1 Fee-for-Service Medicine and Its Discontents

In this chapter I address the pathologies of administratively set fee-for-service medical prices. I then contrast those pathologies with the pathologies of more market-oriented methods in much of the remainder of the book. (Economists sometimes say that everything is relative. "More market-oriented methods" is definitely meant as a relative statement.) Before coming to the pathologies of fee-for-service prices, however, I need to lay some background.

American physicians and hospitals, like those in many other countries, were traditionally paid a fee for each service rendered to a patient. A service in this context was typically narrowly defined—for example, a brief office visit, or the interpretation of an electrocardiogram, or a simple blood test. In a world with little or no insurance it may have been reasonable for economists to make standard competitive market assumptions about how fees or prices for such services were set. After all, at least for physicians' services, there are typically many sellers in most urban locations, and there are many buyers as well. But if standard competitive models applied to medical care pricing, there would be no need for this book. One can question the realism or usefulness of an assumption of competitive pricing even in a world with no insurance, but that is not the world in which we live, and I therefore do not propose to discuss such a world here.[1]

Rather, given the widespread insurance or public delivery that exists in every developed economy, the market for the purpose of price determination works through prices set by insurers or the government.[2] For some analytical purposes it is reasonable to abstract from how price is set,[3] but the questions of resource allocation to and within the medical care economy, which are central to both the economics of medical care and to health policy, require that one address that issue. That is the task of this book.

Are Insurance and Competitive Pricing Compatible?

Under some circumstances the presence of widespread insurance might be compatible with competitive pricing, but those circumstances typically do not obtain in medical care. First, insurers might contract to pay consumers a lump-sum payment, conditional on a certain state of the world.[4] In practice the most plausible form of such insurance would be to condition the fixed amount on the patient's disease or diagnosis. It would make little sense, for example, to give a person with kidney failure in need of dialysis or a transplant operation the same sum as a person with a streptococcal throat infection who simply needed an inexpensive antibiotic. Taking account of search costs, the consumer who received the lump sum could presumably then shop for the provider that offered the best combination of quality and price in treating that disease, just as if shopping for food with a voucher for food. The consumer would be motivated to do so for the usual reason—he or she would keep any savings in price.

In fact, insurance, at least in the United States, functions in exactly this way for some goods such as auto repair. An appraiser observes that the automobile's fender, say, is damaged and pays the consumer a lump sum for the fender's repair. The consumer may then shop among alternative suppliers for a favorable price. Sometimes the insurer may give the consumer a list of auto repair businesses that have agreed to repair the fender for the amount of the lump sum, analogous to a network of providers in health insurance who agree to accept the insurer's payment as payment in full. In the case of auto insurance the consumer typically receives the lump sum whether or not the fender is repaired and may in fact choose not to repair it and use the funds for other purposes, though one might then wonder why the consumer has paid a loading fee to purchase the insurance.[5]

The payment of a lump sum is generally not observed for medical care because of the difficulty of determining the lump sum. A physician's services typically consist of both diagnosis and treatment. At the diagnosis stage considerable expense may be incurred just to establish what precisely ails the patient. Generally it would not be satisfactory to establish a lump sum before any diagnostic measures have been undertaken, because how to proceed will often depend upon further information, as results from laboratory tests or radio-

logic images of various sorts become available. Even at the treatment stage establishing a lump sum is problematic, since the illness may respond in various ways to treatment or the disease itself may worsen or improve independent of treatment, both of which may dictate changes in an initial treatment plan. After initial chemotherapy or radiation treatment, for example, a cancer patient may or may not have a remission. Any consumer paid a lump sum that was to cover the entire treatment of an established diagnosis could thus be left bearing considerable risk. Although one occasionally observes insurance policies that pay a lump sum conditional on a specific diagnosis, they are rare.[6]

The difficulty of setting an appropriate lump-sum payment in advance of treatment exemplifies the information problems in medical care. These information problems shape many of the supply-side pricing institutions I consider in this book.

Second, competitive pricing models might also apply if insurers themselves took bids from providers and channeled consumers to the providers they deemed had provided the most favorable bid, including the quality of the services, as with the list of firms that will repair the automobile fender for the lump sum the insurer allows. This arrangement was not the traditional American arrangement, perhaps because different consumers value different physicians or other providers differently, and it was costly to write a contract that covered only the consumer's preferred physician(s), or because the consumer may not know what physicians he or she would prefer in future states of the world. In other words, consumers were willing to pay something for a free choice of physician—or more precisely were willing to pay for a policy that left them paying the same or nearly the same amount at the point of service irrespective of their choice of physician. A different argument is that organized medicine conspired or lobbied to keep policies without free choice off the market (Goldberg and Greenberg 1978). In any event, free choice of physician was how American insurance policies were structured for many years, and in many other countries consumers still pay the same amount irrespective of their choice of physicians (e.g., Canada). Traditional American insurance also constrained the physician not to bill the patient additional amounts over and above what the insurer paid.[7]

Free choice of physician means the services of almost all physicians must be covered by any insurer who wishes to compete in the

insurance market. In turn this means that in any negotiation over price between a physician and an insurer physicians have substantial bargaining power; in practice, the traditional American insurer named a fee in a take-it-or-leave-it contract that ensured the participation of most physicians. Physicians, seeing the advantages of such arrangements, sometimes successfully pressed for legislation that required insurers to contract with all physicians who would accept a given price ("any willing provider"). Even without legislation widespread consumer purchase of such insurance plans exerted pressure on any remaining physicians to contract with the insurer, even if, as was often the case, the insurer did not allow physicians to bill the patient for any amount the physician could not claim from the insurer. In other countries—for example, Germany and Canada—a physician association negotiates a fee schedule with private or public insurers on behalf of virtually all physicians, and physicians are not allowed to bill patients for additional amounts. As in the traditional American system, the presumption is that the patient should have choice among all physicians, which precludes the insurer from just sending patients to physicians with low bids. In more recent years, however, the rise of managed care in the United States has somewhat constrained patients' freedom of choice through the use of provider networks and drug formularies, as I come to in chapter 2.

Third, price competition among traditional American insurers took place only over a minor portion of the insurance premium, the loading or retention kept by the insurer. The expenditures incurred in the medical care system were largely taken as a given by all insurers. The net result was that private insurers acted as both price and quantity takers in the market for medical services.

Why the United States saw little or no price competition over medical services for several decades is puzzling. The proximate explanation is that a passive, self-insured employer generally paid the premium for the insurance policy on behalf of the employee, or paid a large percentage of it, but that explanation, of course, begs the question of why the employer was passive. Employers could potentially have bargained directly with providers for lower fees, but they faced large transactions costs to do so and until relatively recently, they did not do so.[8] Moreover, as long as employers sought to provide their employees with an insurance policy that covered all or almost all providers, they had little bargaining power.

In the case of the large public Medicare program, the government acted in a similar passive fashion for nearly two decades after the enactment of the program in 1965; as described in what follows, it set fees such that almost all physicians would willingly see Medicare patients, and it simply paid for whatever services those physicians ordered on behalf of their patients; in other words, it made no effort to ration covered services.[9] From the point of view of the physician, ordering a service was equivalent to writing a check on the Treasury.

In other words, the prevailing American model from the 1940s into the 1980s in both private and public insurance was that of indemnity insurance—indemnify the insured after the fact for financial losses suffered. Insurance companies and employers behaved as if medical treatment and the resulting bills were an act of God—like an earthquake or a tornado—and independent of the price paid to the physician for the care. It was as if insurers assumed that physicians treated patients according to a template they learned in medical school and postgraduate training, a template that was invariant to reimbursement. As I show, the evidence does not support this assumption.[10] Economists may think this assumption quaint, but many physicians and others still speak in a language of delivering the services that the patient needs, which appears to leave little role for price.[11] Indeed, virtually all American insurance contracts are written so as to cover services that are "medically necessary."

The failure of insurers to compete on the amounts they paid for medical care services meant there was little or no price competition in the provider market. In particular, the individual provider had little, if any incentive to cut price because the demand facing him or her would be little changed if the fee charged were lower. That was because the price to the insured patient, or the demand price, typically changed only modestly, if at all, when the provider changed the price charged the insurer, or the supply price. Thus, the standard market mechanisms for eliminating rents—or prices above average cost—were weak in the market for medical services, and did not operate at all in the limiting case of insurance that reimbursed patients in full at the margin (Newhouse 1981). The rise of managed care in the United States and the consequent development of provider networks—meaning the ability of the insurer to write a contract that reimbursed patients at a less favorable rate unless they used specific providers—increased the elasticity of demand providers faced and hence reduced rents, as I come to in chapter 2.

Given free choice of physician, fees were set in an administrative transaction between the insurer and the provider. In the United States insurers often named a price that was above the reservation price of most physicians and so ensured their participation. Indeed, the Medicare program, the largest insurance program in the world after its establishment in 1966, building on language developed in the private insurance industry, agreed to pay "customary, prevailing, and reasonable" fees to physicians. Such fees came to be defined operationally by an elaborate set of rules—rules that grew steadily more elaborate over time as is typical of administered price systems. In describing these rules, the word *rococo* comes to mind.

One can presume that the resulting American fees included rents, however, because the passivity of the employer offered little incentive for the insurer to bargain for prices near competitive rates and because of the provider's participation constraint; that is, in the long run payment of less than cost would result in exit from the industry, whereas the industry has in fact been growing substantially.[12]

By contrast, in Canada and Germany, governments have tried to lower fees. Although this can be seen as reducing rents, it might also be seen as holding up physicians who have made investments in their education, because the physician who has invested in training will usually have a distinctly inferior alternative to medicine after the training. Hence, he or she is likely to continue to practice even if fees are lowered. Over time, of course, entry could well be affected if fees are kept below competitive rates. My personal experience with the political economy of the American Medicare program, however, suggests that tax financing is no guarantee of the absence of rents. In that program lobbying by providers for higher rates and other forms of rent seeking are ubiquitous.

Many years ago Vincent Taylor and I proposed what we termed Variable Cost Insurance, a mechanism we thought would bring greater price competition to the provider market (Newhouse and Taylor 1970, 1971a,b). The essence of our idea was that providers should quote a unit price; consumers would then choose a provider and their insurance premium would vary according to the unit price of the provider chosen. Although this idea was something of a precursor to the type of preferred provider and point-of-service arrangements one sees today in the United States, as proposed it had two difficulties:[13]

1. Providers varied substantially in their style of care or volume of services they provided to similar patients.[14] Thus, the cost or "loss" incurred by the insurer if the patient sought care from a certain provider was not necessarily well correlated with the unit price charged by that provider. Directly relating the premium to the named price, therefore, did not internalize the proper incentives.

2. Although most consumers could name a primary care or first contact physician, they could not necessarily name what other physicians they might want to use in various states of the world. For example, if they were diagnosed with heart disease, they might want to consult, or their physician might want to refer them to, a cardiologist, but perhaps only if the heart disease were sufficiently severe; if they then needed surgery, they would want a cardiac surgeon and an anesthesiologist, and so forth. Moreover, they would need not only to name the specialist physicians they would want in each of a vast number of possible states of the world, but also whom they would want to see in some future states of the world, which implied they would have to anticipate how the relevant medical technology might evolve.[15]

Proceeding further along the lines Taylor and I had proposed would have to await future changes in the financing and organization of medical services that would allow payer-driven competition, to use the phrase of Dranove, Shanley, and White (1993).

Rents and Administered Prices

With this as background, I am now ready to turn to the problems caused by administratively set prices. Most of the subsequent chapters take up problems of more competitive or market-oriented arrangements. One of the virtues of such arrangements is their lesser reliance on administratively set prices. In terming that a virtue, however, I am assuming that rents would be less if traditional market-oriented arrangements were more prominent in price setting.[16] Because administered prices are so prevalent, however, it is difficult to provide evidence on that key assumption. In any event, it is important to be clear about the problems of administratively set prices when appraising alternative institutions for determining medical prices.

Table 1.1
Discounted Value of Income and Rate of Return, by Specialty, 1985

Specialty	Present Value (1985 $)	Rate of Return (%)
Pediatrics	1,068,000	−3.9
General and family practice	1,075,000	−3.8
Psychiatry	1,149,000	0.8
General internal medicine	1,229,000	3.4
Medical subspecialties	1,634,000	10.4
General surgery	1,635,000	10.6
Surgical subspecialties	1,864,000	14.1
Radiology	1,888,000	14.4
Anesthesiology	1,944,000	17.9

Source: William Marder, Philip R. Kletke, and Anne B. Silberger, "Physician Supply and Utilization by Specialty" (Chicago: American Medical Association Center for Health Policy Research, 1988), 82.

A principal defect of administered pricing is the presence of rents. Because of the insurers' need to meet the reservation prices of providers as well as the weak incentives of insurers to keep provider fees down, it is plausible that traditional American fees contained substantial rents. But two pieces of evidence support the notion of rents.

Earnings by Specialty

Many years ago Milton Friedman and Simon Kuznets (1945) sought to document rents by contrasting the rate of return to physician training with that of dentist training. Later evidence in that spirit is shown in table 1.1, which gives the present discounted value of lifetime earnings using a 5 percent discount rate, as well as the implied rates of return for various specialties in 1985.[17] Clearly there is a substantial difference among the specialties.

An economist who knew nothing about medical care and who was asked to interpret these differences would probably first ask about nonpecuniary differences among the specialties. Is it the case, for example, that surgeons and anesthesiologists have more onerous working conditions than pediatricians or psychiatrists? In that case the differentials shown in table 1.1 might simply be equalizing differentials. Although one might make such an argument, it seems a bit strained.[18] For example, internal medicine subspecialists (e.g.,

cardiologists, pulmonologists) make almost a third more than general internists, yet their working conditions seem rather similar.[19]

A more plausible explanation than equalizing differentials lies in the insurance arrangements.[20] For many years indemnity insurance in the United States was much more extensive for inpatient services, whereas outpatient services were much less well covered. The rationale was that outpatient services were relatively inexpensive and hence it was not worth the consumer's paying a loading charge to insure them. Thus, if insurers tended to pay fees that included rents and if entry were controlled, those specialists whose work was predominantly hospital based, such as surgeons and anesthesiologists, would tend to earn substantially more per hour or per year than specialists whose work was primarily outpatient based, such as pediatricians.[21] Indeed, pediatrics was often referred to by its practitioners as a "cash and carry" business, because most of the practice was in the office as opposed to the hospital, and there was rather little insurance for services in the office. In sum, although there could be some element of equalizing differentials in the data in table 1.1, the greater incomes for hospital-based specialists are certainly consistent with the role of insurance in inducing rents.[22]

Sticky Prices

A second piece of evidence supporting the notion of rents is the pricing of new procedures, of which the past half century has seen an abundance. When procedures are first introduced, productivity tends to be low, but over time learning-by-doing can greatly improve productivity. Administered prices, however, are notoriously sticky. Thus, a fee, which may be set appropriately for a new procedure, may after several years of being unchanged be substantially above a competitive price because of increased productivity.[23] For example, cardiovascular surgeons and invasive cardiologists earn much more than the average physician. In 1992 cardiovascular surgeons averaged $575,000 and invasive cardiologists $364,000, whereas the average physician earned $182,000 (Center for Research in Ambulatory Health Care Administration 1993; U.S. Bureau of the Census 1999, 134).[24] In both cardiac specialties much of the work is from procedures where productivity has greatly increased, but the administered prices have not much fallen.[25] All this led to a widespread view by the late 1970s that American physician fees for many

procedures were overpriced, whereas fees for evaluation and management (intellectual) services were underpriced, a phenomenon the Resource Based Relative Value Scale (RBRVS), which I come to in what follows, was intended to correct (Hsiao, Braun, Dunn et al. 1988).

Rents in fees have at least two and possibly three negative effects on economic efficiency. First, the rents along with imperfect information create incentives for supplier-induced demand or overservicing (Pauly 1980; McGuire and Pauly 1991).[26] Second, both the rents and any associated supplier-induced demand raise financing requirements—so that premiums or taxes are higher than they otherwise need to be. The greater requirements for financing imply greater deadweight loss, meaning the inefficiency from additional taxes if a tax-based system is used or from the inefficiencies in the labor market if an employment-based premium system is used.[27] Third, and perhaps even more important but also much more speculative, the rents could lead to an excessive rate of technical change (Weisbrod 1991).[28]

One proposed remedy that potentially addresses rents in administered prices is a large deductible.[29] Although usually advocated for purposes of reducing moral hazard, large deductibles could also potentially reduce rents by inducing those consumers who do not expect to satisfy the deductible to shop more carefully for lower prices as in a standard market.[30] (Moral hazard in health care refers to services whose private value exceeds their cost to the consumer but not their total resource cost.)

But the evidence on the efficacy of this approach, which the literature terms demand-side cost sharing, is mixed. The RAND Health Insurance Experiment demonstrated that a large deductible does reduce the use of medical services by about 30 percent relative to no cost sharing. Moreover, for the average person the reduction in demand or use appears to cause little or no adverse consequences (Newhouse and the Insurance Experiment Group 1993). Thus, a large deductible does seem to reduce moral hazard, as its proponents claim.[31]

But such a deductible carries with it a number of drawbacks. First, it clearly increases the financial risk borne by the consumer (Zeckhauser 1970). This seems particularly important in the case of the chronically ill, whose spending may approach or exceed the deductible in each accounting period (typically each year). Indeed, with an

appreciable deductible that must be met each year there is a form of market failure; a person cannot insure against the financial risk of becoming chronically ill. I present some simulation results showing the effect of deductibles on risk in chapter 6.

More important for my purpose in this book, a large deductible does not address the bulk of the problem of rents in administratively determined prices, because, at the size of a deductible that appears reasonable in terms of risk aversion, the share of spending by individuals over the deductible is large, implying that much care at the margin would still be heavily subsidized. In the RAND experiment, for example, some families were randomized to a plan with a $1,000 family deductible in late 1970s dollars; this deductible was reduced for the poor.[32] In this plan 95 percent of the spending was by families that exceeded the deductible.[33] Even in a plan with 25 percent coinsurance, where it took three to four times as much gross spending to exceed the deductible, 85 percent of the spending was by families that exceeded the deductible.[34] Although Milton Friedman (1991) has proposed much larger annual deductibles that approximate median family income, his proposal seems both impractical and undesirable in terms of the risk families would bear. Even in the individual insurance market, one simply does not observe the purchase of such policies on any substantial scale.[35]

Third, the evidence from the RAND experiment was that prices paid to physicians per unit of service were approximately independent of the degree of cost sharing across plans despite considerable variation in price within specialties and sites (Marquis 1985).[36] Thus, consumers in the high cost-sharing plans either did not shop on the basis of price or were ineffectual at finding lower-priced providers. In sum, demand-side cost sharing appears to have a role to play in reducing moral hazard, especially the initiation of episodes of treatment, but it is not sufficient to achieve first best in the medical sector.

I spend relatively little time on demand-side cost sharing in this book.[37] Instead, I focus on supply prices, or the prices that providers receive. This is not to denigrate demand-side cost sharing, which I think has a role to play in medical care financing, but its function is well understood, at least conceptually, by both health economists and general economists alike. On the other hand, outside a coterie of health economists, supply-side cost sharing is less well understood. The more integrated health care delivery systems now emerging in the United States frequently employ some modest demand-side cost

sharing, as one might expect if managed care cannot easily "manage" the initiation of episodes.[38] But aside from its function to steer consumers toward certain ("in network") providers or drugs, it is generally not a large feature of such systems. I therefore mostly abstract from demand-side cost sharing.

Rents in Supply Prices and Their Effects

Rents offer physicians an incentive to deliver more services than an informed consumer might wish (Pauly 1980). The empirical literature on supplier-induced demand, which seeks to establish the degree to which this incentive is acted upon, is lengthy and in my judgment tortured. For my purposes here, I only want to establish that physicians—and presumably other providers as well, many of which are for-profit—do respond to supply prices.

A common genre of study of physician response to supply prices looks at physician behavior in response to variation in fees. One well-known study of this type found that after changes in Medicare fees, both up and down, in Colorado in the 1970s, physicians responded as if their supply curve were backward-bending; that is, in areas in which fees were reduced, the quantity of services increased (Rice 1983). This was interpreted as demand creation in order to maintain incomes. Zuckerman, Norton, and Verrilli (1998), using more recent Medicare data on how physicians responded to changes in fees, confirmed this result. A relatively recent study used cross-sectional variation in relative Medicaid fees for a Cesarean section to ascertain the supply response; in contrast with the Medicare studies, the observed response was normal, meaning that the higher the relative fee, the greater the number of Cesarean sections that were observed (Gruber, Kim, and Mayzlin 1999).[39]

Other studies have observed the behavior of physicians paid by fee-for-service and by other methods. Two of these are particularly notable. Shifting Danish physicians from full capitation (no revenue at the margin for additional services) to partial fee-for-service resulted in more services per visit, fewer hospitalizations, and fewer referrals (Krasnik, Groenewegen, Peterson et al. 1990). Another study observed the behavior of pediatric residents who had been randomized to be paid either by fee-for-service or by salary. Those in the fee-for-service arm of the trial behaved differently; their patients missed fewer recommended visits and exhibited greater continuity

of care (Hickson, Altmeier, and Perrin 1987). These two studies are particularly relevant to the discussion of stinting in chapter 3, where some additional studies of physician behavior are discussed.

Static and Dynamic Efficiency

Although I focus in this book on static inefficiency, or inefficiency at a point in time given medical technology, the sustained increase in medical care costs in almost every developed country means the amount of any welfare loss in the cost increase may be even more important than the inefficiency at a point in time (Weisbrod 1991; Newhouse 1992). Consider a developer of a new medical device, drug, or procedure. If the new product is used predominantly by consumers who are covered by insurance for their marginal dollar (e.g., over any deductible), then the usual market test for innovation is distorted. The informed, fully insured (at the margin) consumer will demand all services that offer any positive expected health benefit, irrespective of their cost. Depending on the supply price, that consumer's physician may well want to deliver all those services. How much inefficiency in the introduction of new products and procedures results, however, is problematic. Although I am skeptical that the welfare loss is as large as often portrayed and therefore skeptical of the claim that cost containment is urgent, I have left the critical topic of dynamic inefficiency mainly outside the scope of this book.[40]

In addition to the possible degree of welfare loss from the rapid rate of technological change in medical care, there is the issue of how new products and procedures will be priced. I have already mentioned the static inefficiency that results from the sticky rents in administered price systems for procedures or products where productivity improves. But new products in a world of administered prices raise another source of possible inefficiency. In a standard market the developer of a new product simply prices the product and puts it on the market. With an administered price system, however, the developer must persuade whatever agency is administering prices to allow a price sufficient to recoup the investment. In one well-known example, the case of cochlear implants to improve hearing, the price Medicare allowed was insufficient, and the product encountered problems in coming to market (Kane and Manoukian 1989).[41] Such an outcome may, of course, have been efficient in this particular case. I return to the issue of regulatory lag in pricing later in this chapter.

The Medicare Program as a Case Study of the Discontents in Administered Fee-for-Service Prices

In addition to rents, several other pathologies are associated with administered prices. I illustrate the nature of these pathologies by focusing on the fee-setting institutions of the traditional U.S. Medicare program. For the most part in this discussion, I ignore the existence of other payers. I do so principally for simplicity but also because for many years, though no longer, American private payers operated in a fashion similar to Medicare. Moreover, although the details of several pathologies are specific to Medicare, the generic problems with which Medicare grapples are common to the payment systems of most developed countries and to American private payers.

I have chosen to focus on Medicare for three reasons. First, in sheer size it is the largest health insurance program in the world. In 1999 it spent $209 billion (2.3% of U.S. GDP) and accounted for around 12 percent of the federal budget.[42] By 2010, on the eve of the postwar baby boom cohort's becoming eligible for Medicare, the Congressional Budget Office (CBO) projects that Medicare will spend 2.9 percent of GDP and perhaps 5 to 6 percent by 2030, though the latter figure is obviously highly uncertain.[43]

Second, some of the Medicare program's best known payment methods, such as paying hospitals an amount per admission on the basis of Diagnosis Related Groups (DRGs), are used by other American insurers and in other countries. Third, I know the Medicare program well, partly from having served on the Medicare Payment Advisory Commission (MedPAC), which recommends payment changes in Medicare to Congress, as well as its predecessor commissions.[44]

The Medicare program is the nearly universal public insurance program for those Americans over 65 years of age, and 85 percent of its monies go to health care for the elderly. The remaining 15 percent pay for the care of certain disabled persons (those who had worked and paid payroll taxes) and those with end stage renal disease (kidney failure). All insurance plans have idiosyncratic features, and as a result the reader, especially the non-American reader, will learn more about Medicare than he or she probably wishes to know. I take this liberty with the reader largely from my desire to be concrete.

The Medicare program now consists of two types of health plans. First there is the traditional program, patterned after traditional American indemnity insurance, which enables beneficiaries to obtain

covered services from almost any physician and hospital to the extent that their physicians deem necessary. Second, there are so-called Medicare + Choice plans, which for my purposes I take to be Health Maintenance Organizations (HMOs).[45] In this chapter I focus more on the traditional program because of its use of fee-for-service pricing, but the HMO part of the program also exhibits some pathologies of administered pricing, as I come to at the end of this chapter. I emphasize issues in the traditional program because it is by far the largest part of Medicare, accounting for 86 percent of the beneficiaries in 2001 and approximately that share of the dollars.

Implemented in July 1966, Medicare spending grew at an annual *real* rate of 7.3 percent between 1970 and 1999, an even more rapid rate than all of the American health care sector, which itself grew at a real rate of 5.3 percent.[46] In July 2000 the CBO projected that Medicare would grow at about a 3.7 percent real rate in the decade to 2010.[47] Even this rate, which seems optimistic given the historical experience, is well above the growth rates of both the entire economy and federal tax revenues, which typically grow at about the same rate as the economy, or 1.5–3.5 percent over longer periods of time.[48]

The future direction and financing of Medicare was a major issue in the 1996 presidential campaign. As a result of that campaign and in response to the projections of future financing difficulties, Congress in 1997 implemented a series of reimbursement reductions and other reforms that reduced the projected rate of growth in spending substantially. Indeed, in the 1997–1999 period Medicare spending actually declined 0.7 percent per year, although no serious observer thinks such a decline can be maintained. Because it is such a large and complex program that affects so many individuals, Medicare will surely remain high on the American political agenda for the forseeable future.[49]

Because traditional Medicare was patterned after the private indemnity insurance of the 1960s and because such insurance was designed to cover the cost of acute medical services, Medicare excludes the cost of chronic, long-term care. Moreover, at that time private insurance typically did not cover outpatient prescription drugs, and Medicare excludes them from coverage as well. In the meantime most private insurance has expanded to include drug coverage, and Medicare coverage for drugs is now under active discussion.[50] Partly as a result of these exclusions, traditional Medicare coverage is now not very generous when compared with employment-based insurance for those under 65; consequently,

many beneficiaries buy individual supplementary insurance, or so-called Medigap policies. Others have such policies provided as a fringe benefit by former employers. Here, however, I want to focus on traditional Medicare's methods for paying providers—in particular, its methods for paying for hospital and physician services, as well as post-acute or post-hospital services such as skilled nursing facilities and home health agencies.[51] Medigap policies are much less important for supply prices, because Medigap simply tends to fill in the prescribed consumer cost-sharing amounts in the underlying program.

Hospital Pricing in Medicare

When it was first established in 1966, Medicare patterned not only its coverage but also its reimbursement methods after private insurance.[52] In the case of hospitals, that meant it paid each hospital a share of the hospital's total allowable costs, where the share was proportionate to Medicare's share of patient-days at the hospital.[53] Under this pricing system hospital costs increased rapidly, rising in real terms by about 10 percent per year from 1970 to 1980.[54] As a result, starting in October 1983 the cost reimbursement system was replaced over a five-year period with the Prospective Payment System (PPS), which reimbursed a fixed amount per case (i.e., per admission).[55] The motive for introducing the PPS was explicitly to increase hospitals' incentives to produce care efficiently. The Report of the Department of Health and Human Services to Congress recommending the system minced no words on this point: "No payment system contains as many intractable undesirable incentives as does the present cost based system" (Department of Health and Human Services 1983, 33).

By fixing a price in advance for a hospital admission, the government was setting unit prices for hospital services, thereby hoping to gain control over total spending. In doing so it had to define what the hospital service was to which the price that it fixed applied.

A Brief Description of the PPS and DRGs

The cornerstone of the PPS was the DRG classification system. In this system, those admitted to a hospital are classified by their principal diagnosis, the diagnosis most responsible for the admission (e.g., a

heart attack), as well as by any major procedures that are performed (e.g., a coronary artery bypass graft operation). Because there are thousands of both diagnoses and procedures, the DRGs are aggregations ("Groups") of the diagnoses and procedures. Altogether there are around 500 groups. In carrying out the aggregation of underlying diagnostic and procedure codes, clinically related problems are kept together; thus, an admission for cancer is not placed in the same DRG as one for coronary heart disease. Subject to this constraint of "clinical coherence," admissions that are of approximately similar cost are grouped together. More specifically, the algorithm that aggregates diagnosis and procedure codes into groups minimizes the variance of within-group cost subject to the constraint of keeping clinically related problems together and a constraint that there be approximately 500 groups.

Given the approximately 500 groups into which all hospital admissions are classified, the next step is to attach relative prices or weights to the groups. Originally this was done on the basis of the average accounting cost of cases in the group, but in 1986 the average of the list price ("charges") was substituted for average accounting costs. Medicare also specifies a conversion factor, or the number of dollars it will pay for a DRG with a weight of 1.0. The size of the conversion factor was initially set on a budget neutral basis; it has subsequently been updated annually by Congress, based upon recommendations from the Secretary of Health and Human Services and the Medicare Payment Advisory Commission.[56] A full technical description of the initial system can be found in Pettengill and Vertrees 1982; see also McClellan 1997 for a description of the system. The changes in the system since 1984 can be found in various issues of the *Federal Register*, although the major outlines of the system have remained intact.

To illustrate, DRG 90 is simple pneumonia and pleurisy without complications, and in 1997 it had a weight of 0.6978. DRG 122 is acute myocardial infarction ("heart attack") without complications and discharged alive; in 1997 it had a weight of 1.1617.[57] Although the two previous examples of DRGs reflect only the patient's diagnosis, other DRGs are based on certain procedures that may be performed during the admission. A patient with an uncomplicated acute myocardial infarction, for example, who had a bypass graft operation with a cardiac catheterization, would not be classified in DRG 122 but rather in DRG 106, coronary bypass graft operation

(1997 weight 5.5564). One who had angioplasty performed rather than a bypass graft would be classified in DRG 112, percutaneous cardiovascular procedures (1997 weight 2.0946). Finally, if a patient has secondary diagnoses (also termed comorbidities or complicating conditions) that are related to the principal diagnosis, the patient is generally classified in a different DRG, reflecting the additional costs of treating such patients. For example, a patient with pneumonia and complicating conditions would not be classified in DRG 90 but rather in DRG 89 (1997 weight 1.1156) and a patient with an acute myocardial infarction with complications would not be classified in DRG 122 but rather DRG 121 (1997 weight 1.6482).

In sum, all Medicare patients in general acute care hospitals are assigned a weight, and in most cases reimbursement to the hospital is proportional to that weight. The average weight across all patients at a given hospital is termed the hospital's Case Mix Index. Thus, hospitals treating patients with more costly diagnoses are paid more. There is nontrivial variation across hospitals in the Case Mix Index; 80 percent of hospitals in 1998 had case mix indices between 1.0 and 1.7. In addition to the Case Mix Index, payments to hospitals are adjusted for the level of wages in the hospital's geographic area.[58] In 1998 80 percent of the hospitals had wage indices between 0.75 and 1.15.

Outlier Payments

For patients with exceptionally costly stays, an additional payment is made equal to 80 percent of the accounting costs above some threshold or deductible. The threshold is set so as to be a given dollar amount above each DRG's mean payment rate.[59] By law 5 percent of total payments are reserved for outlier payments; the outlier threshold is then set such that outlier payments will be 5 percent of the total.

Initially the outlier system defined two types of outliers, one for exceptionally costly patients and one for patients with exceptionally long lengths of stay, even if they were not exceptionally costly. Subsequent economic analysis led to several changes, including abolishing the length-of-stay ("day") outliers and basing payments solely on the costliness of the case (Keeler, Carter, and Trude 1988). The analysis that led to these changes cast the outlier program as insurance at

the case level, with a premium equal to 5 percent of total payments, a deductible equal to the difference between the outlier threshold and the mean payment in the DRG, and a coinsurance rate equal to the difference between marginal cost and 80 percent of average (accounting) cost.[60] From the point of view of minimizing risk it was clearly better to insure against high costs from whatever cause than long lengths of stay. In addition to eliminating long lengths of stay as a basis for outlier payments, the changes standardized the deductible across different DRGs (it had been highly variable) and decreased the coinsurance rate from 40 to 20 percent by increasing reimbursement from 60 to 80 percent of the cost over the threshold. The decrease in coinsurance was an effort to approximate better marginal cost.

The Teaching Adjustment

Two other adjustments are made to a hospital's payments, one for hospitals with teaching programs and one for hospitals serving large proportions of poor patients. I describe only the first here.[61] Hospitals with teaching programs, meaning those with interns and residents, receive two types of supplemental payments from the Medicare program, indirect and direct medical education payments. In the original work underlying the PPS, the following regression, estimated using 1979 data from 5,071 hospitals, was used to set the indirect medical education payment amount (Pettengill and Vertrees 1982):

$$\ln(\text{mean operating cost/case}) = \alpha + \beta_1 \ln(1 + (\text{interns} + \text{residents})/\text{bed}) + \beta_2 \ln(\text{wage index}) + \beta_3 \ln(\text{case mix index}) + \beta_4 \ln(\text{bed size}) + \beta_5 (\text{Dummy variable for metropolitan area} > 1{,}000{,}000) + \beta_6 (\text{Dummy variable for metropolitan area between } 250{,}000 \text{ and } 1{,}000{,}000) + \beta_7 (\text{Dummy variable for metropolitan area smaller than } 250{,}000).[62]$$

In calculating this regression, all the coefficients except β_7 were highly significant.[63] In particular, β_1 was estimated to be 0.569 with a standard error of 0.042. Thus, there could be little doubt that the intensity of the teaching program, as measured by the house staff-to-bed ratio, was correlated with a hospital's per case cost. Before the system was actually implemented in fiscal year 1984, this equa-

tion was reestimated using 1981 data, and the estimated value for β_1 was 0.5795, very close to the 0.569 value with 1979 data.

The additional costs per case at teaching hospitals could stem from many factors. At the time they were often explained as the inefficiency of patient care delivered by residents, who were learning how to treat patients. A common story was that residents would overorder tests. This reason for the additional costs, however, does not stand up well to economic analysis, as I explain subsequently.

Based on the estimated coefficient of 0.5795, the Department of Health and Human Services initially proposed to pay hospitals 5.795 percent more per case for each 0.1 increment in their intern-and-resident-to-bed ratio. These additional payments would be budget neutral; thus, nonteaching hospitals would have their rates reduced to finance them. The remainder of the payment formula, however, did not mimic the regression equation. Most important for these purposes, the payment formula took no account of bed size.[64] That is, two hospitals that were otherwise similar but differed in bed size were each paid the same rate. Omitting bed size from the formula stemmed from the prevailing view that the United States already had too many beds, and so additional beds should not be subsidized. (In light of the subsequent fall in hospital admission rates and lengths of stay, this judgment was surely correct.[65]) The Department of Health and Human Services, however, did not reestimate the regression equation omitting the beds variable to obtain a new estimate of β_1. Any such reestimation would have produced a larger value for β_1, because hospitals with many residents tend to be large (Anderson and Lave 1986). In short, the effect of the department's proposal was to pay the average teaching hospital less than its incremental costs.

The teaching hospitals protested this proposed payment to Congress, which was eager to implement the entire PPS as soon as possible and did not wish to be held up over this issue. As a result, rather than reanalyze the issue, Congress simply doubled the 5.795 percent value to 11.59 percent, taking the additional monies from payments to nonteaching hospitals (i.e., the doubling was budget neutral). Subsequently, the Congress has decreased this percentage value, although as of 2001 it has not come down to the so-called empirical level (i.e., the estimates in later years corresponding to the original 5.795% figure), showing the political difficulty of modifying a formula that simply redistributes money.[66]

Additionally, the Medicare program had from its inception in 1966 paid a share of the so-called direct costs of graduate medical education, where share was defined by the Medicare share of patient days. Most of these direct costs were the salaries of interns and residents, but they also included some faculty salary costs and some overhead costs. These costs were not included in the regression defining indirect costs just described (i.e., they were not part of the dependent variable), although economic theory would suggest that they should have been, because residents bear the cost of general training (Newhouse and Wilensky 2001). Hence, these costs were more properly attributed to patient care, which was the purpose of the indirect adjustment. Put another way, the additional costs at teaching hospitals did not reflect training costs, because those would have been netted out of the salaries paid the residents. Hence, the additional costs the teaching hospitals wrote down on their cost reports reflected something other than teaching.[67]

Excluded Hospitals and Units

Because the initial DRG system did not provide sufficient homogeneity for patients in certain specialty hospitals, patients in those hospitals were excluded from the PPS system. The most prominent types of excluded hospitals were psychiatric and rehabilitation hospitals, as well as psychiatric and rehabilitation units of general acute care hospitals.[68]

Post-Acute Providers

A series of other providers may care for patients after their discharge from the acute care hospital. Such providers include Skilled Nursing Facilities, rehabilitation hospitals or units within hospitals, and home health care agencies. In the mid-1980s, when the PPS was implemented, these providers were relatively small, accounting for only 3 percent of the program's costs. Because there was no analog to the PPS for them at that time, they, as well as hospital outpatient departments, remained largely under cost-based reimbursement, subject to limits or ceilings on reimbursable costs. Costs, however, grew at very high rates after 1988 for reasons I explore in the next section. Since 1997, however, the Health Care Financing Administration (as of 2001 the HCFA was renamed the Centers for Medicare and

Medicaid Services or CMS) has begun transitions to prospective payment systems for the various post-acute care providers.

The PPS and the Pathologies of Administered Pricing

Seventeen years of experience with the PPS now exist, and it appears to be a permanent feature of the Medicare payment landscape. Indeed, the Balanced Budget Act of 1997, a major piece of legislation on Medicare, called for the extension of the principle of prospective payment to many providers not previously covered by it—most notably, hospital outpatient departments, excluded hospitals and units (such as rehabilitation), home health agencies, and skilled nursing facilities—and these additional prospective payment systems are now being implemented. Certainly at a political level, therefore, the PPS is regarded as a successful innovation and much preferable to the cost-based system it replaced. The approbation is partly because the PPS gives Congress more budgetary control over the Medicare program and partly because the fall in hospital-days after the implementation of PPS was interpreted as an increase in the efficiency of the hospital sector.

Some evidence consistent with a decrease in cost and an increase in efficiency is shown in table 1.2; the first two years of the program saw a dramatic fall in patient-days, a drop of 15 percent from 1983 to 1984, and another 12 percent from 1984 to 1985, resulting in a combined 25 percent fall in patient-days.[69] Rogers, Draper, Kahn et al. (1990), in evaluating this change, found only modest adverse health effects, so that the cost savings from the reduction in days was mostly a gain in efficiency.[70] Moreover, independent evidence existed that in the early 1980s the medical services delivered on about a third of patient-days could have been carried out outside the hospital without adverse consequences (Newhouse and the Insurance Experiment Group 1993, chap. 5), suggesting a substantial scope for improved efficiency.[71] The magnitude of the changes in patient-days in the 1983–1985 period shows that how providers are paid can have large consequences for costs and efficiency.

Given the gain in efficiency that it seemingly brought about, it may appear churlish to critique the PPS, but in fact the PPS exhibits many of the pathologies of administered pricing; furthermore, several of these pathologies extend to other Medicare-administered pricing schemes and will be exacerbated by the extension of the principle of prospective payment to other services.

Table 1.2
Substitution of Post-Acute Care for Medicare Inpatient Hospital-Days

Year[a]	Inpatient days per 1,000 beneficiaries	Length of stay (days)	Skilled nursing facility days per 1,000 beneficiaries	Home health visits per 1,000 beneficiaries	Rehabilitation admissions per 1,000 beneficiaries
1981	3,827	10.4			
1982	3,889	10.2			
1983	3,786	9.8			
1984	3,217	8.9			
1985	2,823	8.6			
1986	2,784	8.7	268	1,106	2.8
1987	2,815	8.9	229	1,104	3.3
1988	2,804	8.9	334	1,104	3.7
1989	2,721	8.9	889	1,350	4.0
1990	2,749	8.8	749	2,052	5.1
1991	2,728	8.6	669	2,880	6.0
1992	2,642	8.4	812	3,763	6.6
1993	2,474	8.0	948	4,661	7.2
1994	2,436	7.5	1,006	6,020	7.8
1995	2,317	7.0	1,053	7,125	8.8
1996	2,056	6.5	1,053	7,546	
1997	1,979	6.2	1,519	7,519	
1998	1,895	6.1	1,527	4,590	
Average annual growth rate	−4.1%	−3.1%	15.6%	12.6%[b]	12.1%

Sources: Inpatient Days through 1993, *Health Care Financing Review*, "Statistical Supplement, 1996," Table 23. Inpatient Days, 1994 and 1995, *Statistical Abstract of the United States, 1997*, 115–116. 1996–1998 inpatient days from ⟨http://www.hcfa.gov/stats/stats.htm⟩. Length of stay through 1996: *Health Care Financing Review: Medicare and Medicaid Statistical Supplement, 1998*, 206; 1997–1999 calculated from Medicare Payment Advisory Commission, "Report to the Congress," March 2001, Table B.1. Other values calculated from Prospective Payment Assessment Commission, "Medicare and the American Health Care System," June 1997, chapter 4. SNF values for 1997 and 1998 and home health value for 1997 are unpublished data from the Health Care Financing Administration. 1996–1998 data for rehabilitation admissions are not available.

[a] Calendar year for hospital-days through 1993; fiscal year for other values. 1994 value from *Statistical Abstract* because 1994 value in Statistical Supplement excludes managed care enrollees and so is biased upward.

[b] Value through 1997 is 20.5 percent. The sharp decline in visits in 1998 reflects some undetermined mix of greater anti-fraud enforcement efforts and changes in payment that were effective in October 1997.

Average, Not Marginal Cost

The intent of the PPS is to pay average cost, not marginal cost, almost certainly because average cost is easier to calculate. Because the general view, supported by some empirical evidence, is that the marginal cost of hospital services is less than the average cost, paying average cost could induce additional hospitalization and is not efficient.[72]

Moreover, even the calculation of average cost is distorted in two ways. First, the average costs that are used are accounting costs, and the allocation of joint costs between the inpatient unit, to which the PPS applies, and other units of the hospital, such as the outpatient department, is arbitrary. Second, initially only operating costs were paid prospectively; capital costs were passed through so that hospitals had an incentive to substitute capital for operating inputs. Much like the Averch-Johnson (1962) effect in utility regulation, therefore, hospitals responded by increasing their capital intensity, from about 6 percent of total costs to 9 percent.[73] In 1991 the Congress mandated a ten-year transition to inclusion of capital payments in the PPS, which has now been completed. The length of the transition indicates the degree to which losing hospitals needed to be protected (or were successful in demanding that the political process protect them). Inclusion of capital costs in the administratively set price, however, further emphasizes that accounting costs will differ from economic costs because of the considerable degree of arbitrariness in accounting for capital costs, for example, depreciation life.

Economies of Scale

A further problem arises because small hospitals, which tend to dominate in rural areas, have higher-than-average costs. Their higher costs arise in part from stochastic demand; standard queuing theory models show that occupancy rates will be higher at larger hospitals, other things equal, as in fact they are empirically. Higher average costs can also arise from various indivisibilities. The PPS, however, does not adjust for the higher costs; implicitly it is attempting to force hospitals to an efficient scale and scope.

Doing so, however, fails to consider both the travel costs that might be imposed on rural residents from closing small hospitals, as well as the political economy of reducing federal payments to a

given community. As a result, there are numerous exceptions to the PPS for small rural hospitals.[74] These exceptions, which apply to over 20 percent of all American hospitals (though a much smaller percentage of the beds), dilute the power of the PPS for those hospitals. In other words, substantial elements of cost reimbursement exist for those hospitals.

Within DRG Heterogeneity

Many of the DRGs, especially the medical DRGs as opposed to the surgical DRGs, have substantial within-DRG variance. Coefficients of variation over 1.0 are common. There is thus scope for selection of profitable cases, although rather little has been detected.[75]

The within-group heterogeneity in cost is, of course, affected by the total number of DRGs; hence, that the number of DRGs was administratively set to around 500 was an important choice.[76] The limitation to 500 categories was justified on the grounds of administrative simplicity and understandability. The mapping from ICD-9 codes to the 500 categories, however, is done by computer, using a program known as the "Grouper," so the incremental gain in simplicity and understandability from limiting to 500 groups is far from clear. Nonetheless, the number of categories has only expanded by around 5 percent over the first fifteen years of the program.

Little analysis has been done on the optimal number of groups. The fundamental constraint on disaggregation is the accuracy and stability of the weight that is used, because the weight is estimated from the costs or charges for the cases in the DRG.[77] Indeed, some of the 500 DRG categories are aggregated for the purpose of assigning a weight, because the number of cases within the DRGs are deemed too few to develop reliable weights.[78] Thus, there are only about 350 unique weights in the current PPS.

The Medicare Payment Advisory Commission, in its March 2000 report to the Congress, analyzed the expansion of the number of categories to around 1,420 and documented that there is a nontrivial gain in payment accuracy from doing so (Medicare Payment Advisory Commission 2000a). The value 1,420 was chosen because there is an existing system that uses approximately that number of groups whose effects could be analyzed without having to develop an entirely new classification system.

Although a system with more groups would be more accurate at the level of the individual case, the chances that it would ever be introduced are uncertain, because it redistributes monies away from small rural hospitals, which tend to serve the lower-cost cases within DRGs. Congressmen from rural districts, not surprisingly, wish to see federal tax monies continue to flow to their districts, and so such redistribution will be unwelcome. In effect, expanding the number of DRGs would reduce the "export" earnings from Medicare services delivered in the district. (Since the rest of the country pays for those services through taxes, they are analogous to exports.) If the system with more groups were to be introduced, it would almost certainly have to be with a long transition or provisions to hold losing hospitals harmless for a time, just as was the case with introducing capital costs into the lump-sum payment. This is another example of the political difficulties of redistributing money in an established program.

Technological Change and Regulatory Lag

Determining a reasonably precise figure for average cost, let alone marginal cost, is not straightforward because of the rapidity with which modes of treatment for a given diagnosis change. For the most part Medicare cannot observe the prices of a competitive market, because the private market itself is distorted by extensive insurance coverage. Moreover, the private market price appears to reflect the actions of Medicare. Since the inception of the PPS in 1984 there is a negative correlation of -0.84 ($R^2 = 0.70$) between annual Medicare payment/accounting cost margins and private payment/accounting cost margins (figure 1.1). This is consistent with a game in which Medicare moves first and then hospitals contract with private payers given the Medicare price.[79] Private-payer contracts will reflect the constraint that hospitals must recover their joint costs such as the salary of the CEO.

Furthermore, the size of any update may determine the amount of technological advance that will be put in place. Congress took note of the cost of technological change in the legislation that established the PPS. By statute the executive branch and the Prospective Payment Assessment Commission (now the Medicare Payment Advisory Commission) were to recommend annual update amounts to the Congress based on the following factors: "changes in the hospital market basket [an input price index unadjusted for quality change in inputs], hospital productivity, technological and scientific advances,

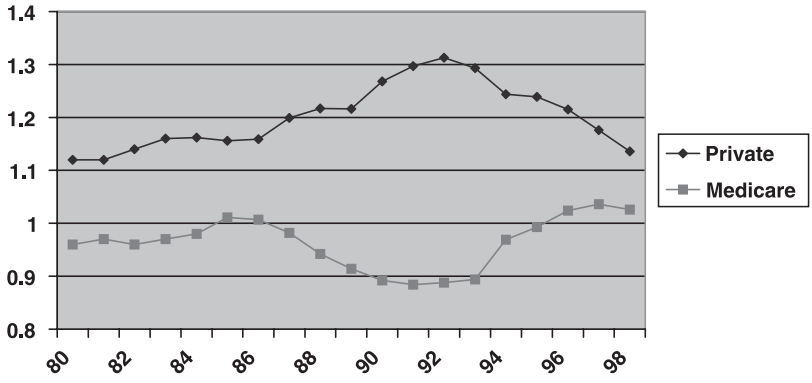

Figure 1.1
Ratio of Medicare and Private Payer Reimbursement to Accounting Cost. The R^2 for all years is 0.47. For the years of the Prospective Payment System, 1984 and later, it is 0.70. (*Source*: Prospective Payment Assessment Commission, "Medicare and the American Health Care System: Report to the Congress," June 1996, 21, and June 1997, 23. Medicare Payment Advisory Commission, "Report to the Congress: Selected Medicare Issues," June 2000 (Table C-12)).

the quality of health care provided in hospitals (including the quality and skill level of professional nursing required to maintain quality care) and [the] long-term cost effectiveness in the provision of inpatient hospital services."[80] Although this list was appropriate conceptually, in practice the only factor that can be measured in reasonably straightforward fashion is the input price index (ignoring the caveat about input quality change). The measurement of productivity, quality, and cost effectiveness founders on the difficulty of assessing the quality of care, and the measurement of technological and scientific advance begs the question of which advances one might wish to pay for.[81]

Regulatory lag may slow technological advance below its optimal level. First, the supplier of the new good or service may need to persuade the insurer that the product should be covered. Assuming it is covered, in the case of a new hospital supply, the initial DRG payment will not incorporate the cost of the good, but if it diffuses anyway, the DRGs in which it is used will increase in relative weight. In the case of physician services or outpatient supplies, which are paid for in a more disaggregated fashion than the DRG payment, payment will require issuing a new billing code, which may result in delay. For further discussion of lags in Medicare's coverage of new products, see Newhouse 2002.

Unbundling

A hospital that is paid a case rate has an incentive to break out services from the bundle for which the fixed rate is paid if it can bill additional amounts for the services it breaks out and thereby increase reimbursement. Because additional revenue was available for post-acute care use, hospitals acted upon this incentive by shifting care from the inpatient unit to the post-acute setting. As a result, although the initial effect of the PPS in 1984 and 1985 was simply a decrease in patient-days with little change in post-acute care use, between 1988 and 1996 there was a further 16 percent reduction in inpatient days and a large increase in so-called post-acute care days (Skilled Nursing Facility or SNF, home health, rehabilitation hospitals and units), as well as hospital outpatient care (tables 1.2 and 1.3).[82]

Table 1.3
Medicare Program Spending for Outpatient Facility Services, 1983–1997

Year	Outpatient department payments (billions of 1996 dollars)
1983	$4.8
1984	5.4
1985	6.2
1986	7.1
1987	8.2
1988	9.0
1989	9.3
1990	10.0
1991	11.0
1992	12.1
1993	13.2
1994	14.6
1995	15.8
1996	16.6
1997	16.9
1998	18.1

Source: Medicare Payment Advisory Commission, "Report to the Congress," March 1999, Table 6.1, and "Report to the Congress," June 2000, 36. GDP deflator used to convert to 1996 dollars. Data exclude payments to physicians and ambulatory surgery centers.

Many SNFs, as well as rehabilitation units, are physically located in a hospital; thus, the patient using post-acute services may simply be discharged from a general medical and surgical floor and wheeled on a gurney to another floor of the hospital building. The hospital, of course, collects not only the DRG payment for inpatient services but also the additional revenue for the post-acute services. Importantly, the PPS system for several years did not adjust the DRG rate for this unbundling. Between 1998 and 2002 the hospital update framework used by the Medicare Payment Advisory Commission included a "site-of-service" adjustment, the intent of which was to adjust (or "rebase") for the unbundling, but as of 2001 the Medicare Payment Advisory Commission estimated that only about two-thirds of the unbundling was adjusted for.[83] Moreover, the initial windfall gains from the unbundling have remained with the hospital industry.

The Balanced Budget Act of 1997 sought to discourage unbundling by modifying hospital payment in ten DRGs that made frequent use of post-acute care (e.g., stroke, hip fracture). Specifically, if a patient was in one of ten DRGs, used post-acute care, and stayed in the hospital less than the geometric mean stay for the DRG, the hospital was paid a per diem rather than a per case payment for the hospital stay.[84] This effectively changed the incentive at the margin to keep the patient in the hospital and made payment more neutral between a marginal day in the hospital and a marginal day of post-acute services. Another way to put this is that for some patients, payment policy became lower powered (Laffont and Tirole 1993).

Treatment of New Entrants

All administered price systems face the problem of how to treat new entrants, and the remarkable growth of post-acute care after 1988 was facilitated by generous reimbursement of new entrants. New entrants were reimbursed their costs for at least their first two years, subject to rather generous limits, and some of them were allowed to keep a portion of any subsequent cost reductions, further increasing the incentive for high initial costs.[85] Because output at these new facilities often grew over time, any economies of scale added to their profit, as did learning-by-doing.

The generous treatment of new entrants led to a dramatic increase in the number of facilities. For example, between 1990 and 1996 the number of skilled nursing facilities increased 7 percent annually and

the number of home health agencies increased 9 percent annually (Prospective Payment Assessment Commission 1997a, 105). The Balanced Budget Act of 1997 considerably tightened the reimbursement for new entrants into post acute care, but the horse was long since out of the barn (Medicare Payment Advisory Commission 1998).

Many of the new entrants, especially entrants into the home health industry, were for-profit firms. By 1998 58 percent of the agencies were for-profit firms, whereas prior to 1981 such firms were not permitted to be Medicare home health providers; at that time most home health care agencies were visiting nurse associations (VNAs).[86] It was easier, of course, for for-profit firms to raise the capital to exploit the profitable opportunities for new entrants that Medicare offered.[87]

Interactions among Different Payment Systems for Sites that Are Substitutes

For many years only inpatient hospital care was reimbursed on a prospective system; post-acute care providers were paid largely on the basis of cost, as were hospital outpatient departments. Thus, in principle profitability was unaffected by where post-acute care was received. Now, however, not only post-acute providers but also hospital outpatient departments are being moved to prospective payment systems. This raises the distinct possibility that patients will be cared for in the site that yields the maximum profit, although that may be neither the most appropriate site clinically nor the most convenient for the patient.

Consider, for example, a stroke patient who needs speech therapy. The therapy could be delivered in the acute care hospital's rehabilitation unit or in a freestanding rehabilitation hospital, in the SNF (either the hospital's or a freestanding SNF), in the outpatient department, or at home. But with each of these facilities paid at a different fixed rate for the same procedure, the amount of reimbursement from giving the therapy in a different place could differ substantially. The differing rates arise for two reasons. First, rates reflect the average cost of the cases treated in the different facilities, and the case mix of the providers differs—home care patients, for example, tend to be in the best health. It is unlikely that the case mix systems that are being implemented can fully adjust for the differences across sites.[88] Second, accounting costs of the facilities differ

because some of them are part of hospitals, and the PPS has given hospitals an incentive to shift as much joint cost as possible to these facilities and away from inpatient care in order to take advantage of cost reimbursement for these facilities.[89]

These problems could be addressed by paying a lump sum for the entire hospital episode including the post-acute care, a so-called bundled payment. Although paying a lump sum raises concerns about possible stinting (i.e., such a system may be too high powered), the prospective payment systems for post-acute care that are being developed make lump-sum payments to the specific post-acute providers; hence, the incentive to stint would not be any greater under this system than under the system being put in place. The incentive to stint could be mitigated by basing a portion of the payment on the number of services delivered, an issue I take up with respect to health plan payment in chapters 5 and 6. Existing post-acute providers, however, are strongly opposed to bundling, fearing that rents they now receive may be eliminated if hospitals could contract for services.[90] Their opposition has succeeded in blocking this reform.

A similar, but more intractable problem arises with payment of hospital outpatient departments in part because there are several relevant margins; outpatient department care may substitute for inpatient care, for care in the physician's office, or for care in an ambulatory surgery center.[91] For example, reimbursement for many outpatient procedures can differ substantially, depending on the site. Three examples are given in table 1.4. Medicare is currently changing its methods for reimbursing hospital outpatient departments; there is little reason, however, to believe that the new system will be more neutral across sites than the old.

Distortion of the Market for Interns and Residents

The teaching adjustment appears to have led to a major expansion of the number of house staff, as accords with a simple model of a subsidy in a competitive market. Before examining that model, I note that the labor market for residents approximates a competitive market. On the buying side there are around 1500 teaching hospitals throughout the country, with perhaps 100 to 200 being "major" teaching hospitals. On the selling side there are over 20,000 first-year

Table 1.4
Reimbursement Rates for the Same Outpatient Procedure in Different Sites, 1998

CPT (Procedure)	Office practice expense	Ambulatory surgery center	Hospital outpatient department
56350 (Hysteroscopy)	$95	$481	$675
58120 (D&C)	115	458	720
58340 (Catheterization and introduction of saline solution for hysterosonography)	356	n/a	155

Sources: Federal Register, June 5, 1998, June 12, 1998, September 8, 1998.

residents at any point in time. Collusion on either side of the market would therefore appear difficult.[92]

Figure 1.2 shows that the teaching adjustment has worked more or less as one would have expected a subsidy in a competitive market to work. Given the approximately fixed number of U.S. medical school graduates (around 16,000 per year), hospitals have increased the number of residents by hiring graduates of medical schools outside the United States and by lengthening the training period. The number of residents is up about 30 percent since 1985 (table 1.5).[93] The figure has been stable since 1993, suggesting that by that time a new equilibrium had been reached.

Rent Seeking

Any administered price system is vulnerable to rent seeking. With Medicare taking around an eighth of the federal budget, the amount of redistribution in even seemingly small changes in reimbursement can be substantial. Hence, a strong incentive exists for provider groups to attempt to influence policy, and virtually every provider group engages in some sort of lobbying effort on its behalf. This is not surprising, but it means that more efficient reimbursement systems may not be feasible because of distributional considerations. Several examples of the political difficulty of changing the system, such as increasing the number of DRGs, have already been mentioned, and there are many other examples.

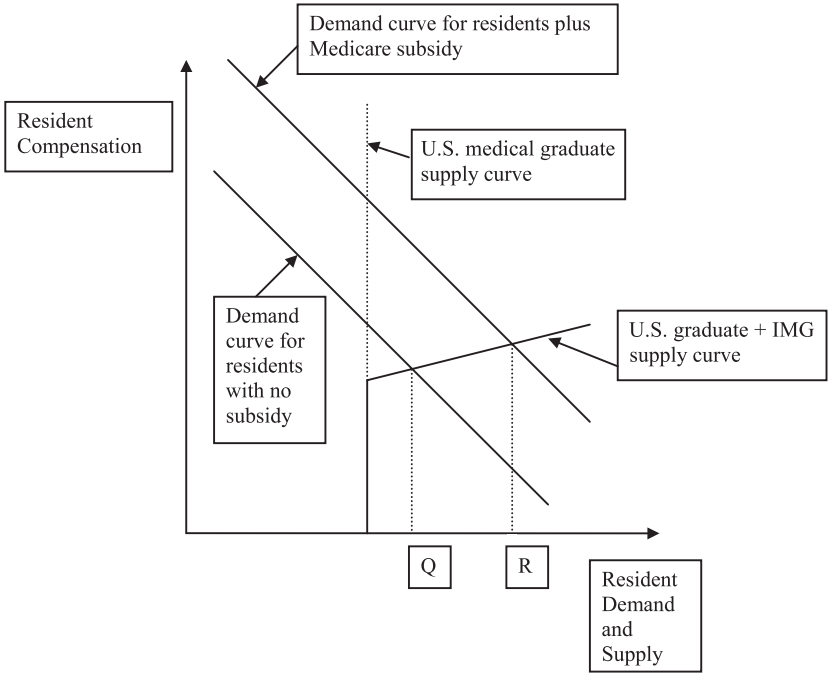

Figure 1.2
Demand and supply of residents, with and without the Medicare subsidy. With the subsidy equilibrium quantity increases from Q to R, IMG = International Medical Graduate.

How High-Powered Is the PPS?

I end this section on hospital pricing by commenting on the popular impression that payment under the PPS is independent of a provider's actions. In other words, much writing treats the PPS as if it were a high-powered payment system that, except for the outlier payments, pays a lump sum per case that is independent of provider actions. Moreover, because the outlier payments are only 5 percent of the total dollars, their influence is often thought to be modest. Mark McClellan (1997), however, has shown that this impression is misguided for two principal reasons.

First, in over 40 percent of the DRGs, payment is related not only to diagnosis but also to the performance of specific procedures. Most of these are related to the performance of surgical procedures. Virtually all the DRGs that have been added since the beginning of the PPS are treatment rather than diagnosis related. Thus, by performing

Table 1.5
Numbers of House Staff, by Year

	Number of residents		Percent international graduates	
Year[a]	First year	Total	First year	Total
1980	18,702	61,465	21	20
1985	19,168	75,518	14	17
1990	18,322	82,902	19	18
1991	19,497	86,217	24	20
1992	19,794	89,368	25	20
1993	21,849	97,370	27	23
1994	21,949	97,832	27	24
1995	21,372	98,035	26	25
1996	21,394	98,076	25	25
1997	21,808	98,143	24	26
1998	21,732	97,383	24	26
1999	22,320	97,989	26	26

Source: 1980–1992, Physician Payment Review Commission, "Annual Report, 1997," 352. 1993–1998, Rebecca S. Miller, Marvin R. Dunn, and Thomas Richter, "Graduate Medical Education, 1998–1999," *Journal of the American Medical Association*, September 1, 1999, 282(9): 856. 1999, Sarah E. Brotherton, Frank A. Simon, and Sandra C. Tomany, "U.S. Graduate Medical Education, 1999–2000," *Journal of the American Medical Association*, September 6, 2000, 284(9): 1122, Figure 1. The figures on first-year residents in the first source and second source for 1993–1995 are discrepant; the second source has been used. (The figures on total residents are the same.)
[a] Year is the academic year beginning in year shown; for example, 1996 is academic year 96–97.

the procedure the hospital incurs higher costs but also receives higher reimbursement. For example, a patient who has suffered a heart attack and is not catheterized and does not have a bypass operation or angioplasty will be classified in DRG 121, 122, or 123, whereas a patient who has one of those procedures will be classified in a (much) higher-weighted DRG, as noted previously.

McClellan (1993) shows that such treatment-based DRGs might be optimal under reasonable assumptions about demand and production technology. In particular, if hospitals have some market power and can impose capacity constraints, making DRGs purely diagnosis based can lead to underinvestment in treatment intensity. For example, if care for all heart attacks were reimbursed at the same rate, the hospital may well not make the investment in the specific capital needed to support catheterization units, bypass grafting, and angioplasty, because all of the incremental costs associated with those

procedures would reduce its residual dollar for dollar. This result is in the same spirit as a model of Chalkley and Malcolmson's (1998) that I describe in chapter 3. In their model and in the absence of agency, quality will be at the minimal legally permitted level if there is some element of prospectivity in reimbursement (i.e., if reimbursement is not fully cost-based).

Second, the outlier system makes the system lower powered than it otherwise would be, because for outlier cases reimbursement depends upon the quantity of services delivered. And McClellan (1997) showed that the effect of the outlier system on the power of the PPS is substantially larger than might be conveyed by its 5 percent share of payments.

More generally, McClellan tried to measure the extent of cost sharing in the PPS.[94] Using variation in cost and reimbursement across patients in 1990, he found that an additional dollar of reported Medicare costs at an average hospital was associated with 55 cents of additional Medicare reimbursement. This proportion varied substantially with patient demographic characteristics, treatment choices, and diagnoses. McClellan focused on the implications of this dependence for technological change and cost increases, but his presumption is that the degree of cost sharing should vary with elasticities of supply (Ramsey pricing on the supply side) and also with the degree of agency.[95] In particular, to the degree that hospitals are less responsive to reimbursement levels, reimbursement can be a greater function of cost.

Important for my purposes in this book, McClellan found that the information in diagnoses, as aggregated in the DRG system, explained less than 20 percent of the variance in both reimbursement and cost across all cases. In other words, a prospective system based purely on diagnosis does not match cost variation at the patient level very well. (Of course, the amount of observed cost variation could well be less if payment were not a function of procedure.) Indeed, as already noted, this is presumptively why the PPS includes treatment-related reimbursement features; otherwise the system would discourage the provision of potentially beneficial treatments.

By contrast, procedure as embodied in the DRG system explained around 30 percent of the variance in cost across cases in both 1987 and 1990; thus, how the PPS reimburses hospitals for the specific services given to a patient is important. This 30 percent of explained variance was almost entirely attributable to surgical admissions;

reimbursement for medical admissions was nearly invariant to what is done in the hospital.

Although only accounting for five percent of the total dollars, outlier payments explained 30 percent of the variance in reimbursement in 1987 and fully 46 percent in 1990. The noteworthy amount of variance explained by outlier payments implies that there is substantial variation in reimbursement across patients that neither diagnosis nor major procedure can explain, a theme to which I return in the discussion of partial capitation in chapters 5 and 6.

That the PPS is not independent of the actions of a provider should not necessarily be treated as a distortion, an essential consideration for subsequent chapters. Although the lack of independence could well lead to higher a higher cost of production for treating certain illnesses, full supply-side cost sharing, meaning in this context placing all the risk of a higher cost admission on the hospital, is likely not optimal, as I attempt to show in subsequent chapters. I am, however, getting somewhat ahead of the story.

Physician Pricing in Medicare

Medicare uses a different administered pricing method for physician services, and so the resulting pathologies differ from those for hospital services. Unlike the PPS, which classifies hospital treatment into one of 500 groups, there are over 7,000 different codes for physician services, each of which has its own price. In other words, payment for physician services has a much more disaggregated basis of pricing than does payment for hospital services.

For the first twenty-six years of the Medicare program, the prices Medicare actually paid for each physician code were based on "usual, customary, and reasonable" fees, the definition of which was sufficiently complicated that it is not worth describing here. Suffice it to say that the scheme was initially based upon existing fees, with complex rules for updating the fees. Like many administered price schemes, the method grew more complex and more unwieldy over time. Largely in response to the view that the relative prices of procedures (e.g., surgery) were too high relative to so-called evaluation and management services (e.g., taking a patient history), the RBRVS was launched in the 1980s to create a new set of relative prices.

The scale relied upon physician ratings of relative "work" within specialty; for example, a sample of general surgeons was asked to

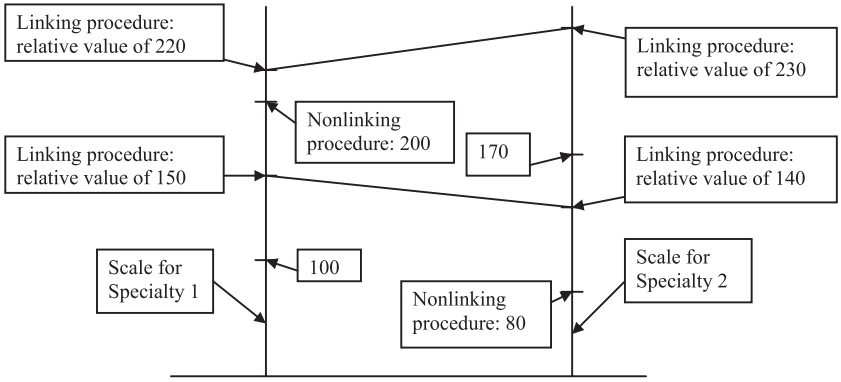

Figure 1.3
A schematic of relative prices in the RBRVS. The nearly horizontal lines connect values for two linking procedures and minimize the sum of squared errors for prices of specialty 2, given that specialty 1 has prices of 150 and 220 for the linking procedures: $(150 - 140)^2 + (220 - 230)^2 = 200$. The other tick marks show relative prices of nonlinking procedures.

rate the relative work in repairing an inguinal hernia relative to an appendectomy and a cholecystectomy. Payment for each procedure was then to be proportional to work. Empirically, physicians had a substantial degree of consensus on relative prices or weights for procedures within their own specialty.[96]

In order to set relative prices across specialties, an effort was made to find "linking" procedures that two or more specialties commonly perform (e.g., both neurosurgeons and orthopedic surgeons perform laminectomy) or procedures that a panel composed of physicians from many specialties regarded as equivalent work. Relative prices for such linking procedures were then set so as to minimize the sum of squared errors for the linking procedure prices. Figure 1.3 illustrates the procedure by showing relative values for a small number of hypothetical procedures, including two linking procedures. For more detail on the construction of the RBRVS see Hsiao, Braun, Dunn et al. 1988; Hsiao, Yntema, Braun et al. 1988; Becker, Dunn, and Hsiao 1988, and Braun, Yntema, Dunn et al. 1988.

What Is the Optimal Fee?

Setting aside for now issues of unobserved or noncontractible physician effort to minimize the cost of achieving a given outcome, an

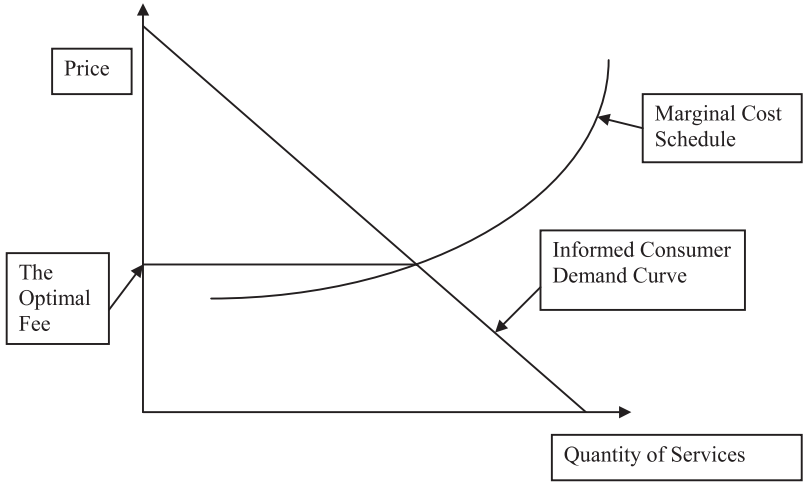

Figure 1.4
The optimal fee.

issue to which I return in chapters 3 and 5, a basis for determining the optimal fee for a single physician service was set out by Mark Pauly (1980) and is shown schematically in figure 1.4. The model depicted shows the demand curve of an informed consumer and a physician-level marginal cost curve; the model assumes a competitive supply side.[97] Suppose the physician has some discretion over the level of demand for the service and responds to the level of fees in the intensity of treatment he or she recommends to the patient. The physician is assumed not to want to induce demand for services that are of little or no value to the patient, but will do so if the reward, or the excess of fee above marginal cost, for doing so is sufficiently high.[98] The key point is that where the marginal cost curve cuts the informed consumer's demand curve, the physician has neither an incentive to deliver excess services nor an incentive to stint (underserve).

How Well Does the RBRVS Match the Optimal Fee?

Although the developers of the RBRVS intended that it match the outcome of a competitive market (Hsiao, Braun, Dunn et al. 1988), which would be the optimal fee shown in figure 1.4, in practice it cannot do this for many reasons.

The Conversion Factor and Updating

At most the RBRVS only yields appropriate relative fees. Congress then sets a conversion factor that translates the relative value into a dollar value. In doing so, it has to take account of the recommendations of the Secretary of the Department of Health and Human Services and the Medicare Payment Advisory Commission.[99] When the transition to this new set of relative prices began in 1992, the initial conversion factor was set so as to be budget neutral. Because of the high likelihood of rents in the prior system, however, the resulting fees were almost surely above competitive levels. In other words, budget neutrality implied that any prior rents were simply redistributed across services.

In principle, any initial rents could have been squeezed out of the prices over time by adjusting the rate of increase of the conversion factor downward. Subsequent updates to the conversion factor, however, were set on a formulaic basis. The initial system was termed the Volume Performance Standard (VPS) system, and the current system is termed the Sustainable Growth Rate (SGR) system. The essence of both these systems is to make the conversion factor an inverse function of the total number of units of service, so that total spending will equal a budgeted figure.[100] That is, if the quantity of services rises above what was planned or desired, the price paid per service will be reduced, so as to keep total spending constant.

Congress's dominant motive for adopting this method was to impose budgetary control on Medicare spending for physician services, which had grown in real terms by 8.8 percent per year from 1970 to 1990; by contrast, the real growth in the economy was 3.2 percent per year and in federal tax revenue was 3.0 percent per year.[101] Initially Medicare financing for physician services was financed equally between federal general revenues (mainly personal and corporate income taxes) and elderly premium payments.[102] But the 8.8 percent real annual increase in spending was not only greater than the growth in tax revenue; it was also much greater than the growth rate of the elderly's income. In a contest between the elderly and (the mainly nonelderly) taxpayers, the elderly won; Congress acted to shield the elderly from the increase by steadily lowering the share of physician spending to be financed by elderly premium payments from 50 to 25 percent of the total. Thus, the financing of this increased spending on physician services fell increasingly upon

federal general revenues, and in fact by the time the RBRVS was adopted, Part B of Medicare, 75 percent of whose dollars covered physician services, had become the largest domestic program financed from general revenue.[103]

When the VPS constraint on total spending was adopted in 1992, however, it was portrayed as giving physicians an incentive to "control" the volume of services, which were generally viewed as excessive because of the rents in the fee-for-service system.[104] A collective incentive to hundreds of thousands of physicians, of course, makes no economic sense because of the free-rider problem; to any individual physician the loss of income from "controlling" his volume of services swamps the increase in his fees from such control, because the fee increase is averaged over hundreds of thousands of physicians. This collective incentive, however, turned out to have a pernicious side effect, another instance of the importance of the political economy of the program. To understand the side effect requires a little background.

Although a principal rationale for the VPS was to control spending, it had a feature that was designed to allow for growth in spending from new capabilities in medicine. A five-year moving average of past quantity increases was computed, and each year this moving average entered the calculation of the target rate of increase. Thus, other things equal, if the five-year moving average had increased 3 percent per year, the target rate of increase was to increase 3 percent per year. The logic was that the additional quantity of services was attributable to new procedures and devices, which should be available to the elderly, and an average growth rate would approximate the monies that should be devoted to financing increased capabilities. In other words, quantity-increasing technical change was assumed to occur at a reasonably constant rate.

After the program was enacted, however, surgeons seized upon the rhetoric that the VPS was a collective incentive to hold down services to argue that surgical services should not be in the same pool as services supplied by internists and others for purposes of computing the rate of increase of fees. Their position was prompted by a more rapid growth in the quantity of nonsurgical than in surgical services.[105] Congress, responding to the surgeons' argument that they should not be responsible for or disadvantaged by other physicians' "profligate" use of services, allowed a separate conversion factor for surgeons effective in 1991.

Recall that the growth rate in fees in any given year is inversely related to the growth in the quantity of services. As a result, the separate surgical conversion factor combined with the slow growth in the quantity of surgical procedures initially gave the surgeons very high increases in fees, approaching 10 percent per year in real dollars for a few years. From the logic of the collective incentive, this was a "reward" for keeping down volume. Because the marginal cost of surgery almost certainly did not rise by anything like 10 percent in real terms, any rents that surgeons were receiving initially increased.[106]

Recall, however, that the expenditure target was a function of a five-year moving average of past quantity or volume increases (the technological change factor). As a result, after a few initial years of large fee increases from the small changes in quantity of services, the years with little change in quantity started to reduce the five-year moving average of quantity increase, which meant the target for the overall increase in spending on surgical services was to fall. Indeed, the fall would have been so large that, had the formula continued, surgeons' fees would have actually decreased in nominal terms. Implicitly the formula was sending a signal that quantity-saving technical change was occurring in surgery. Moreover, Congress, in its frustration over quantity and spending increases among all physicians, had enacted provisions that arbitrarily reduced the spending target below what it otherwise would have been, which added to the potential decrease in the surgeon's fees.[107]

Partly because such a fall could have jeopardized surgeons' willingness to supply services to Medicare beneficiaries, Congress in 1997 abolished the VPS system and replaced it with the SGR system. This new system retains the expenditure target feature of the prior system, but ties the growth of the target to the growth of GDP rather than to the five-year moving average of the growth rate in quantity. Thus, fee increases now depend on the growth in the quantity of services relative to GDP growth. Although the demise of the VPS is partly a case study in the difficulty of managing an administered price system, it also illustrates the point that there is nothing in either the five-year-moving-average method (VPS) or the change-in-GDP method (SGR) for updating the conversion factor to suggest that the resulting dollar fee will equal marginal cost.[108]

Moreover, by 1997 a number of years of differential updates to surgical services and evaluation and management services had

occurred, which caused the conversion factors applied to different services to differ by 21 percent. This difference undid the logic of the linking across procedures performed by various specialties described earlier and thus threatened the integrity of the RBRVS. In the 1997 Balanced Budget Act, Congress therefore mandated that there be one conversion factor for all physician services, as in the original system. This was done in a budget-neutral fashion, so fees for surgical procedures did fall.

Average Costs, Not Marginal Costs

Just as with the PPS for hospitals, the physician pricing system intends to estimate average cost rather than marginal cost. But setting physician fees to approximate marginal cost is probably even more important than for hospital services because of the physician's potentially greater ability to induce services.

Because of lumpiness in ancillary personnel hours, capital equipment, and office space, average costs probably do not equal marginal costs for many physician services.[109] If marginal and average cost differ, two problems arise. First suppose that marginal cost is always below average cost because economies of scale are not exhausted over the relevant range and that Medicare is the only payer. Then the RBRVS method must pay an additional amount that exceeds marginal cost in order to keep physicians in business. If this amount is paid as a per unit subsidy (i.e., a simple increase in the fee for each service), an incentive to induce demand will arise. With other payers and Medicare moving first in setting price, the subsidies will come from other payers and in practice are likely to be paid as a per unit subsidy, resulting in the same outcome.

Alternatively, suppose that there is a textbook U-shaped cost curve, or perhaps economies or diseconomies of scope, and that some physicians are not operating at the minimum point where marginal and average costs are equal. Because of spatial differentiation among physicians, among other reasons, such differences could exist in equilibrium in a private market. The RBRVS method ignores these economies or diseconomies and implicitly attempts to force all practices to the least cost scale and scope. As in the case of rural hospitals, in small markets the least cost scale and scope may not be feasible because the market is too small.

Geographic Adjustment

The Medicare program is a national administered price system. Medical care markets, however, are mainly local. Moreover, Medicare as a large but generally not dominant payer must compete with private insurers in attracting physicians to serve its beneficiaries. The fact of competition is dramatically underscored by physician reaction to the Medicaid program, the American program for certain low-income and disabled persons. Many states have set Medicaid fees well below both the private market and Medicare, and, as a result, many physicians will not accept Medicaid patients.[110]

Because Medicare competes in local markets, it would optimally set a price appropriate for each local market. If its fees fall too far below fees paid by private insurers in any given market, the Medicare program could begin to look like the Medicaid program. This is, however, politically unthinkable. Numerically the beneficiaries of the Medicaid program are mainly low-income women and children, many of whom do not or, in the case of the children, cannot vote. Moreover, the elderly beneficiaries of the Medicare program vote at higher rates than any other age group, and for many of them Medicare is a decisive issue in determining their vote. It is similarly an important issue for their adult children (Blendon, Altman, Benson et al. 1995). Thus, the law setting up both the Prospective Payment Assessment and Physician Payment Review Commissions to monitor Medicare gave the commissions a special charge to monitor "access." As a result, the annual reports of the commissions have generally contained a chapter concerning access. Although access has a number of meanings, the principal meaning in this context is that Medicare beneficiaries should have no trouble seeing a physician if they wish to.

The result is that Medicare fees must not fall far below private fees in each market. But the Medicare fees are set centrally. They do have a geographic adjustment factor, which is an index of input prices. Inevitably, however, local markets will have more or less competition. Towns of under 25,000 inhabitants, for example, rarely have more than one hospital, and even considerably larger towns may have only one hospital. Similarly, smaller towns may have only one specialist of a given type. For very specialized services, such as burn services, there may be only one facility for a substantial

geographic area. Consequently, private prices in local markets may contain varying degrees of markup over a competitive price, but Medicare has no mechanism to account for this variation. As a result, the difference between Medicare and private prices almost certainly varies from market to market, although the difficulty of obtaining transaction prices for private payers makes it hard to ascertain the degree of variation. To prevent access problems in any market, however, the variation implies that Medicare will pay rents in more competitive markets in order to match rents in private fees in less competitive ones.

Overhead Expense

On average in the United States about half of a (nonsalaried) physician's gross income goes to pay practice expenses and the other half is net or take-home income. Many expenses, such as rent, utilities, and accounting, however, are joint costs across the services or product lines of the physician, and there is no nonarbitrary method of allocating such costs to one of the thousands of disaggregated services for which Medicare is setting prices. Moreover, any such allocation could readily lead to fees above marginal cost for the specific service.

Given the possibility of demand inducement if fees exceed marginal cost, Wedig (1993) and I (Newhouse 1991) have both proposed that Medicare adopt a supply-side variant of Ramsey pricing.[111] Under this proposal Medicare would allocate joint expenses disproportionately to services where inducement is high, or so-called nondiscretionary services. In practice, however, the relative degree of inducement across services is unknown, and with thousands of services will never be known with great precision.

The RBRVS was developed only for the so-called work component of the physician's fee, that is, the nonoverhead half of the fee. There remained the issue of how to incorporate "practice costs," or overhead costs, into the Medicare fee schedule. The Health Care Financing Administration (HCFA) delayed for many years implementing so-called resource-based practice costs.[112] Although the work component of the RBRVS was implemented between 1992 and 1996, implementing resource-based practice costs did not commence until 1999 after Congress mandated it in the 1997 Balanced Budget Act. In the interim the HCFA continued to use the older "usual, cus-

tomary, and reasonable" portion of the fee schedule for practice cost. The HCFA initially sought to gather empirical evidence on practice costs through a survey of physicians, but its attempt failed for a variety of reasons, including low response. Ultimately the agency relied upon groups of physician-specialists to estimate practice costs that could be directly attributed to procedures. These initial estimates are now being refined by advisory panels of physicians, although the HCFA retains final authority. Practice costs that could not be directly attributed to specific services were ultimately included proportionately, thereby keeping fees above marginal cost for a given service.

Problems of Cross-Specialty Linking

Although members of a given specialty tend to agree among themselves about the appropriate relative fee for the various services they render, there is sharp disagreement across specialties on the relative fees for services that different specialties provide. The disagreement largely reflects the historical income differentials across specialties (table 1.1). Those specialties that primarily provide so-called evaluation and management services (e.g., taking histories, making diagnoses, recommending a course of treatment) argue that they are underpaid relative to those specialties that primarily perform procedures (e.g., surgery, endoscopy, radiologic services, etc.). Not surprisingly, the specialists who perform the procedures and who have substantially higher incomes, do not agree that they are overpaid. As a result, interspecialty fee differentials are particularly contentious.

As described earlier, relative fees across specialties were set using so-called linking procedures, with an algorithm that minimized the sum of squared errors of relative fees (figure 1.3). Such a procedure is appropriate if two conditions are satisfied: the services provided by the two specialties actually are identical (or are deemed equivalent), and all relative fees within each specialty are a function only of a hypothetical true value and random measurement error.

But are services with the same procedure code when performed by different specialties in fact identical? Consider a physician taking the history of a patient who presents to the physician with a fever of undetermined origin. The physician is taking the history in order to determine how to proceed in treating the patient. The diagnostic skill of the physician should vary across specialties, because of variation

in the amount of training. For this particular problem an infectious disease specialist would have had the most training and therefore would likely make the correct diagnosis most frequently, perhaps followed by a general internist, a family practitioner, and finally a (now disappearing in the United States) general practitioner. The logic of the RBRVS, however, is that a visit is a visit, regardless of the specialty that supplies it or the medical problem for which it was supplied, and thus Medicare should pay the same amount.[113] "Equal pay for equal work." Although exaggerated to make the point, this is something like saying that because all cars provide transportation, they are all similar.

Furthermore, there are over thirty specialties whose fees must be linked to each other. Some of these specialties, for example, ophthalmology and anesthesiology, mainly provide services that other specialties do not provide.[114] As a result, among the actual linking procedures are several that are judged to be equivalent across specialties. It seems evident that errors can enter at this point, and that therefore the relative values across specialties can be in error.

Estimation Errors

Ignore all the foregoing problems but assume that those responsible for determining the administered price are trying to set a price equal to marginal cost.[115] Doing so will surely be fraught with error. For openers, if the pricing system is very disaggregated, there are thousands of services to price. The data base required for reliable estimation would have to be very large. Moreover, one would need to control for any variation in the quality of service that affected marginal cost, but current measures of quality are almost surely not sufficient for this task and are better in some dimensions than in others.[116] And even if one could satisfy oneself that quality variation were adequately controlled for, ongoing changes in medical technology and learning-by-doing will change production possibilities and costs. As a result, even valid empirical estimates of marginal cost may be out-of-date by the time they are available.

In sum, for many reasons physician fees in practice are not likely to equal the optimal fee described previously but rather contain rents that will induce demand and cause distortions in financing. This is particularly the case in programs such as traditional Medicare

that offer "free choice" of physician and hence have little ability to negotiate for lower prices or improved performance by threatening not to contract with a given provider. In the United States traditional private insurance also paid providers on a fee-for-service basis and offered free choice of physician—indeed, Medicare was patterned after private insurance plans of the 1960s, especially Blue Cross and Blue Shield plans. Thus, most of the traditional American system for financing care suffered from the distortions of administered fee-for-service prices. In recent years, however, American private insurers have moved sharply away from the traditional model, as described in chapter 2. I conclude this chapter by describing how Medicare pays health plans and summarizing several ills of administered medical care prices in general and fee-for-service prices in particular.

Health Plan Pricing in Medicare

Although this chapter has principally focused on fee-for-service pricing, I end it with a discussion of how Medicare sets prices to health plans. This will serve as prelude to subsequent chapters.

During its first decade Medicare did not contract with HMOs on a prospective basis, but, beginning with a demonstration at one health plan, it began to do so in 1976 (Eggers 1980). Contracting on a prospective basis did not move past demonstration status, however, until April 1985, nearly two decades into the program.[117] At that time Medicare began to pay any qualified plan a fixed amount per enrollee per month, and all enrollees were allowed to choose the option of a health plan, provided a plan existed in their area whose enrollment was open. For the first several years of contracting, however, enrollment in prospectively paid health plans was small, only about two or three percent of Medicare beneficiaries. The overwhelming number of Medicare beneficiaries remained in the traditional Medicare program. Starting in the 1990s, however, both the number of plans paid prospectively and number of beneficiaries in such plans grew rapidly until 1997, when it has stabilized around a value of about 15 percent (figure 1.5) and has recently begun to decline (not shown).

Several features of Medicare contracting with HMOs deserve comment. The feature that the analytic literature focuses on most intensively is risk adjustment, and I begin there.

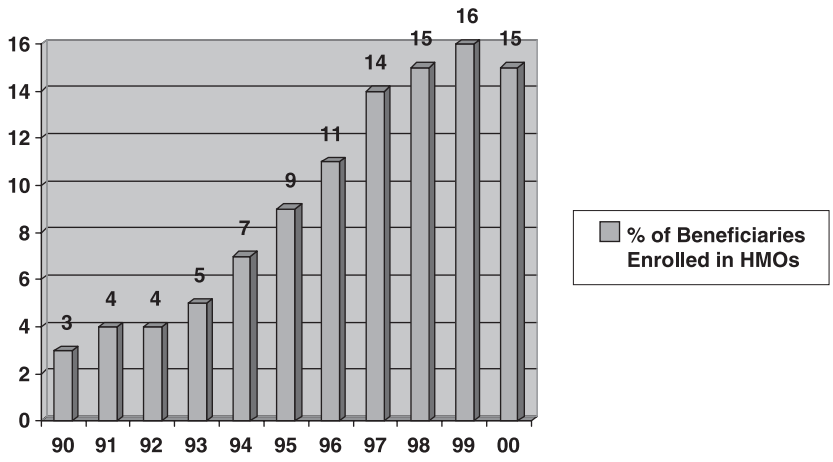

Figure 1.5
Percentage of medicare beneficiaries enrolled in HMOs. (*Source*: Medicare Payment Advisory Commission 1998, p. 5, and unpublished data.)

Risk Adjustment

The prospective amount that Medicare pays plans for each enrollee, termed the Adjusted Average per Capita Cost (AAPCC), varies with the characteristics of the enrollee, just as the PPS varies with the DRG of the patient who is admitted. Before 1998 the AAPCC was computed as follows. First, a national average payment for traditional Medicare was computed, the United States per Capita Cost (USPCC). This was then mapped to the county level by multiplying the ratio of a five-year moving average of traditional Medicare's payments in each county to the same five-year average nationally; the five-year moving average was used to achieve stability, because spending per beneficiary can fluctuate greatly from year to year in small counties.[118] This yielded an estimate of per beneficiary spending in each county, which was reduced by 5 percent as an estimate of HMO savings that might be returned to the government. Finally, demographic adjusters were introduced in the form of a rate table. From 1984 until 1998 the cells of the table pertaining to elderly enrollees were described by age (5 groups), sex (2 groups), Medicaid status (eligible or not), and institutional status (yes or no) for a total of forty groups.[119] The multiplier for each cell was simply per beneficiary spending in that cell relative to the national average. For

example, if 65- to 69-year-old noninstitutionalized males who were not eligible for Medicaid spent 90 percent of the national average amount, the multiplier for them was 0.9.

Starting in 2000, Medicare began making a transition to include Principal Inpatient Procedure-Diagnostic Cost Groups (PIP-DCGs) as an additional risk adjuster.[120] PIP-DCGs are a capitation analog to DRGs that incorporate diagnostic information into the payment. For example, all else equal, a man with staphylococcus pneumonia would have higher expected spending during the upcoming year than a man with no chronic disease. Although the old formula would have paid the same amount for the man with pneumonia as for a similar man with no chronic disease, in the new formula the plan will receive more for a man with pneumonia.[121]

When adding PIP-DCGs as a risk adjuster to the demographic variables it used previously, the HCFA lacked reliable outpatient diagnostic data. As a result, only inpatient diagnoses will "count" for reimbursement purposes (hence, the "I" in "PIP"). This nonneutrality between inpatient and outpatient care means there is a potential for substantial moral hazard, namely, patients being admitted to the hospital in order to qualify for additional reimbursement.

To reduce moral hazard, the HCFA took three steps. First, the DCGs are being implemented in a prospective fashion. That means a hospital diagnosis in a given year will not affect plan reimbursement until the following year. Moreover, the increase in reimbursement in the following year will only be the average cost of persons with that diagnosis (or in that DCG) in the following year. As a result, the average cost of the initial hospitalization will not be reimbursed. In other words, a diagnosis in year t will increase reimbursement in year $t+1$ by an amount that reflects expected spending in year $t+1$ by those who received a given diagnosis in year t. But the health plan does not receive the increment in year $t+1$ if the enrollee dies or disenrolls, further reducing the potential gain from hospitalizing a person in year t. Second, certain diagnoses that the HCFA has deemed particularly susceptible to moral hazard, so-called discretionary admissions, will not receive additional payment; they will be assigned to the lowest cost group. Third, diagnoses made during one-day stays will also be assigned to the lowest cost group. These measures probably make the cost of the great majority of admissions greater than the additional revenue the plan would receive from a higher classification. These steps to mitigate moral hazard, however,

Table 1.6
Likelihood of a Claim with the Given Diagnosis in 1995, Conditional on a Claim for that Diagnosis in 1994

Diagnosis	Likelihood (%)
Hypertension	59
COPD	62
Stroke	51
High cost diabetes	58
Dialysis	56
Quadriplegia/paraplegia	52
Coronary artery disease	53
CHF	61
Dementia	59
Rheumatoid arthritis	58

Source: Medicare Payment Advisory Commission 1998, vol. 2, chap. 2.
Note: COPD = Chronic Obstructive Pulmonary Disease; CHF = Congestive Heart Failure.

come at a price; they impair the ability of risk adjustment to reduce selection, a topic I return to in chapter 6.

The unreliability of outpatient diagnoses was shown in a 1998 analysis by the Medicare Payment Advisory Commission (1998). The Commission analyzed outpatient claims from 1994 for selected serious, chronic diagnoses (table 1.6) and determined the probability that there would be an outpatient claim with that diagnosis in the following year. Those who died during 1994 or 1995 were eliminated from the sample. Although a small percentage of the individuals with these diagnoses may have made no visits in the following year, the data in table 1.6 suggest massive undercoding of diagnosis in the outpatient data. This is not surprising because diagnoses made on an outpatient claim do not affect payment.[122] By contrast, inpatient diagnosis coding is reasonably accurate, since it is used to determine the DRG. Moreover, the accuracy of inpatient coding is audited, and coding that results in overpayment is subject to penalties for fraud.[123]

If outpatient diagnosis were to be used as a risk adjuster, therefore, adjustments in payment would have to be made for coding changes. When DRGs were initially implemented in 1984, substantial problems were found with upcoding, or coding the same patient with additional or more serious diagnoses in order to justify additional reimbursement. This was generally not fraudulent; with the

introduction of DRGs it paid hospitals to be more careful with diagnosis coding, including hiring a higher level of personnel to do the coding. As a result, Medicare payments in the initial years of the PPS were several percent above what was intended (Carter, Newhouse, and Relles 1990). Such upcoding is a once-and-for-all change and in principle can be subsequently adjusted by reducing updates, although the initial increased Medicare payments were in fact not recouped.[124]

Because the DRG system has been in place for many years, coding of inpatient diagnoses is now reasonably stable. Indeed, the HCFA estimates that in 1997 and 1999 there was no overall change in coding practices and in 1998 there was a modest amount of downcoding, probably because of increased auditing efforts (Medicare Payment Advisory Commission 2000b). Hence, introduction of risk adjustment for health plan payments based on inpatient diagnosis should not change coding practices.[125] But when outpatient diagnostic coding is introduced, potentially large upcoding (relative to the information in the claims that is used to norm the system) may be observed. For this reason it seems prudent to phase in risk adjustment based on outpatient diagnosis. Because only 10 percent of the payment is currently based on diagnosis, it probably pays providers to engage in rather complete coding, but the risk to Medicare from large payouts is reduced by an order of magnitude.

Other Features of Health Plan Payment

Through 2001 Medicare beneficiaries were able to disenroll from their health plan every month, whereas those under 65 in employer-based insurance are generally locked into their choice of plan for the year. The ability to change plans monthly is intended as a Medicare beneficiary protection, but it clearly increases the likelihood of adverse selection because beneficiaries can react quickly to changes in their health status. Starting in 2002 Medicare has moved to a modified annual enrollment period.[126]

Medicare requires that plans cover at least the services that would be covered in traditional Medicare, although plans are allowed to provide additional services. Plans are allowed to use cost sharing, and they are allowed to substitute a premium for cost-sharing amounts, but the actuarial level of the combination of premium and cost sharing cannot be above the actuarial level of the cost sharing

in traditional Medicare. In practice, this constraint has rarely been binding.

Although plans may reduce the amount of cost sharing at the point of service (and starting in 2003 they may charge lower Part B premiums), they may not actually pay individuals to enroll. As a result, there is no direct price competition among plans, and more important, there is no price competition with the traditional Medicare program.

Plans are de jure limited in the profit rate they can make. In particular, if the observed profit rate is above an adjusted profit rate that the plan earns on its commercial business, the so-called Adjusted Community Rate (ACR), the excess is to be returned to beneficiaries in the form of additional benefits or returned to the government.[127] Needless to say, any excess is always returned in the form of additional benefits. In most metropolitan areas, however, this constraint on profits is not binding because competition forces plans to pass most rents through to beneficiaries and because the adjustments to the rate used to calculate any excess profit are somewhat arbitrary.[128] That competition forces plans to pass through rents in their payments to consumers is shown by the variation in the value of additional benefits provided by plans according to the level of plan payment. As table 1.7 shows, additional benefits are markedly higher in high-payment areas where rents to plans are likely to be highest.[129]

This large disparity in benefits across areas led to a change in policy in 1997, which might be classified as another example of the pathology of administered pricing. Over a five-year period the AAPCC is moving to a 50-50 blend of the county rate and the national mean; additionally, all counties with an AAPCC below a certain level will receive at least a certain threshold or floor payment, and all areas are guaranteed at least a 2 percent update (3% in 2001). Reimbursement in traditional Medicare, however, is unaffected. As a result, payment within local markets is no longer neutral between health plans and traditional Medicare, except for those plans whose payment is near the national mean. As a result, this policy change will unbalance local markets and give enrollees in high-rate areas an incentive to shift back toward traditional Medicare as their supplementary benefits are cut. I return to the Medicare experience with health plan pricing in chapters 5 and 6.

Table 1.7
Standardized Extra Benefits as a Function of Plan Payment, 1996

Decile	Plan Payment Index	Standardized Extra Benefits
U.S. average	1.0	$77
10	1.29	121
9	1.15	86
8	1.09	80
7	1.06	86
6	1.03	92
5	0.99	78
4	0.94	68
3	0.88	57
2	0.82	53
1	0.75	48

Source: Prospective Payment Assessment Commission, "Medicare and the American Health Care System," June 1997, Table 2.8. Plans are grouped in deciles of equal numbers of plans according to the level of the AAPCC. The value of extra benefits is the actuarial value of any waived premium for noncovered services and reduced cost sharing divided by the hospital wage index for the area.

Some Lessons from the Medicare Experience

The Medicare-administered price methods illustrate several points about administered prices in general and fee-for-service reimbursement methods in particular:

1. If one assumes a single universal insurance plan, a fee-for-service reimbursement system, and free choice of physician, it is extremely difficult, if not impossible, to set up marketlike institutions that yield outcomes approximating the desirable features of usual markets. In general, there will be little or no incentive for patients to find physicians or other providers who offer the same service or same outcome at a lower cost, and providers will face nearly perfectly inelastic demand curves. The insurer must therefore set supply prices administratively.

2. It is extremely difficult in practice for the insurer to approximate optimal supply prices, especially when technology continues to change.

3. Both capitated and fee-for-service administered price systems are likely to include rents. These cause deadweight losses from additional

financing requirements and also offer incentives to use real resources to influence policy to allocate rents in one direction rather than another.

4. Related to the previous point, the political economy of the program tends to promote rents for those providers who are most effective in the political process, which in turn influence their behavior.

5. Fee-for-service pricing requires setting many prices, over 7,000 in the case of physician services alone. Setting administered prices is inevitably fraught with error, and because of lags in adapting to technological change, the extent of the error increases as pricing systems age. The United States has not, for example, "rebased" the PPS after more than fifteen years.[130] Given a fixed level of resources to devote to administering the system, it seems likely that the errors in price setting in a disaggregated fee-for-service system will exceed those in a more aggregated system simply because there are fewer administrative resources to focus on the accuracy of each price.

The errors in fee-for-service pricing lead to the discontents of fee-for-service medicine. Rents offer an inducement to overserve. Having to set many prices administratively inevitably results in distortions among relative prices of different services, which can be especially problematic when those services are substitutes, as in the case of whether a service should be performed on an inpatient or outpatient basis or where a post-acute care service should be delivered.

Partly because of the problem of pricing thousands of disaggregated services without creating incentives to overserve and other distortions, the traditional indemnity commercial insurance market in the United States has shrunk enormously in favor of managed care and capitated payment of health plans. The health plans, in turn, tend to negotiate prices with providers. For the same reason there is a desire on the part of many in Congress to increase the use of capitation and decrease the use of fee-for-service in both the Medicare and Medicaid programs.

Capitation payment to integrated health plans is the subject of much of the remainder of the book, but the fee-for-service reimbursement method is still dominant in traditional Medicare. Furthermore, fee-for-service is used by many managed care plans to pay individual providers, so material in this chapter is relevant even in a world of competing health plans that are paid by capitation.

Despite these drawbacks, many theoretical advantages of disaggregated payment systems exist relative to more aggregated systems such as capitation. In the rest of the book I emphasize two of those advantages in particular. First, capitated and other more aggregated systems are more vulnerable to selection problems because the payment for an episode of illness does not correspond to how sick the patient is. Second, because the product on which payment is based ("necessary care") is harder to define and contract for, payment systems that are more aggregated than fee-for-service are more vulnerable to stinting and unbundling. These drawbacks of capitated methods must be set against the disadvantages of fee-for-service methods. This is the conundrum of the pricing of health care services, which is taken up in detail in chapters 3 through 6.

A Remark on the Use of Salary

Some reformers consider that the use of salaries to pay physicians or other health care personnel is a way around the problems associated with fee-for-service and capitation that this book describes. Because I do not take up problems in administering a salaried system elsewhere in the book and because I do not believe salaried physicians are a solution to the health care conundrum, I make two remarks here.

First, the organization employing salaried physicians or other health professionals must receive its funds through some sort of budget. A capitation payment can be considered a per person budget, so the issues associated with capitation considered in chapters 3–6 are not so dissimilar from those associated with an overall budget. The key is the implicit incentives in the budget-setting process, and how those incentives are transmitted to influence employee performance within the organization. In any event, the basis for salary increases and promotion are surely important incentives at the individual level. Second, a salaried system requires methods to minimize shirking and more generally monitor and reward appropriate behavior. Thus, the information problems discussed in this book apply to salaried systems as well. For more material on incentives within firms, see Prendergast 1999.