction, better captures the tone of the empirical literature.

The hypothesis of physician-induced demand, that physicians alter the patient’s preferences in their own interest, threatens the economist’s basic market paradigm, and undermines the normative implications that underlie economic recommendations about market policy. Positive and normative economic analysis proceed readily when consumers have stable preferences. Then, combinations of price and quantity can be interpreted in terms of the interests of the consumer and in terms of market efficiency. Throughout the 1970s, economic policy in health care focused on the demand side and

50 A patient begins with the prior (knowledge? impression?) that the physician is average. Then, if the first outcome is good, the posterior estimate of the physician is improved, and the patient stays. If the first outcome is bad, the posterior drops. In the absence of switching costs, the patient then leaves the physician since the average doctor is always available. A survey of the (rational) patients would reveal that all thought their doctor was average or above in terms of quality.
relied on the basic demand-management paradigm. National health insurance was the main issue, and the operational question for economic policy was the degree of price subsidy that should be contained in national policy. The positive model of utilization determination was demand, and the normative theory of demand was used to derive implications for health insurance coverage. [See Newhouse and the Health Insurance Experiment Group (1993) for a report on this line of work.] Rice (1998) contains an extended discussion of the issues. But if physicians could induce demand, the policy superstructure build over the theory of consumer demand was at risk. As Reinhardt (1989, p. 339) argued in this context, "The issue of physician-induced demand obviously goes straight to the heart of probably the major controversy in contemporary health policy, namely the question whether adequate control over resource allocation to and within healthcare is best achieved through the demand side or through regulatory controls on the supply side." Dyckman (1978, ii), referring to physicians' ability to induce demand, contended that "normal market forces are weak or nonexistent."

The PID hypothesis has direct policy implications, many of which run counter to the normative implications of conventional models of demand and supply. For example, training fewer surgeons will reduce rates of unnecessary surgery in the US [Schroeder (1992)]. Prohibiting physicians from owning testing equipment will reduce excessive rates of testing [Hillman et al. (1992)]. To make physician fee policy within a fixed budget, it must be anticipated that fee reductions induce quantity increases, so fees must be reduced even more than otherwise [PPRC (1991)] to maintain a balanced budget. The Health Care Financing Administration assumed in Medicare's fee reforms that half of any payment reduction would be offset by a volume increase in Medicare [PPRC (1991)]. The demand-inducement assumption cost all physicians 6.5 percent of fees in 1992 (an effect compounded over the years). Identifying the empirical importance of the PID effect, if any, in various policy contexts is obviously an important objective of health economics.

Fortunately, at a conceptual level at least, there is agreement about what constitutes PID. We adopt the following definition taken from the careful writing on the subject over the past two decades:

Physician-induced demand (PID) exists when the physician influences a patient's demand for care against the physician's interpretation of the best interest of the patient.

It is important to keep two distinctions in mind when applying this definition. The first is the distinction between useful agency and inducement. Fuchs (1978, p. 36) early on defined demand inducement as above, in relation to the consumer's optimal consumption point, leaving open scope for influence in the interest of the patient distinct from

\[51\] The volume offset was assumed to be asymmetric: only price reductions, not price rises would be offset.

inducement. Thus, if a physician influenced a patient to move towards the consumer's optimal point this would not be inducement, only useful agency.\textsuperscript{53} Eisenberg (1986, p. 57) defines inducement as "prescription of services that a well-informed consumer would not want to use." Pauly (1980) makes use of the same concept in his definition of a "perfect agent." The physician assists the patient to demand "exactly those quantities of care of various types that the patient would have chosen if he had the same information and knowledge the physician has." (1980, p. 5). A similar idea of the "perfect agent" is contained in Culyer (1989) and Williams (1998). Even Frech's negative characterization of demand inducement is still consistent with our definition: "As more physicians crowd into a market, they give more fraudulent advice and raise the demand for health care" [Frech (1996, p. 84)]. The upshot of these definitions is that showing influence is not enough to establish "inducement." As Newhouse has said, the question of PID is not a matter of introspection, of "thinking back on one's last visit." (1978, p. 60). (After all, who has not been influenced by physician recommendations?) PID requires, in essence, a finding of "undue" influence. The empirical methods for identifying "undue influence" will be discussed below, but essentially stem from the idea that if a change in the physician's return from inducement (e.g., fees go up) stimulates a change in influence (more surgery recommended), we have evidence for PID.

The second distinction is between utilization and demand, a distinction that has become more salient with the growth in supply-side cost sharing and managed care, rationing devices that do not rely on controlling costs by decreasing quantity demanded. A physician can influence utilization without influencing demand. Here is an example. Patients treated in an HMO may receive less treatment. This could be interpreted as a PID-type mechanism -- a decrease in demand caused by the physician. At the price they were paying and with a fully informed demand patients would have demanded the extra treatment but the physician influenced them otherwise and lowered their demand. Alternatively, it could be evidence of rationing -- the HMO physicians simply ration the care, not allowing patients to have all they want. The HMO patients have the same demand as the non-HMO patients, it is simply unsatisfied. An empirical finding of an HMO effect or an effect of prospective payment or managed care, as is now common in the literature, is not sufficient to establish PID in the Fuchs/Pauly sense. Utilization has been affected, but it is not clear that demand -- the function relating price to desired quantity -- has shifted. This quantity setting is of course the first of the three ways we have identified for a physician to influence utilization, not the third way that is of immediate concern here.

The literature outside of economics is without soul searching about whether physicians influence demand. It is nearly universally considered obvious that of course they do. The concern in this literature is usually with identifying the factors, such as socioeconomic status of patients, that lead physicians to direct patients to different courses

\textsuperscript{53} This mechanism -- supplying information to the patient that changes demand -- has been identified and discussed above as the third mechanism by which a physician can influence quantity used.
of therapy. A recent study by the medical sociologist John McKinlay and his colleagues [McKinlay, Potter and Feldman (1996)] described creation of a series of videotapes to study among other things the influence of insurance status, payment method to the doctor, and socioeconomic status of the patient on diagnosis and recommended treatment. The videotape method allowed the investigators to present exactly the same medical information to physicians by different patients and to physicians in two payment conditions. One of the medical conditions studied was chest pain. Insured patients were more likely to be given a cardiac diagnosis, with greater subsequent resource use than the gastrointestinal or psychogenic alternatives. Physicians practicing in an HMO setting were less likely to recommend a follow-up visit for chest pain. Diagnosis and recommended treatment were regarded as physician decisions in this study, and these were significantly influenced by nonmedical factors.54

Evidence such as this, while interesting for many purposes, does not make the distinctions between influence and undue influence, and between demand shifting and quantity-setting that underly the economic definition of PID. Physicians may have superior knowledge and can help the patient by conveying this information to the patients, or, perhaps, more directly by simply choosing on the patient’s behalf and foregoing the effort of education and persuasion. (Some patients may not want to make choices or pay for this effort if they believe the physician is acting on their behalf.) In the McKinlay et al. study, patients with insurance are more likely to be given the higher cost cardiac diagnosis. Physicians may be acting as proper agents of the patients, giving the higher cost cardiac diagnosis more frequently to insured patients because these patients face a lower out-of-pocket price for care and may be more willing to demand more aggressive therapy. Physicians anticipate that patients have a downward sloping demand curve and recommend more aggressive treatment for the insured. The second finding illustrates the ambiguity of interpretation of findings of less quantity in an HMO. Is the lower quantity achieved through demand shifting or simple quantity rationing?

In terms of the economic view of PID, there are theoretical reasons to believe that PID, in the way we have defined it, exists to some degree. Consider a physician who is giving the “optimal” amount of information to a patient, and the patient is using his optimal quantity. An envelope theorem argument can be made that around this point, a small increase up or down in quantity has a small impact on consumer welfare, because the consumer is near his privately optimal point. The physician, by contrast, may gain or lose money (depending on the payment incentives) from inducing the patient to demand more or less. Whatever particular model we assume about physician motivation, the nature of the tradeoff presented to the physician – I can gain income by a change that has a very small effect on the welfare of my patient – implies that the physician will be doing some demand inducement. In a recent overview of health economics, Blaug (1998, p. 567) contends that, “it is only the quantitative impact . . . of supplier-induced

54 Socioeconomic factors, such as race, have been shown in many studies to influence rates of treatment. See, e.g., Lee et al. (1997).
demand that is a bone of contention among American health economists." As Pauly (1980, p. 51) puts it: "Other things equal, physicians would rather tell the truth, but they would be willing to surrender some accuracy for some amount of money income." Once that tradeoff is admitted, it is hard to avoid the conclusion that the physician will be inducing some demand.

5.1. Theory of demand inducement

The theory of demand inducement has received some but not extensive treatment in health economics, surprising in light of the attention paid to the concept in the empirical and policy literature. Evans (1974) proposed that physicians maximize a utility function including income and inducement as arguments, and the disutility of inducement limited the physician's income generation. Fuchs (1978) graphically represented inducement as physicians' ability to shift a market demand curve, without addressing the mechanisms or the limits of inducement.

Any seller gains from a higher demand, and unless there is some cost to inducement, a doctor pursuing net income would induce demand to an infinite extent. It is necessary, therefore, in models of demand inducement, to introduce some limit or cost to inducement. Stano (1987a) takes one direction, making the natural analogy between inducement and advertising. He assumes that inducement has a real resource cost (like advertising) and is limited by the profit calculations of doctors in the presence of diminishing returns. More common are approaches that follow Evans, where inducement is regarded as inherently unpleasant, and limited by the psychic costs the physician bears when she gives advice to the patient slanted toward her own self interest. This conception of the cost of inducement fits well with definitions of inducement we have been working with. Only influences on demand that push the patient away from the optimal consumption point impose psychic costs on the doctor.

There is a distinction in the literature between models of inducement that limit inducement within a profit maximization context [Dranove (1988), Stano (1987a)] and those that incorporate a disutility of acting against the best interest of the patient [Evans (1974), Fuchs (1978), McGuire and Pauly (1991), Gruber and Owings (1996), Zweifel and Breyer (1997), Carlsen and Grytten (1998)]. Although the discussion of physician objectives other than profit maximization is primarily a subject for Section 6, we note here that the empirical literature on PID is set predominantly within models of physician utility maximization.

56 As we discussed in Section 4, Dranove (1988) proposed a model of inducement wherein the physician exploits her superior informational position, but is limited by the loss in credibility she suffers by being too aggressive in inducing demand. This model is useful for showing that with asymmetric information, we can expect some inducement. For many purposes we need to go beyond a model that shows there will be demand inducement in equilibrium to address whether the degree of demand inducement changes with changes in the conditions of the market for physician services.
McGuire and Pauly (1991) formalized the ideas of Evans and Fuchs in the context of a model intended to draw the implications of PID for physician response to fee changes. Inducement was limited by physician disutility. Gruber and Owings (1996) expanded on the McGuire and Pauly model by adding a parameter to capture the overall demand and supply conditions. The expanded model can be used to interpret the two main types of empirical studies on PID: physician response to changes in MD/population ratios, and physician response to fee changes.

Modifying the McGuire and Pauly model along the lines of Gruber and Owings, we can write the physician's utility maximization problem as follows:

$$\begin{align*}
\text{Max } U &= U(Y, I), \\
\text{where } Y &= N(m_1x_1(i_1) + m_2x_2(i_2)), \\
I &= N(i_1 + i_2). \\
\end{align*}$$

(5.1)

The physician has utility $U$ which depends on her net income $Y$ and the total inducement she conducts, $I$. $U_Y > 0$; $U_I < 0$; $U_{YY} < 0$; $U_{II} < 0$. She sees $N$ patients who use services 1, 2. Quantity of each service is $x$, affected by the level of inducement $i$. $x' > 0$; $x'' < 0$. $m$ is the margin for each service equal to the difference between the fee the doctor is paid and the cost of the service. Other factors influencing demand, such as patient cost sharing, are suppressed since they do not change.\(^{57}\) The physician chooses $i_1$ and $i_2$ to maximize utility.\(^{58}\) Deriving the first order conditions with respect to $i_1$ and $i_2$, and rewriting them, we can describe utility maximization as follows:

$$m_1x_1' = m_2x_2' = -U_I / U_Y.$$  

(5.2)

The marginal (dollar) return to inducement for each service must be equated to the marginal psychic cost (in dollar terms) of inducement. The parameter $N$, while not explicit in (5.2) is of course one of the arguments in the function for $U_Y$ and $U_I$. The pair of equations in (5.2) can now be used to interpret the effects of a change in the number of patients per doctor as a result of, say, increasing the number of physicians, or a change in fees, that is, a change in the margins.\(^{59}\)

---

\(^{57}\) The model in (5.1) could be regarded as a generalization of (4.2). Effort could be reinterpreted as "inducement," and in addition to affecting utilization, enters into utility directly.

\(^{58}\) A number of elements of this model were contained in a paper by Van Doorslaer and Geurts (1987). Their paper looked at the effects of relative prices and income on physiotherapists' treatment decisions. The interpreted the effects in terms similar to McGuire and Pauly.

\(^{59}\) In a regulated fee environment, Dranove's (1988) model based on physician net income maximization predicts the effect of changes in fees on inducement to be zero. Express a physician's net income as $Y = N(i)x(i)m$, and following Dranove, let the number of patients be a negative function of $i$. It is obvious that maximization of net income by choice of $i$ is unaffected by $m$. The physician simply chooses an $i$ to maximize the total volume of services she provides. This would be true in a multiple payer context as well, though some cross elasticities could be generated if physician services were produced at non-constant costs.
A decrease in $N$, as might be brought about by an increase in the supply of physicians, affects only the third term, $-U_I/U_Y$. Decreasing $N$ decreases the quantity of inducement (in total) as well as decreasing income, thereby decreasing the marginal disutility of inducement and increasing the marginal utility of income. Together, these changes reduce the value of $-U_I/U_Y$. The amount of inducement necessary to bring the return to inducement into equality with this new value must therefore increase. Note that this effect depends on an income effect—the changing tradeoff between $I$ and $Y$ as income changes. Empirically then, the impact of a change in the number of physicians per capita, as studied in many papers on physician-induced demand, is essentially the product of the change in supply on income and the income effect on inducement.\(^{60}\)

Suppose now the payer for service one (imagine this to be Medicare) reduces its fee, reducing $m_1$. Service two, paid by another payer, does not change its fee. The effect of this can be thought of in two parts, an income effect and a substitution effect. The income effect comes about because $-U_I/U_Y$ will fall because of the increase in $U_Y$. This will tend to increase inducement for both services 1 and 2.\(^{61}\) There is also a substitution effect which comes about because a reduction in $m_1$ reduces the return to inducement in sector 1. This effect will tend to reduce inducement for service 1 and increase it for service 2 to restore the equality in (5.2). Thus, the effect of a fee reduction in service 1 on inducement for service 1 is ambiguous, depending on income and substitution effects, but unambiguous for service 2—inducement should increase—because income and substitution effects work together. When service 1 is small and the margin is low compared to 2, substitution effects can be expected to dominate the income effect.\(^{62}\)

The empirical literature on demand inducement can now be reviewed within this general framework, beginning with papers that study the change in inducement working through an income effect alone.

5.2. Physician-to-population ratios, income effects, and inducement

The early papers testing for PID were concerned with what Pauly (1980) labelled an “availability” effect. Does an exogenous increase in the availability or supply of physicians increase the demand for physicians’ services? Demand and supply analysis implies that greater supply should increase quantity demanded through a price effect

\(^{60}\) Pauly (1980, p. 51) notes that an income-maximizing physician would give the same advice irrespective of the change in the physician-population ratio. To allow for a quantity effect from inducement, “we must enlarge the set of arguments in the physician’s utility function.” A goal of income maximization does not admit income effects. A utility function is necessary for that.

\(^{61}\) Even if the income effect was completely dominant and the physician pursued a target income, the TI model does not imply that all income will be recovered from the service experiencing the fee reduction. In general it will be distributed among all the services a physician supplies. See Section 6.

\(^{62}\) McGuire and Pauly (1991) work out the comparative statics here for cases in which the income effect is all that matters (TI) and when the income effect does not matter at all, and the physician can be regarded as simply maximizing profits. This is discussed more in Section 6 below.
Demand and supply analysis also implies through the process of geographic mobility of physicians that areas with high levels of demand ought also to have high levels of supply. Research on PID tested an additional causal effect. According to the PID hypothesis, areas in which supply is large in relation to demand, physicians’ incomes are depressed, and they make seek to regain some ground by inducing demand.

Empirical tests of the effect of supply on utilization through demand inducement have taken place at the market and individual level. The first studies were at the market level, and these also suffer most from methodological limitations. From a theoretical standpoint, if the market is regarded as monopolistically competitive, the predictions of the effect of a change in a supply parameter (such as physicians per capita) is generally ambiguous, because the position of the demand curve at the physician level may affect the elasticity and the markup, with uncertain effects on the direction of price and quantity change [Reinhardt (1978), Frank (1985)]. When the physician with market power can also set “quality,” the picture is even more clouded. Feldman and Sloan (1988) show that in this case, quantity can go up or down in response to supply shocks. Changes in the number of physicians per capita do not have a direct effect on the return to inducement, but only affect the utility or disutility of inducement through an income effect. Tests of PID examining physician/population ratios are in effect testing the joint hypothesis of induced demand and income effects. If the income effect itself is weak, then inducement may not change even if inducement is going on.

In terms of data, studies have been subject to a number of criticisms relating to the unobservability of key variables, and the possible correlation of these with the key independent variable, physicians per capita. Supply will tend to equal demand in a competitive equilibrium, so unless demand side variables are adequately controlled for, supply will be correlated with utilization because of market equilibrium conditions [Auster and Oaxaca (1981)]. Major determinants of demand and supply, including input prices, and even the money or time price of health care are sometimes omitted from empirical models. Most studies have used cross-sectional variation in supply, leaving open the possibility of severe bias from unobserved variables. Market definition issues (e.g., border crossing) introduce measurement error and possible bias [Dranove and Wehner (1994)]. Using a physician-to-population ratio as a measure of an exogenous income shock in a cross section is dubious [Gruber and Owings (1996)].

From the first, it should be noted, researchers were aware of many of these limitations and took what steps they could to contend with the difficulties. Fuchs (1978) in the first major paper of this type, studied the effect of the supply of surgeons in 22 metropolitan areas in 1963 and 1970. He chose surgery because time price is presumably less of an issue for these serious procedures, and quantity can be fairly readily measured. He used a TSLS regression to replace the actual number of surgeons by a fitted valued using metropolitan/nonmetropolitan area, hotel receipts, and percent white in the population in the first stage regression. A ten percent increase in the number of surgeons increased the rate of surgery by three percent, according to his estimates.

Cromwell and Mitchell (1986) address the same question as Fuchs, using the same methodology with more years, more and smaller geographic areas and better controls,
and find results consistent with the earlier analysis, though with a reduced estimated inducement effect. Following up on an idea of Green's (1978) that some physician markets might be in shortage or surplus, and that we ought to be able to see an availability effect more readily in a shortage area, Cromwell and Mitchell found a much larger effect of supply on use in areas of high surgeon workload (elasticity of 0.28) than in areas with low workloads (elasticity of 0.09). Carlson and Grytten (1998) point out that a bigger availability effect in "shortage" areas may be more consistent with a rationing effort than with PID. Using a similar methodology, Birch (1988) and Grytten et al. (1990) find that the number of dentists per capita has a strong effect on the volume of dental visits. Cross-sectional variation in physician-to-population ratios used to estimate a demand inducement effect have left skeptics unpersuaded on theoretical [Feldman and Sloan (1988)] and empirical grounds [Phelps (1986)].

Two recent papers have enlivened the literature on PID and market-level effects of physicians per capita. Dranove and Wehner (1994) tested for "induced demand" in a case where it surely does not exist: the effect of the obstetrician-to-population ratio on the volume of births. They reasoned that if they found evidence for induced demand using the techniques common in the earlier studies of surgery, the results of those studies would be suspect. Mimicking the Fuchs and the Cromwell and Mitchell methodology, Dranove and Wehner indeed did find evidence that obstetricians appear to induce births (in an economic sense), substantiating the omitted variable and border-crossing criticisms of the earlier studies.

Gruber and Owings (1996) also looked at births, studying the income shock to obstetrician/gynecologists resulting from the 13.5% fall in fertility among US women occurring over the period 1970–1982. They reasoned that an income effect should lead obstetrician/gynecologists to induce demand for the more lucrative Caesarian section procedure over vaginal deliveries. The timing and magnitude of the decline differed across states, and Gruber and Owings found a strong correlation between within-state declines in fertility and within-state increases in Caesarean section rates, though the absolute magnitude was small. A 10 percent fertility drop corresponded to an increase of 0.6 percent in the probability of a C-section. Gruber and Owings (1996, p. 113) calculate that obstetrician/gynecologists replace about 10% of the fertility-caused income drop by an increase in C-sections. (These results do not preclude other income-recovery effects. Obstetrician/gynecologists could have changed the level of demand inducement for the gynecologist side of their practice as well. Births account for about half the income of obstetrician/gynecologist specialists, though physicians tend to specialize in either obstetrics or gynecology.)

53 In the case of the correlation between quantity of physicians and use, a number of authors [Freh (1996)] have speculated that this can be explained by rationing. If prices are controlled, e.g., in Medicaid or in other plans, use is not demand-determined. An increase in supply increases use but by decreasing rationing, not through a PID effort. Comanor (1988) found a strong availability effect (Paul's term) in Ontario between 1972 and 1975. Adding specialists to an area increases use of specialty care almost one-for-one. A result like this would be consistent with demand rationing as well as PID.
54 For discussion of European empirical studies, see Zweifel and Breyer (1997).
Gruber and Owings make a good case that within-state fertility declines can be regarded as an exogenous shock to demand and income, and therefore represent a good test of demand inducement occurring through an income effect. Extensive specification checks attempting to rule out alternative explanations for the correlation between within-state fertility declines and C-section rates lend confidence to the demand-inducement interpretation.65

Some research uses individual level data for which market level measures of supply could be taken as exogenous.66 Pauly (1980) used data from the 1970 Health Interview Survey from the National Center for Health Statistics, with information on over 100,000 persons. He predicted that inducement effects should be largest for the least well-informed consumers, those with low income in big cities (big cities because information on each physician is less available). He split his sample and found an availability effect for ambulatory care (not hospital care) for the group he predicted, but this was small. Pauly was explicit in recognizing that he only measured a change in inducement with a change in physician supply. "If they [physicians] do manipulate information in an effective way, they do so to approximately the same extent regardless of how many of them there are in an area, and regardless of how busy they are" (p. 16).

In a pair of papers, Rossiter and Wilensky (1983, 1984) distinguished between physician-initiated and patient-initiated visits, and between visits for more or less discretionary procedures. When health insurance and other patient level controls were taken account of, the effect of physicians per capita on use was small, and statistically significant only for more discretionary procedures. [See Stano (1987a) for critique and discussion of these studies.] Scott and Shiel (1997) later studied GPs in Australia and found spotty and small effects of GP density on physician-initiated follow up visits.

The PID hypothesis can be looked at for its implications for supply per physician as well as demand per patient. Viewed from the supply side, a strong PID effect should allow a physician to insulate herself from competition by inducing more to make up for the fewer patients to go around. The Newhouse et al. (1982) results discussed in Section 2 on location are inconsistent with full insulation. McCarthy (1985) studied the number of visits supplied by primary care practitioners in 1975 using data from the AMA about price, waiting time, and other information about the physicians' practice, as well as market-level data on physicians per capita. In most specifications, full insulation (consistent with powerful PID) could not be ruled out, but his estimates of this effect were not very precise.67 McCarthy (1985) also estimated that demand for individual primary care practitioners to be highly price responsive, around –3. If demand is very elastic, the physician's price-cost margin is also likely to be low. In this setting, inducing demand has little payoff to the physician in comparison to where marginal induced visits bring in high profits [Stano (1987b)].

65 For a corroborating study using fees but in the same clinical area, see Keeler and Brodie (1993).
66 The argument is that individual level unobservables are less correlated with overall market demand.
67 He noted that density of primary care practitioners was highly correlated with other important explanatory variables.
The basic premise behind PID is that physicians may exploit the information gap between themselves and patients. If so, as Pauly noted in his study of high- and low-educated consumers, more PID (e.g., more surgery) should be observed where the information gap is greater. Bunker and Brown (1974) reasoned that the smallest information gap should be between physicians and patients who were themselves physicians or their families. Rates of surgery, including "discretionary" procedures like appendectomies and hysterectomies were actually higher among Stanford University Medical School faculty and their spouses than among a group of other professionals and their spouses, controlling for age and sex. It may have been, as Hay and Leahy (1982) speculate, that the physician families may have faced lower prices because of professional courtesies or better insurance, or had easier access to better services, perhaps explaining their higher rates of use. To address this, Hay and Leahy used survey data, more extensive controls, but also found the physician-families used more services. (When professional courtesy was reported these few observations were dropped but may not have been fully reported since there were only four occurrences in 7800 respondents.)

With the exception of the recent study by Gruber and Owings, the evidence for PID with the "availability effect" is equivocal. Recall, however, what those studies test: that there is an income effect on PID. Absence of an income effect on PID does not, of course, imply that PID is not taking place, or may not respond to other exogenous market changes.68 By the late 1980's, physician fees were being set by payers, and changes in fees were being examined for their effect on PID.

5.3. Fees and inducement

After the success of Medicare's price control program for hospitals embodied in the DRG system in the early 1980's, the largest payer in the US turned its attention in the later part of the decade to developing a price-setting policy for physicians (Pauly (2000)). Medicare had imposed various price controls in its program over the years, but had never undertaken full-scale "rationalization" of prices. The existing pattern of prices were regarded as irrational and distortionary, contributing to the over use of invasive procedures [Hsiao et al. (1988a)]. Physicians, by and large, went along with fee reforms because Medicare promised to conduct the reforms in a revenue-neutral fashion (like DRG reform), to rationalize, not reduce, prices, at least on average.

The threat of PID in response to fee changes made Medicare actuaries skittish, so much so that they interpreted "revenue neutrality" to require a cut in average fees. The problem was that when the anticipated (by the actuaries) demand inducement was figured into the revenue-neutral calculations, fees had to be reduced by an extra 6.5 percent to keep Medicare's books in balance. The actuaries figured that each 1% reduction in a fee would lead to 0.5% increase in volume. The actuaries' predictions were not borne

68 As Phelps (1997, p. 214) points out, "none of these studies could show that inducement was not occurring, but only that an alternative explanation existed."
out. The dreaded "volume offset" failed to materialize at least at the aggregate level. During the period 1991–1996, when the new fees were phased in, the price-adjusted volume increase for surgery (where fees were reduced) was lower than for primary care or evaluation and management services (where fees were raised) [Medicare Payment Advisory Commission (1998)].

In terms of research, studying PID in the context of changes in regulated fees has some advantages over the earlier literature concerned with an availability effect. As we noted above in Section 5.1, a change in a regulated fee changes the physician's fee/cost margin and directly changes the incentives to induce, without relying on the transmission of an impact through a potential income effect. Even profit-maximizing physicians may respond to margin changes by changing inducement. Income effects may matter, of course, but may not be as empirically important as substitution effects. In the case of an own-fee change, income and substitution effects of a fee change work in opposite directions, a fee reduction tending to increase inducement for all services because of an income effect, but making inducement less remunerative tending to decrease inducement for the service directly affected. Income and substitution effects work in the same direction for cross-fee effects, suggesting such cross-effects as a promising area for empirical research.

Fee change studies also have advantage over availability effect studies from an empirical perspective. Fee changes follow from regulatory policy that can more be readily regarded as exogenous to the physician's practice or the consumer's demand. Hadley and Lee (1978) report that the utilization growth in Medicare during the price freeze of 1972–74 was so great that the rate of growth of total costs exceeded the growth after prices were unfrozen in 1975. The alternative explanation, common to a number of such studies, is that a price freeze reduces the price to the consumer facing a given circumstance so that it is a demand response. In Medicare, balance billing was falling over this period as well, possibly accounting for some rise in demand.

In another market-level study of provincial billings in Ontario, Canada over the period 1975–1987, Hurley et al. (1990) studied fees and rates for 28 procedures. No discernible pattern in the negative and positive responses to own-fee changes was found. Hurley and Labelle (1995) later found no evidence for a utilization response to fee changes in Canada. Escarce (1993b) also found a mix of positive and negative responses when responses were studied procedure by procedure. This may reflect the ambiguity of theoretical predictions about the direction of change for an own-price effect. The Canadian data have the advantage of being generated in a single payer system with no balance billing. Billing data could be aggregated to the practice level, and income and substitution effects could perhaps be separately identified, by for example, variation in the baseline composition of certain procedures at the physician level. Rochaix (1993) found that GPs in Quebec subject to quarterly income thresholds at which fees were lowered, responded by changing the mix of services they supplied to patients.

69 See also Holahan and Scanlon (1978), Feldman and Sloan (1988).
In a well-known study, Rice (1983) examined rates of procedures per encounter with physicians in Colorado following administered price changes. If we view the encounter as an episode of care, the experiment is meant to investigate how fee changes affected the physician’s “practice style.” In 1977, Medicare in Colorado began to set fees according to state-wide averages, which had the effect of reducing fees for the previously higher paid physicians in the Denver–Boulder area, and increasing fees for physicians located elsewhere. Although Rice did not couch his model in terms of income effects and own- and cross-price effects, he did regress measures of the quantity of surgical services, medical services, and ancillary services, on changes in prices paid for each of these services at the physician level. In some regressions he included only own-price effects and in others included some cross effects. For surgical services, the own-price effect was negative and smaller in absolute value than the also negative cross-price effect from medical services. This pattern is consistent with income and substitution effects working in the same direction for the cross effect, and in opposite directions (but the income effect winning out) for the own-price effect. For surgical and ancillary services, the pattern of own and cross effects (where estimated) did not appear to be consistent with an income and substitution effect model.70

Nguyen and Derrick (1997) studied “overpriced procedures” for which Medicare reduced fees in 1990. They improved on the earlier literature by aggregating effects to the medical practice level. Using 1989 quantities as weight, they constructed a price index for each physicians practice and examined the impact of this price index on an index of Medicare volume. They did not disaggregate the income and substitution effects of the price change but interpreted their results in these terms. Overall there were no significant volume responses (income effects just balanced by substitution effects) but for the 20% of physicians who experienced the largest price reductions, there was a significant negative net income effect. For these physicians, a one percent reduction in price led to an increase in volume of about 0.4%.

Surgeons were most adversely affected by Medicare fee reform. Thoracic surgeons were projected to lose 26% of their income [assuming constant volume; see Levy et al. (1990)], making them one of the hardest hit groups, and also, therefore, one of the groups best to study to look for income effects. Yip (1998) applied the McGuire and Pauly model to thoracic surgeons in New York and Washington state.71 She measured the total income impact of a set of Medicare fee reductions for “overpriced procedures” and included this in a series of procedure-level regressions for Medicare and private insurance patients along with measures of procedure price and other covariates. She found strong evidence that Medicare fee cuts led to increased volumes by thoracic surgeons.

---

70 In a companion paper, Rice and McCall (1983) study the effect of the Medicare rate changes on physician pricing behavior and the willingness of the physicians to accept the Medicare payment as full payment. There, the responses were consistent with conventional economic behavior, rate increases led to price increases and increased willingness to accept Medicare fees as full payment (accept "assignment").

71 These two states included physician identifiers in these discharge abstract files. Also, thoracic surgeons practice is primarily hospital-based.
to both Medicare and private payers, and that their effect worked through the income effect. Taking responses across the board into account, Yip estimates that thoracic surgeons recouped 70% of the income lost by price inductions by a volume increase.\footnote{Yip (1998) studied both the increase in volume for a given procedure and the switch between less and more generously reimbursed procedures.}

5.4. Other evidence bearing on PID

In addition to changes in numbers of physicians per capita and changes in fees, other research on physician behavior also bears on physicians' persuasive powers. A difficulty in interpreting these studies is in determining which of the three quantity-setting mechanisms is operative.

5.4.1. Defensive medicine

"Defensive medicine" is when a doctor conducts procedures in order to protect herself against litigation [Danzon (2000)]. A economically pure instance of defensive medicine would be when a procedure provides no benefit to the patient or involves risk to the patient, but the physician recommends the procedure anyway for selfish reasons. Obviously, a fully informed and price-taking patient paying any part of the cost or submitting to the risk would not agree to such a procedure. If such procedures are observed, some physician quantity setting power must be in place. Note that this is not necessarily demand inducement, but could be the simply the first quantity setting mechanism identified in Section 3.

Most malpractice goes undetected, unpunished, and uncompensated. In one well-known study [Locasio et al. (1991)], 98 percent of patients injured by negligence (malpractice) did not sue. At the same time, physicians perceive the risk of malpractice to be a significant one. Lawthers et al. (1992) report that doctors overestimate the risk of being sued by a factor of three, and the risk of being sued contingent on committing malpractice by a factor of 30. In qualitative responses, doctors regularly report that fear of malpractice motivates their actions. Blendon et al. (1993) found that 32 percent of US physicians admit that they "often" do more than is clinically appropriate; another 29 percent admit to "sometimes" doing too much. Four of five doctors from the Lawthers et al. study agree that they order extra procedures to protect themselves against malpractice. Some older empirical studies of lab tests asking doctors to respond quantitatively to how much of lab activity is due to malpractice get estimates of from nothing to negligible amounts [Garg et al. (1978), Lusted (1977), Werman et al. (1980), Hirsh and Dickey (1983)]. More recently, Kessler and McClellan (1996) studied the effect of malpractice

\footnote{In a similar study, Tai-Seale et al. (1998) also used hospital level data in Medicare and private insurance to study volume responses to Medicare fee-reductions to "overvalued procedures." The directions of effects were generally consistent with the income and substitution effect framework, although there were many data limitations, and many estimated effects were insignificantly different from zero.}
liability reform on the treatment of heart disease among Medicare beneficiaries. They concluded that reforms, hypothesized to reduce fear of liability among doctors, caused a 5–9 percent reduction in medical expenditures.

5.4.2. Self referrals

When physicians have financial ownership in testing and therapy facilities to which they can refer patients, they refer more often [Hillman et al. (1990, 1992)], and patients may be treated more intensively [Mitchell and Sass (1995)]. Self ownership in a regulated price environment is like a fee increase, the interpretation in terms of PID is thus similar to Section 5.3. Extremely high rates of use of prescriptions in connection with physician office visits in Japan is partly attributable to physician dispensing authority and how regulated office visit fees there [Scherer (2000)]. The temptations physicians face in generating income at patients’ expense when physicians own ancillary facilities is recognized by restrictions in federal statutes [see Getzen (1997, pp. 146–148)]. American physicians are prohibited from owning pharmacies, and are restricted in their rights with respect to other types of referral destinations.

5.5. Summary comments on PID

Returning to the Fuchs/Pauly definition of PID, two things must be established for evidence on physician control or influence to support the PID hypothesis. First, the exercise of control must be in the interest of the physician, not the patient. This criterion, even without a gold standard of what patients ought to get, seems to have been met by the studies reviewed here. Adding up the evidence, on obstetricians doing more C-sections, surgeons doing more bypass operations, physicians referring more frequently to their own labs, and other studies, makes a convincing case that doctors can influence quantity and sometimes do so for their own purposes.

The second criterion from the Fuchs/Pauly definition is that the physician exercise quantity control by influencing the patient’s demand, not by quantity setting through rationing, or via the quantity setting power available to the monopolistic competitor. Consider first the availability–effect literature. The profit-maximizing, quantity-setting monopolistic competitor is not influenced by income effects. If adding more doctors to a market is associated with more aggressive quantity setting as in Gruber and Owings, the net income or profit-maximizing model is contradicted, and the PID model is supported. It is possible to come up with other income–effect mechanisms to explain the availability effect finding, but all of these must also sacrifice the objective of income maximization. It also is possible to construct other explanations for the association

74 There are many possibilities. The physician might, for example, choose quantity trading off her own interest with the interest of the patient (Section 6.1). The income effect would then have the effect of alternating this tradeoff more in favor of her own interests. The general idea is that the physician has some other objective that becomes less important in relation to her own interests when her income falls.
between availability and quantity per patient, involving time price/access to patients, or even of effects via the market power of physicians being enhanced by more physicians in a region [Pauly and Satterthwaite (1981)]. Without more direct evidence for these possibilities, the income effect-inducement hypothesis gains support from this segment of the availability-effect literature. Much of the availability-effect literature is, however, vulnerable to the statistical artifact argument [Dranove and Wehner (1994)].

In the fee-effect literature, if fees fall for a small part of a physician’s practice, the physician’s response will not be revealing regarding PID. If quantity falls, this might be PID or standard supply response. If quantity goes up, it is impossible to interpret this as an income-effect driven increase in inducement – the income effect would play out for all services, and for the segment of the practice where fee fell, would have to work against the substitution effect [McGuire and Pauly (1991)]. There are a few papers [e.g., Yip (1998)] for which Medicare fee changes had a large enough income impact on physicians for there to be credible income effects at work. What about demand-shifting versus quantity-setting? For a fee fall, the quantity-setting model (or even a simpler price-taking consumer model) generates a negative fee-quantity prediction if the fee reduction reduces the price to the patient and loosens the demand constraint. This could have been operative in some of the earlier work on Medicare fees [Rice (1983)], but is unlikely in the later period studied by Yip (1998) when there was more supplemental insurance by the elderly, and for procedures where the surgeons’ fee was a small part of the financial and personal cost.

In sum, there is a large volume of research on PID that is supportive of PID but not highly discriminatory between the PID hypothesis and theories with fixed patient preferences. Some recent papers have refined the tests for PID using income effects and have provide more direct support for the PID idea. It is worth recalling at this juncture, as Phelps (1997) and others have pointed out, that the studies discussed are about changes in the intensity of PID following income changes. PID could well be empirically important but simply not respond to income effects. A no-effect finding in the literature therefore does not contradict the existence of PID. It is hard to dispute Fuchs (1996, p. 3), in his presidential address to the AEA reviewing the state of health economics, where he says that, “Despite many attempts to discredit it [citing Dranove and Wehner (1994)], the hypothesis that fee-for-service physicians can and do induce demand for their services is alive and well [citing Gruber and Owings (1996)].”

The “fee-for-service” physicians in Fuchs’ quote are now increasingly being paid by managed care plans, and their response to this environment presents new opportunities to test PID. Empirical cases are appearing in which insurance benefits to patients are improved, and use falls due to other managed care changes implemented at the same time [Ma and McGuire (1998)], strong evidence for the existence of some physician control over quantity. Determination of the discharge status of the patient ("satisfied,"

---

75 This leaves open the issue of how such findings are to be interpreted. One possibility is suggested in Section 6.3.3 below.
“rationed”), and how this changes with managed care would be one fruitful direction for distinguishing between PID and rationing hypotheses.

It seems appropriate to return to the question of what difference it makes if physician quantity control is exercised through manipulation of demand/preferences or a more direct quantity control without changing demand. For some purposes it does not matter. From the normative standpoint of evaluating whether a change in fees exacerbates overutilization, the comparison is with respect to some standard of efficiency—a cost-effectiveness standard from clinical research or the mystical perfect-agent demand curve. If a physician increases quantity beyond the desired point, it is irrelevant for the planner whether the mechanism is preference shift or direct quantity setting.

From the standpoint of running a market in health plans, however, the distinction between PID or direct quantity setting does matter. Much of contemporary health policy is based on the assumption that a market of competing health plans can be the basis of an efficient health care system. For this strategy to succeed, consumers must be able to evaluate and choose plans in their self-interest. If plan A rations me too tightly, I leave it and go to plan B. This process will work if I can evaluate when I am rationed (or given too many things I know I don’t really need). It works, in other words, if physician quantity setting is seen for what it is by consumers. If, however, physicians use PID in competing health plans, and patients are persuaded to like what they get, the basic mechanism for encouraging efficiency in rationing among competing managed care plans breaks down. Given the direction of health policy in the US and many developed countries, it is more important than ever to know if, when physicians do less in managed care, they persuade patients that this is good for them, or if patients are knowingly choosing a plan that rations them to efficiently contend with the prisoners’ dilemma of moral hazard in health insurance.  

In his presidential address, Fuchs (1996, p. 8) also reported the results from a sample of American health economists, economic theorists, and physicians. 68 percent of health economists, 77 percent of theorists, and 67 percent of practicing physicians agreed with the following statement:

Physicians have the power to influence their patients’ utilization of services (i.e. shift the demand curve), and their propensity to induce utilization varies inversely with the level of demand.

The majority of respondents agree that two things are true, that physicians induce demand and they do so in a way that does not reflect simple profit or income maximization. In the next section, we turn to the question of the objectives being pursued in physician decisionmaking.

76 Of course, alternative directions for policy, such as relying on a “single-payer” system without competition, also have their own set of distortions and inefficiencies.
6. Other physician objectives

Many years before his survey of economists and physicians, Fuchs warned in his influential book (1974, p. 60), *Who Shall Live?:* “A common mistake is to think that the behavior of physicians can be understood only in terms of their desire to maximize income.” A “charity hypothesis” [Kessel (1958)], concern for medical ethics [Arrow (1963)], desire for interesting cases [Feldstein (1970)], and target income [Newhouse and Sloan (1972), Evans (1974)] had all been proposed by the time Fuchs was writing. One way or another, these alternative hypotheses incorporated a concern for patient health or economic welfare into physician objectives.

Neoclassical economic analysis derives predictions about behavior of suppliers by assuming that suppliers make decisions in order to maximize the profit of the firm. Reservations about the profit maximization assumption are of course not confined to physicians and the field of health economics. “Alternative theories of the firm” relying on utility maximization, bargaining models, or behavioral models, with parallels in the health literature, have been around for many years. While this literature has had only a limited impact on mainstream economics, physicians seem a particularly good candidate for alternative objective functions. Physician-firms are most often “owner-operated” permitting the ready indulgence of utility-related objectives. Physicians enjoy very high incomes, and if a falling marginal utility of income renders income less attractive in relation to other objectives, physicians are likely to be susceptible to an income effect. Physician decisions affect their customers in a profound way – the “cost” to the client of physicians’ opportunistic behavior may be much higher than it is for other suppliers with comparable informational advantages. Therefore, physicians, facing a tradeoff between what is good for my customer and what is good for me, may be induced to sacrifice more income because of the steep tradeoff. Finally, as Arrow (1963) pointed out, the social norm shared in large degree by the buyers and the sellers of health care is that physicians enjoy special independence in decisionmaking in exchange for an understanding that they act in the patient’s best interest.

6.1. Medical ethics as a constraint on choices

Medical ethics involve a wide range of issues from confidentiality, duty to inform patients, end-of-life treatments, as well as our concern about the role of ethics in routine decisions about health care use [Anderson and Glesnes-Anderson (1987)]. Medical ethics command physicians to *primum no nocere* (first, do no harm), and more actively, to “act in the patient’s best interest” [Hiller (1987)]. Arrow (1963) thought that the efficient and fair allocation of health care depended on physicians behaving ethically, not exploiting patients’ vulnerability. Mechanic (1998) argues that the reciprocal relationship of ethical behavior and trust contribute to effective medical care. The imposition of an active third-party payer into clinical decisionmaking through managed care has been alleged to threaten physician’s loyalty to patients [Rodwin (1993), Emanuel and
Goldman (1998), Mechanic (1998), raising new questions in the debate about the role of ethics in medical care.\footnote{Mechanic and Schlesinger (1996) speculate that financial incentives managed care plans give physicians incentives to reduce treatment interfere with the “trust” between physicians and patients that contributes to effective medical care. A physician with incentive to do less than necessary may, however, be as trustworthy as one who has incentives to do too much.}

One interpretation of medical ethics is that in a situation in which there are many choices of how to treat a patient, ethics dictate that the physician chooses the “medically correct” way to proceed. If ethics operated in this way, health care choices would not respond to economic incentives on either the demand and the supply side, as they evidently do. Ethics could fit within a model in which incentives matter by serving to cull some but not all alternatives from what a physician could choose. It might be unethical, for example, for a physician to conduct a Caesarean section in some situations, unethical not to in others, and in some third set of conditions, “ethics” might not come into play in the decision yes or no. This view of the role of ethics seems to fit Hillman’s (1990) meaning: “Whereas most physicians will act in the patients’ best interest when the medical decision is clear-cut, the effect of financial incentives may be more important in areas where the correct decision is not clear.”

Formal treatments of ethics in medical decisionmaking are few. Ma and McGuire (1997) study a model in which ethics are represented by a lower bound on the health benefits a physician is willing to provide to a patient. Physicians control one input (non-optional visit) and the patient controls the other (contractible visits). If the physician anticipates that a patient will not choose extensive treatment, perhaps because of the prices the patient is paying, the physician feels compelled to make up for this by putting in more of the input she controls. A payer can take advantage of a physician’s ethical constraint by setting up a payment system that puts the physician in the position of being forced to take more effort to make sure the patient attains an acceptable outcome.

6.2. Utility and the patient’s best interest

An “ethic” has the flavor of a dictate or a constraint – once the constraint is binding, other objectives of the physician become irrelevant. Perhaps for this reason, most papers in health economics do not use a constraint to represent ethics, but instead represent physician concern for patients with a utility function including as an argument something valued by the patient (quantity, quality) or the patient’s utility itself. In this construction, the physician’s ethically driven concern for patients is subject to being traded off against self-interest.\footnote{For an early model of this type, see Woodward and Warren-Bolton (1984). There, “ethics” took the form of a cost felt by the physician as the actual treatment diverged from the “medically appropriate” treatment.} Models with physician induced-demand reviewed above in Section 5 often use physician concern for patients, Pauly’s “internal conscience,” as a brake on inducement. Eisenberg (1986, p. 57) expresses the role of patient welfare in
a way consistent with the argument-in-the-utility-function interpretation, "A substantial part of the physician's satisfaction with practice is fulfilled by serving successfully as a patient's advocate." He reviews evidence to show that prices patient's pay affect the physician's behavior, implying a concern for patients' overall well-being (not just health) on the part of the physician.

Other papers endow physicians with a utility function of the form $U(\pi, B)$, where $\pi$ is the physician's net income, and $B$ is the benefits or utility the patient receives [e.g., Chalkley and Malcolmson (1998), Ellis and McGuire (1986), Ma and Riordan (1998), Rosenthal (1998)]. An attractive feature of such a formulation is that it can be used to derive a "supply curve" of services to a patient as a function of the degree of supply-side cost sharing. Substituting a payment system with a parameter for supply-side cost sharing readily generates comparative statics [Rosenthal (1998), see also Jennison and Ellis (1987)]. With assumptions about the form of $U(\cdot)$, supply-side payment systems can be solved for the form which maximizes social surplus (normally, simply $\pi + B$).

Another interpretation of $U(\pi, B)$ is that it represents a Roth–Nash solution to a cooperative game between a patient and a physician disputing what quantity the patient should consume. The patient demands $x_d$ on the basis of his insurance coverage, the physician desires to supply $x_s$ on the basis of her payment. Maximizing $U(\cdot)$ with respect to $x$ represents the axiomatic solution to the bargaining [Ellis and McGuire (1990)]. This interpretation can be regarded as a generalization of Program II in Section 3. The maximand expressed in Equation (3.5) maximized profit subject to a constraint on consumer utility. In terms of a bargaining model, the consumers $NB^D$ is the "point of minimal expectations," and all the bargaining power rests with the physician [Roth (1979)]. More generally, bargaining power is shared between the two parties and the quantity settled upon will fall between the quantities desired by the two parties.

6.3. Target incomes

A radical alternative to the assumption of profit maximization in the theory of the firm is the so-called "behavioral theory," pioneered by Simon (1958). Behavioral theory claims firms operate by "rules of thumb" rather than by maximization, with targets for rates of return or markups over cost. In health economics, a prominent behavioral theory proposes that physicians make decisions to maintain a "target income."

---

80 In Ellis and McGuire (1986), the utility function is additive, although the weights on profit and patient benefit need not be equal.
81 In the simple case of quadratic utilities of both the patient and the physician and equal bargaining power, the Roth–Nash solution turns out to be the simple average of the two quantities [Ellis and McGuire (1990)].
6.3.1. Background

The target income (TI) hypothesis explains why higher physician-to-population ratios (presumably a measure of supply in relation to demand) can be associated with a higher price of physician services, not a lower one, as simple price theory would suggest. Suppose physicians only set a price high enough so as to attain some target. They could make more by charging a higher price, but choose not to, perhaps because of concern for patients’ welfare. As more physicians appear in a market and patients are spread more thinly among the available suppliers, physicians must raise prices to maintain the target income. TI behavior, in the 1970s, reflected restraints in pricing. Physicians were “humanitarian” in Farley’s (1986) term, not fully exploiting their price-setting power unless they were forced to by competitive pressures. The TI hypothesis was taken very seriously by health economists and policy makers. The federal government sponsored a conference and published a volume on the supply and pricing issue, The Target Income Hypothesis and Related Issues in Health Manpower Policy (Department of Health, Education and Welfare 1980).

TI behavior, when used to explain physician response to competition from more physicians was not connected to demand inducement (as later it came to be). TI behavior was not about quantity setting, it was about pricing, an explanation for unexploited monopoly power. Indeed, if physicians could induce demand, they could have done more for the fewer patients competition left them, and not raised prices.

In the 1980’s, when direct fee-setting replaced increased supply as the mechanism used by regulators to limit physician prices, TI was used to explain another empirical anomaly, the negative correlation between fees physicians were paid and the quantity supplied (Rice 1983). If physicians wanted a target income and income was P x Q, if you reduced P, then they would increase Q. During the 1980s, writers proposing TI explanations linked it to PPI. Physicians could set quantity because they could induce demand. Interestingly, TI behavior was no longer a form of benevolence. Mr Hyde took over from Dr Jekyll. TI frustrated policies designed to contain health care costs. Physicians were using their power to influence patient utilization for their own (the physicians’) interests in order to counter the well-intentioned regulation of high fees.

6.3.2. Target income or income effects?

In the context of fees and demand inducement, simple TI theory is generally presented as a behavioral, i.e., a nonmaximizing, theory. In the terminology used in this chapter, physician net income is: (p x − c)x. Call p x − c the margin, m, on services. If the physician chooses x by inducing demand so as to hit a target, T, then

\[ x = T / m. \]  

(6.1)

82 An exception in Feldstein’s (1970) discussion of the reasons why he found unconventional relations among price and quantity when attempting to estimate demand and supply curves for physicians.
It is obvious that $dx/dm < 0$; indeed the implication of (6.1) is that the elasticity of $x$ with respect to the margin is $-1$.

There are two severe problems with this theory [McGuire and Pauly (1991)]. First, the idea of a “target” is problematic. It is difficult to explain why physicians would pursue a target income in the first place, difficult to explain how targets are set, and difficult to explain the evident differences in targets across individuals and over time. No one has established that the distribution of incomes cross-sectionally and over time among physicians are any different from what would be expected with conventional maximizing behavior.

Second, the formulation of target income theory in (6.1) is simply conceptually inadequate for explaining physician behavior in typical US markets in which physicians supply services to many payers. Suppose a physician has a target $T$, but supplies services to payer 1 and payer 2, and we designate the margins and quantities for each by subscripts. Then, the combinations of $x_1$ and $x_2$ that satisfy the target are:

$$T = m_1x_1 + m_2x_2. \tag{6.2}$$

This equation does not yield a unique solution for $x_1$ and $x_2$. There are an infinite number of combinations of quantities that satisfy a target for any pair of margins. The TI theory based on (6.1) does not generalize: it is incapable of generating quantity predictions for more than one payer. There is no unique solution, and no testable comparative statics from (6.2) regarding the effect of fees on quantities in a context like the US. The behavioral TI theory does not tell the physician how to choose among the many combinations of $x_1$ and $x_2$ (and more generally the large set of services she supplies) to hit a target.

McGuire and Pauly (1991) propose a utility maximizing framework in which a physician can set quantities for multiple payers through demand inducement. They show that target income behavior and profit (or income) maximization lie at opposite ends of a spectrum of income effects. When income effects are all that matter around a certain point, physicians act so as to pursue a target. When income effects are absent, physicians maximize income.\footnote{Income effects on behavior are generated by changes in the marginal utility of income. Targeting behavior emerges when these changes are very dramatic, that is, when the derivative of the marginal utility of income with respect to income is minus infinity. This accords with the sense of a TI. When income is less than the target, its marginal utility is very high, when income is above the target its marginal utility is very low. Thus, around the target, the marginal utility must fall steeply. And as is well-known, when there are no income effects, firm behavior is not income maximizing.} Furthermore, this framework can be used to generate comparative statics. When one price falls, income generation will be pursued differently with respect to all services supplied, along the lines of the income and substitution effect discussed above in Section 5.3.

It is unnecessary to view TI as an “alternative” to profit maximization. The idea of a “reference” income goes back to Feldstein (1970) who discussed the impact of a
proposed price ceiling imposed by regulation on physician fees in terms of income and substitution effects. Evidence from the literature can be assessed from the point of view of what it says about the magnitude of the income effects, not with regard to a yes/no issue of TI behavior.

Taking this perspective, the recent demand inducement literature [Gruber and Owings (1996), Yip (1998), Tai-Seale et al. (1998)] provides evidence for and explicit discussion of an income effect. The early fee and availability effect literature was debated in terms of a target, not an income effect [see Wedig et al. (1989), Rice and Labelle (1989), Feldman and Sloan (1988)]. Income effects on physicians could be estimated in the same way as in labor economics, with information on non-employment income, such as assets or a spouse’s income. Sloan (1975) and Hurdle and Pope (1989) studied physician supply decisions and the effect of non-practice income, both studies finding evidence for small income effects. Rochaix (1993) found no effect of outside income on supply.

Rizzo and Blumenthal (1996) analyze a survey of young physicians in which questions were asked that could shed light on physician motivations. The young doctors were asked what income they considered to be “adequate,” considering the stage they were at in their career. Rizzo and Blumenthal treated this reported income as a target, and found that when physicians were away from this target, they pushed prices higher, tending to move in the direction of the “target.” This paper recalls the earlier spirit of the TI literature in which physicians exercise restraint in pricing (not pushing as far as they might) as their income reaches the “adequate” range.\(^\text{84}\)

6.3.3. Revenue targeting from a participation constraint

“Targeting” behavior can stem from another source that does not require an assumption that physicians pursue a “target” income. The targeting discussed now can also explain targeting within a single payer. Suppose, following the presentation in Section 3.3 above that a physician’s cost depends on some activities which are reimbursed in a payment system and some which are not directly reimbursed. In the notation introduced above, cost is \(C(x, e)\) where \(x\) is paid upon and \(e\) is not. The revenue function can be expressed as \(R(x)\). The physician must make a decision about whether to accept a certain patient, or patients from among a class, perhaps defined by a payer. To accept a patient, a participation constraint, as it is referred to in the contracting literature, must be satisfied. We can normalize the required profit to be 0, and express the participation constraint as:

\[
R(x) - C(x, e) \geq 0. \tag{6.3}
\]

\(^{84}\) Rizzo and Zeckhauser (1997) reexamine the same data and find that these young doctors increase their hourly earnings more the farther they are away from the target on the downside, but above the target, the distance from the target does not matter. Although the findings are cast in these papers in terms of TI behavior, evidently this is not literally correct since doctors are not at the target, but making tradeoffs to get closer, behavior that can be understood within the more conventional approach of income effects.
Note that this constraint (6.3) has elements of a "target." A physician must gain a certain profit per patient to justify taking on the patient. This follows simply from recognizing that a physician has a certain opportunity cost of time. The (unreimbursed) effort that goes into caring for this patient could be spent elsewhere.

Suppose there is a class of patients (e.g., Medicaid birth-related clients) seen by a physician for whom the fees are reduced by regulation, violating the physician's participation constraint at the old values of $e$ and $x$. The physician has several choices. She can drop the patients and refuse to participate in Medicaid. The physician can "upcode," labeling the procedures she does in a more elaborate fashion, perhaps risking censure or penalty. She can cut back on the time she spends per visit (reduce $e$). If some prices are unregulated but associated with the use of this patient, these prices can be raised to satisfy a participation constraint. Finally, she can generate more billing by increasing the quantity of reimbursed services supplied (increase $x$, if $R' > C_x$). Gabel and Rice (1985) refer this set of physician responses as the "price of paying less." Medicaid and Medicare experience less physician participation as fees are reduced. Physicians upcode in response to fee regulation [Berry et al. (1980), Yip (1998)]. Danson et al. (1984) found evidence directly consistent with the operation of participation constraint: when physician fees for an office visit were limited, physicians compensated by raising fees on the associated lab tests to retain the overall net revenue associated with a visit.

Quantity response at the episode or payer level can be understood as one of the set of things a physician can do to satisfy a participation constraint. Rice's (1983) empirical work on Medicare, for example, is essentially an episode-level analysis. The fee-effect on the participation constraint expressed here is another explanation for the observed $P - Q$ relation.

The participation constraint route to targeting avoids an implausible assumption about motivation. It also avoids the logical gap in TI theory in terms of multiple payers. A participation constraint applies to each payer. Therefore, a "target" behavior observed for one small payer can be explained. Note that unexploited income generation must be available to a physician for any of this set of responses to emerge (except for dropping the patient). Fraud (upcoding), price, or quantity-setting, must be limited by other forces, such as disutility, as in the Evans (1974) or McGuire and Pauly (1991) framework.

7. Conclusion

In the neoclassical theory of the firm, the firm sets price and quantity in order to maximize profit subject to the constraint of market demand. Every phrase in the paradigm has been questioned in the course of this chapter. Do physicians maximize profit? There is abundant evidence that in some circumstances physicians are prepared to trade off income against welfare of the patient. Furthermore, this tradeoff is affected by income effects, in a manner consistent with conventional views of labor supply.

There is not enough evidence, however, to justify keeping the "target income" theory alive. The theory is logically incomplete in a multiple payer environment; there is
no evidence to support this extreme version of income effects; there is a theoretically superior way to generate target-like behavior even within a single payer by invoking a participation constraint.

Following from the view that physicians' tradeoff of other (patient-centered) objectives with income depends on their level of income, the weight on income in physicians' utility may be changing. As managed care plans succeed in limiting the prices charged and quantities supplied, physician income objectives may become paramount. The profit-maximization assumption may be becoming more applicable to physician behavior.

Are physicians constrained by market demand? The answer to this is "yes," even while noting that there are several mechanisms physicians have to influence quantity provided. The understanding of "market demand" must first of all extend beyond the conventional demand curve. In general, even in the most simple models of physicians with some market power, the demand curve, relying as it does on price-taking patients, does not describe prices and quantities in this market. While the "demand curve" has limited use, market demand is still an essential concept. If physicians set quantity only by virtue of the nonretradability of their services, patient benefits (market demand) constrains this activity. If physicians move demand by undertaking nonreimbursed activities perceived as "quality" by patients, demand considerations, in particular how the physician attracts patients, remain relevant. If physicians "induce" demand through a persuasive activity when patients have less information than the physician, market demand response can still limit what even the most self-interested physician can get away with.

A large body of credible research establishes that physicians set quantities, and they do so partly in response to self-interest. An important question for research is to decompose the source of the quantity setting. The welfare economics of quantity setting due to nonretradability within a fixed demand, observable quality or effort, or unobserved persuasive activity are very different. Interpreting the consequences of quantity-setting for purposes of policy depends on assessing the relative strength of the three potential sources.

Do physicians even set price and quantity? Prices for "visits" are easily observable and contractible, and within the grip of third-party payers. Physicians are subject to market forces like other workers, so the prices chosen by health plans are probably best regarded as being determined by demand and supply. Quantities are another matter. The difficulty of third parties' contracting on outcomes (even if the patient observes a signal related to outcomes) means that physicians are certain to remain with discretion about quantities, even when measured in simple terms like "visits" of various types. Economic models, abstracting the complexity of medical decisions into a single dimension of "quantity," give the impression that treatment decisions are more easily monitored and controlled than they really are. Acknowledging the many elements composing treatment—the many inputs, the sequence of events, the observable and the behind-the-scenes activities—leads inevitably to the conclusion that the simple monitoring and incentives devices used by payers leave a great deal of authority about treatment with the physician.
References

 Arrow, K.J. (1971), Essays in the Theory of Risk Bearing (Markham, Chicago).


Eisenberg, J.M. (1986), Doctors' Decisions and the Cost of Medical Care (Health Administration Press, Ann Arbor, MI).


Lusted, L. (1977), "A study of the efficacy of diagnostic radiologic procedures: final report to the National Center for Health Services Research" (National Center for Health Services Research, Rockville, MD).
Newhouse, J.P. (1976), The Economics of Medical Care: A Policy Perspective (Addison Wesley, Reading, MA).


